Cloudera Runtime 7.3.3

Apache Impala Reference

Date published: 2020-11-30
Date modified: 2026-11-13



https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2026. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Data Durability Considerations....5
Erasure Coding Overview....5
Verify the Related Query Option....5
Enable EC Replication....5
Verify the EC Policies....6
Understand Query Performance....9
Monitor EC Metrics....9
Performance Considerations....10
Recommended configurations....11
Performance Best Practices....11
Query Join Performance....13
Table and Column Statistics....15
History based optimization [Preview]....26
Runtime Filtering....28
Distribute Runtime Filter Aggregation....31
Min/Max Filtering....31
Bloom Filtering....32
Late Materialization of Columns....33
Partitioning....34
Partition Pruning for Queries....37
HDFS Caching....40
HDFS Block Skew....43
Understanding Performance using EXPLAIN Plan....45
Understanding Performance using SUMMARY Report....46
Understanding Performance using Query Profile....47
Impala AI query profile analyzer....48
Planner changes for CPU usage....49
Calcite planner in Impala [Preview]....50
Enabling the Calcite planner [Preview]....51
Planner changes to improve cardinality estimation....51
Caching Codegen Functions....55
Scalability Considerations....56
Scaling Limits and Guidelines....61
Dedicated Coordinator....62
Hadoop File Formats Support....66
Using Text Data Files....68
Using Parquet Data Files....70
Using ORC Data Files....79
Using Avro Data Files....80
Using RCFile Data Files....83
Using SequenceFile Data Files....84
Storage Systems Supports....84
Impala with HDFS....85
Configure Impala Daemon to spill to HDFS....87
Impala with Kudu....88
Configuring for Kudu Tables....89
Impala DDL for Kudu....89
Impala DML for Kudu Tables....97
Impala with HBase....98
Impala with Azure Data Lake Store (ADLS)....102

Impala with Amazon S3....105
Specifying Impala Credentials to Access S3....109
Impala with Ozone....109
Configure Impala Daemon to spill to Ozone....110
Ports Used by Impala....111
Migration Guide....111
Setting up Data Cache for Remote Reads....116
Managing Metadata in Impala....116
On-demand Metadata....116
Automatic Invalidation of Metadata Cache....118
Automatic Invalidation/Refresh of Metadata....118
SQL transactions in Impala....121

Data Durability Considerations

You can now read from Ozone data stored with Erasure Coding (EC) feature. EC can be considered as an alternative to replication.

Erasure Coding Overview

The Ozone EC feature provides data durability and fault-tolerance along with reduced storage space and ensures data durability similar to the Ratis THREE replication approach.

Supported Engines

EC supports the following data processing engines: Impala, Hive and so on.

Before you begin

When you use EC as the data durability scheme, recovery happens in the background and requires no direct input from a user. But you must be aware of the following highlights to use EC.

- Note the limitations for EC.
- Determine which EC policy you want to use: See Understanding Erasure Coding Policies for more information.
- Determine if you want to use EC for existing data or new data. If you want to use EC for existing data, you need to replicate that data with distcp.
- Verify that your cluster setup meets the rack and node requirements described in Best Practices for Rack and Node Setup for EC.

Limitations

- Certain operations are not supported on erasure coded files due to substantial technical challenges.
- Querying an erasure coded data might have an impact on the performance. If you notice performance degradation you can troubleshoot it by running the Query profile command. It will include erasure coding related information in the profile. This will help you to quickly identify the cause.

Verify the Related Query Option

Impala already supports the basic functionalities of erasure-coding. Before this release, EC was turned off by default because Impala lacked metrics, observability and DDL commands related to erasure coding.

Procedure

1. For this release, the query option ALLOW_ERASURE_CODED_FILES is enabled by default.
2. You may use the ALLOW_ERASURE_CODED_FILES query option to disable the support of erasure coded files in Impala. When the ALLOW_ERASURE_CODED_FILES query option is set to FALSE, Impala returns an error when a query requires scanning an erasure coded file. Type: BOOLEAN Default: TRUE

Enable EC Replication

You can enable EC replication configuration cluster wide or at the bucket level or at file levels. The details of which can be found here.

Verify the created keys

You can use the following Ozone subcommand to verify that the keys in the bucket are erasure coded.

```
ozone sh key ls s3v/ecbucket
```

The resulting key description should contain a similar attribute:

```
"replicationConfig" : {
"data" : 3,
"parity" : 2,
"ecChunkSize" : 1048576,
"codec" : "RS",
"replicationType" : "EC",
"requiredNodes" : 5
}
```

Set up EC policies

Once you have set up the EC replication configuration at your desired level, you can set up supported EC policies, using the Ozone EC subcommands, for a new directory or an already existing directory. If you want to use EC for existing data, set up an EC policy for the already existing directory and then copy the data to the directory using the distcp tool or Cloudera Manager's Backup and Disaster Recovery process.

Verify the EC Policies

You can verify the EC policies you set up on the File / Table / Partition level using the Impala commands.

File level

If you set EC policies on the file level, use the SHOW FILES <TABLE-NAME> command to display if a file is an EC file, and what the EC policy is. If a file is erasure coded, you will notice an additional column “EC policy” with the related information.

```
show files in ect
+-------------------------------------------------------------------------
-----------------------------------------+------+---------------------+--
-----------+
| Path
                                         | Size | Partition           | EC
 Policy    |
+-------------------------------------------------------------------------
-----------------------------------------+------+---------------------+--
-----------+
| hdfs://localhost:20500/ec-warehouse/ect/date_col=2009-01-01/1d46b6cbf747
c7a8-a818238e00000000_1347890556_data.0.txt | 338B | date_col=2009-01-01 | R
S-3-2-1024k |
| hdfs://localhost:20500/ec-warehouse/ect/date_col=2009-01-02/1d46b6cbf747
c7a8-a818238e00000000_588711848_data.0.txt  | 353B | date_col=2009-01-02 | R
S-3-2-1024k |
| hdfs://localhost:20500/ec-warehouse/ect/date_col=2009-01-03/1d46b6cbf747
c7a8-a818238e00000000_140483252_data.0.txt  | 353B | date_col=2009-01-03 | R
S-3-2-1024k |
| hdfs://localhost:20500/ec-warehouse/ect/date_col=2009-01-04/1d46b6cbf747
c7a8-a818238e00000000_795616328_data.0.txt  | 353B | date_col=2009-01-04 | R
S-3-2-1024k |
| hdfs://localhost:20500/ec-warehouse/ect/date_col=2009-01-05/1d46b6cbf747
c7a8-a818238e00000000_234666509_data.0.txt  | 353B | date_col=2009-01-05 | R
S-3-2-1024k |
```

```
+-----------------------------------------------------------------------
-------------------------------------------+------+---------------------+--
------------+
Fetched 5 row(s) in 0.07s
```

Table level

If you set the EC policies on a table level directory, use the DESCRIBE EXTENDED <TABLE-NAME> statement to verify if the table had an EC policy set or not. If the directory has an erasure coding policy, it will be listed under storage information.

```
describe extended ect
+-------------------------------+-----------------------------------------
------------------+----------------------+
| name                          | type
                  | comment               |
+-------------------------------+-----------------------------------------
----------------+----------------------+
| # col_name                    | data_type
                | comment                 |
|                               | NULL
                | NULL                    |
| id                            | int
                 | Add a comment          |
| bool_col                      | boolean
                | NULL                    |
| tinyint_col                   | tinyint
                | NULL                    |
| smallint_col                  | smallint
                | NULL                    |
| int_col                       | int
                 | NULL                    |
| bigint_col                    | bigint
                | NULL                    |
| float_col                     | float
                | NULL                    |
| double_col                    | double
                | NULL                    |
| string_col                    | string
                 | NULL                   |
| timestamp_col                 | timestamp
                | NULL                    |
|                               | NULL
                | NULL                    |
| # Partition Information       | NULL
                | NULL                    |
| # col_name                    | data_type
                 | comment                |
|                               | NULL
                | NULL                    |
| date_col                      | date
                | NULL                    |
|                               | NULL
                | NULL                    |
| # Detailed Table Information | NULL
                 | NULL                   |
| Database:                     | default
                | NULL                    |
| OwnerType:                    | USER
                | NULL                    |
| Owner:                        | michael
                | NULL                    |
```

```
| CreateTime:                   | Tue Nov 15 15:10:21 PST 2022
                 | NULL                  |
| LastAccessTime:               | UNKNOWN
                 | NULL                  |
| Retention:                    | 0
                 | NULL                  |
| Location:                     | hdfs://localhost:20500/ec-warehouse/ect
                 | NULL                  |
| Erasure Coding Policy:        | RS-3-2-1024k
                 | NULL                  |
| Table Type:                   | EXTERNAL_TABLE
                 | NULL                  |
| Table Parameters:             | NULL
                 | NULL                  |
|                               | DO_NOT_UPDATE_STATS
                 | true                  |
|                               | EXTERNAL
                 | TRUE                  |
|                               | OBJCAPABILITIES
                 | EXTREAD,EXTWRITE      |
|                               | STATS_GENERATED
                 | TASK                  |
|                               | TRANSLATED_TO_EXTERNAL
                 | TRUE                  |
|                               | external.table.purge
                 | TRUE                  |
|                               | impala.lastComputeStatsTime
                 | 1668549634            |
|                               | numRows
                 | -1                    |
|                               | totalSize
                 | 27923                 |
|                               | transient_lastDdlTime
                 | 1668548361            |
|                               | NULL
                 | NULL                  |
| # Storage Information         | NULL
                 | NULL                  |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySi
mpleSerDe         | NULL                  |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat
                 | NULL                  |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyT
extOutputFormat | NULL                  |
| Compressed:                   | No
                 | NULL                  |
| Num Buckets:                  | 0
                 | NULL                  |
| Bucket Columns:               | []
                 | NULL                  |
| Sort Columns:                 | []
                 | NULL                  |
|                               | NULL
                 | NULL                  |
| # Constraints                 | NULL
                 | NULL                  |
+-------------------------------+----------------------------------------
------------------+----------------------+
Fetched 50 row(s) in 0.02s
```

Partition level

For partitioned tables, use the SHOW PARTITIONS statement to list out the EC policy of each partition directory. If you set up EC policy at the partition level, you will find the policy information under the new column EC policy.

```
show partitions part_text;
+-------+-------+-----+-------+--------+---------+--------------+--------
-----------+--------+--------------+-------------------+
| year  | month | day | #Rows | #Files | Size    | Bytes Cached | Cache Rep
lication | Format | EC Policy    | Incremental stats |
+-------+-------+-----+-------+--------+---------+--------------+--------
-----------+--------+--------------+-------------------+
| 2014  | 1     | 1   | -1    | 4      | 25.16MB | NOT CACHED   | NOT CACHED
         | TEXT   | RS-3-2-1024k | false             |
| 2014  | 1     | 2   | -1    | 4      | 26.22MB | NOT CACHED   | NOT CAC
HED          | TEXT   | RS-3-2-1024k | false             |
| 2014  | 1     | 3   | -1    | 0      | 0B      | NOT CACHED   | NOT CACHED
         | TEXT   | RS-3-2-1024k | false             |
| 2014  | 1     | 4   | -1    | 0      | 0B      | NOT CACHED   | NOT CAC
HED          | TEXT   | RS-3-2-1024k | false             |
| Total |       |     | -1    | 8      | 51.38MB | 0B           |
         |       |     |       |        |         |              |
+-------+-------+-----+-------+--------+---------+--------------+--------
-----------+--------+--------------+-------------------+
```

Understand Query Performance

Querying an erasure coded data might have an impact on the performance. If you notice performance degradation, you can troubleshoot it by running the PROFILE command. You can also use Cloudera Manager to quickly identify the cause.

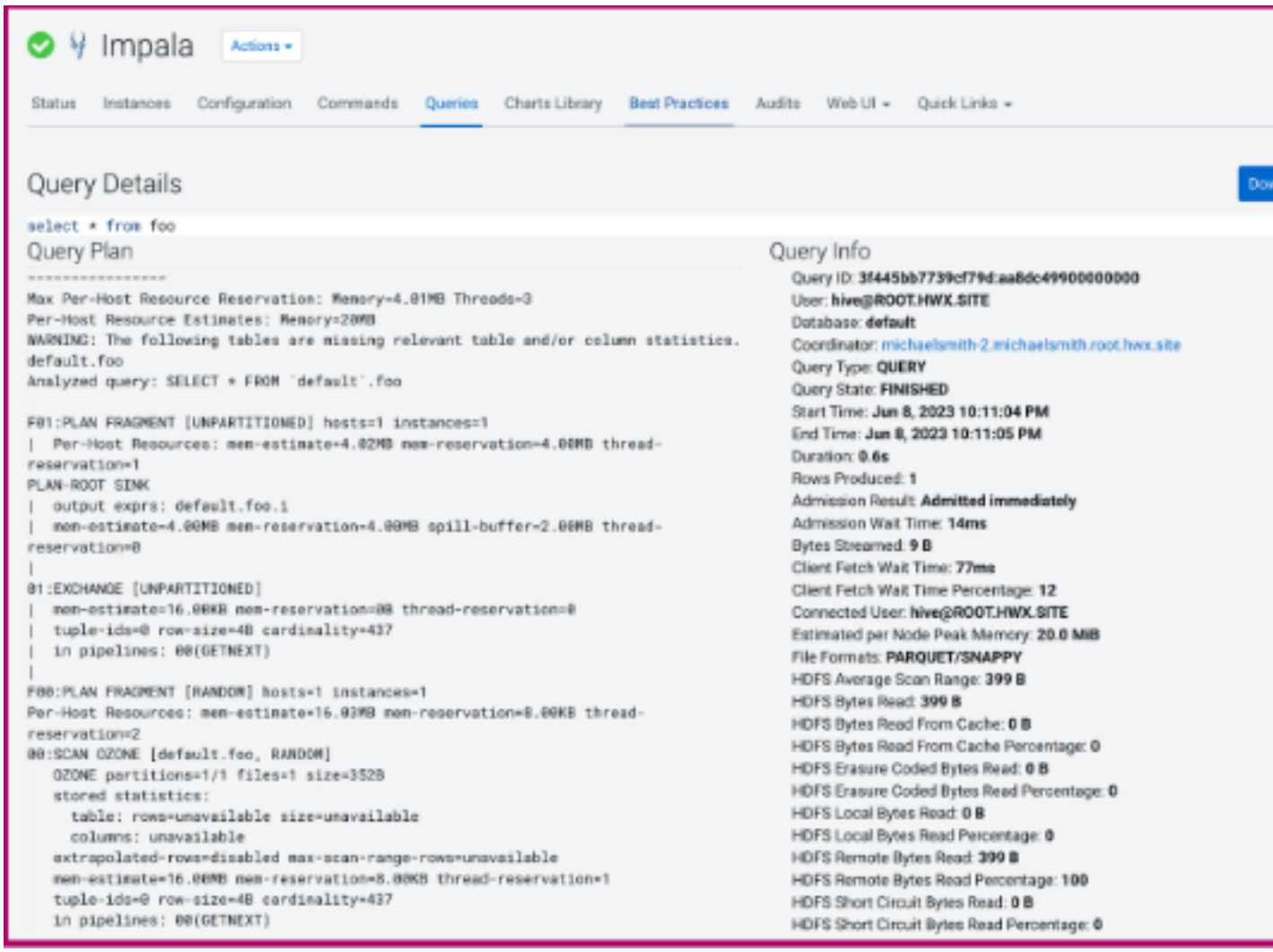
```
[localhost:21000] functional> select * from functional.alltypes;
…
[localhost:21000] functional> profile;
```

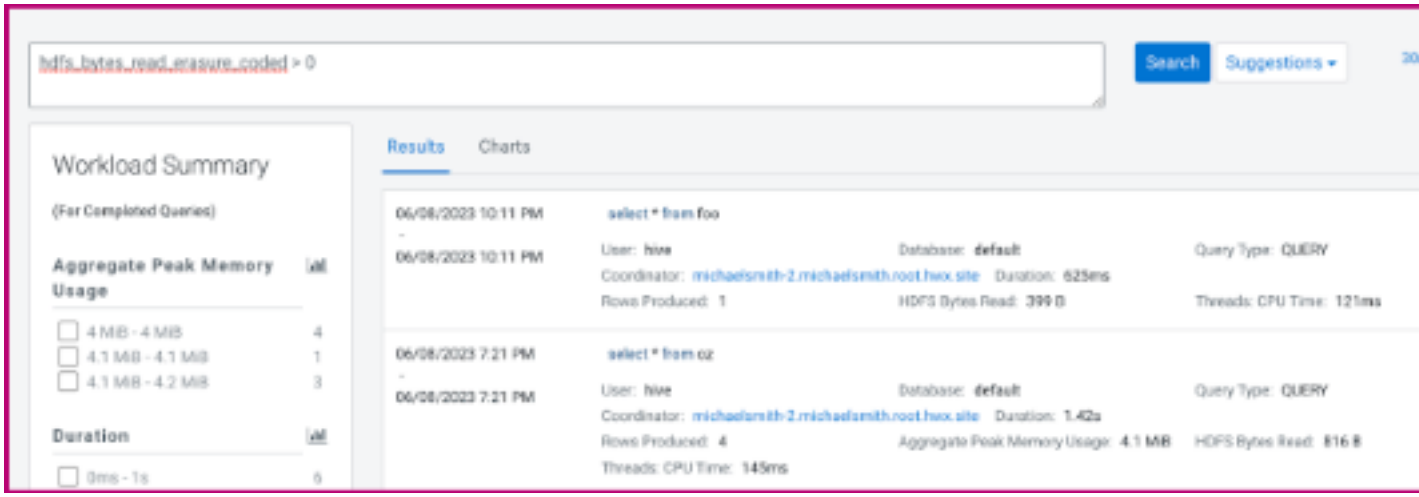
The PROFILE statement gives you an outline of the logical steps that a query will perform. It will include erasure coding related information in the profile as BytesReadErasureCoded. In the PROFILE output, look for the excerpts ExecSummary and Query Timeline, and then drill down into the particular fragment details to understand the query performance.

Monitor EC Metrics

Under Query Details in Cloudera Manager, you can check whether a query read erasure coded data via HDFS Erasure Coded Bytes Read.

You can search for queries that read erasure coded data with the hdfs_bytes_read_erasure_coded property.





Performance Considerations

The following sections explain the factors affecting the performance of Impala features, and procedures for tuning, monitoring, and benchmarking Impala queries and other SQL operations.

Recommended configurations

Following are the recommended configuration setting for the best performance with Impala.

- Set the --use_local_tz_for_unix_timestamp_conversions startup flag and the --convert_legacy_hive_parquet_utc_timestamps startup flag both to true. Setting these startup flags to true ensures that the timestamps between Hive and Impala match. See TIMESTAMP Data Type for more details.
- Always set the --idle_session_timeout and the --idle_query_timeout timeouts for the Impala daemon (impalad). Ensure that the setting for idle_session_timout is less than the setting for the timeout set for your load balancer. See Setting the Idle Query and Idle Session Timeouts for impalad for details.
- Set the --fe_service_threads startup option for the Impala daemon (impalad) to 256. This option specifies the maximum number of concurrent client connections allowed. See Startup Options for impalad Daemon for details.
- Increase the --num_metadata_loading_threads startup option to 64 to improve metadata loading performance. See Configuring Impala Startup Options through Cloudera Manager for more information.

Performance Best Practices

Use the performance guidelines and best practices during planning, experimentation, and performance tuning for an Impala-enabled cluster.

Choose the appropriate file format for the data

Typically, for large volumes of data (multiple gigabytes per table or partition), the Parquet file format performs best because of its combination of columnar storage layout, large I/O request size, and compression and encoding.



Note: For smaller volumes of data, a few gigabytes or less for each table or partition, you might not see significant performance differences between file formats. At small data volumes, reduced I/O from an efficient compressed file format can be counterbalanced by reduced opportunity for parallel execution. When planning for a production deployment or conducting benchmarks, always use realistic data volumes to get a true picture of performance and scalability.

Avoid data ingestion processes that produce many small files

When producing data files outside of Impala, prefer either text format or Avro, where you can build up the files row by row. Once the data is in Impala, you can convert it to the more efficient Parquet format and split into multiple data files using a single INSERT ... SELECT statement. Or, if you have the infrastructure to produce multi-megabyte Parquet files as part of your data preparation process, do that and skip the conversion step inside Impala.

Always use INSERT ... SELECT to copy significant volumes of data from table to table within Impala. Avoid INSERT ... VALUES for any substantial volume of data or performance-critical tables, because each such statement produces a separate tiny data file.

For example, if you have thousands of partitions in a Parquet table, each with less than 256 MB of data, consider partitioning in a less granular way, such as by year / month rather than year / month / day. If an inefficient data ingestion process produces thousands of data files in the same table or partition, consider compacting the data by performing an INSERT ... SELECT to copy all the data to a different table; the data will be reorganized into a smaller number of larger files by this process.

Choose partitioning granularity based on actual data volume

Partitioning is a technique that physically divides the data based on values of one or more columns, such as by year, month, day, region, city, section of a web site, and so on. When you issue queries that request a specific value or range of values for the partition key columns, Impala can avoid reading the irrelevant data, potentially yielding a huge savings in disk I/O.

When deciding which column(s) to use for partitioning, choose the right level of granularity. For example, should you partition by year, month, and day, or only by year and month? Choose a partitioning strategy that puts at least 256 MB of data in each partition, to take advantage of HDFS bulk I/O and Impala distributed queries.

Over-partitioning can also cause query planning to take longer than necessary, as Impala prunes the unnecessary partitions. Ideally, keep the number of partitions in the table under 30 thousand.

When preparing data files to go in a partition directory, create several large files rather than many small ones. If you receive data in the form of many small files and have no control over the input format, consider using the INSERT . .. SELECT syntax to copy data from one table or partition to another, which compacts the files into a relatively small number (based on the number of nodes in the cluster).

If you need to reduce the overall number of partitions and increase the amount of data in each partition, first look for partition key columns that are rarely referenced or are referenced in non-critical queries (not subject to an SLA). For example, your web site log data might be partitioned by year, month, day, and hour, but if most queries roll up the results by day, perhaps you only need to partition by year, month, and day.

If you need to reduce the granularity even more, consider creating “buckets”, computed values corresponding to different sets of partition key values. For example, you can use the TRUNC() function with a TIMESTAMP column to group date and time values based on intervals such as week or quarter.

Use smallest appropriate integer types for partition key columns

Although it is tempting to use strings for partition key columns, since those values are turned into HDFS directory names anyway, you can minimize memory usage by using numeric values for common partition key fields such as YEAR, MONTH, and DAY. Use the smallest integer type that holds the appropriate range of values, typically TINY INT for MONTH and DAY, and SMALLINT for YEAR. Use the EXTRACT() function to pull out individual date and time fields from a TIMESTAMP value, and CAST() the return value to the appropriate integer type.

Choose an appropriate Parquet block size

By default, the Impala INSERT ... SELECT statement creates Parquet files with a 256 MB block size. (This default was changed in Impala 2.0. Formerly, the limit was 1 GB, but Impala made conservative estimates about compression, resulting in files that were smaller than 1 GB.)

Each Parquet file written by Impala is a single block, allowing the whole file to be processed as a unit by a single host. As you copy Parquet files into HDFS or between HDFS filesystems, use hdfs dfs -pb to preserve the original block size.

If there is only one or a few data block in your Parquet table, or in a partition that is the only one accessed by a query, then you might experience a slowdown for a different reason: not enough data to take advantage of Impala's parallel distributed queries. Each data block is processed by a single core on one of the DataNodes. In a 100-node cluster of 16-core machines, you could potentially process thousands of data files simultaneously. You want to find a sweet spot between “many tiny files” and “single giant file” that balances bulk I/O and parallel processing. You can set the PARQUET_FILE_SIZE query option before doing an INSERT ... SELECT statement to reduce the size of each generated Parquet file. (Specify the file size as an absolute number of bytes, or in Impala 2.0 and later, in units ending with m for megabytes or g for gigabytes.) Run benchmarks with different file sizes to find the right balance point for your particular data volume.

Gather statistics for all tables used in performance-critical or high-volume join queries

Gather the statistics with the COMPUTE STATS statement.

Minimize the overhead of transmitting results back to the client

Use techniques such as:

- Aggregation. If you need to know how many rows match a condition, the total values of matching values from some column, the lowest or highest matching value, and so on, call aggregate functions such as COUNT(), SUM(), and MAX() in the query rather than sending the result set to an application and doing those computations

there. Remember that the size of an unaggregated result set could be huge, requiring substantial time to transmit across the network.
- Filtering. Use all applicable tests in the WHERE clause of a query to eliminate rows that are not relevant, rather than producing a big result set and filtering it using application logic.
- LIMIT clause. If you only need to see a few sample values from a result set, or the top or bottom values from a query using ORDER BY, include the LIMIT clause to reduce the size of the result set rather than asking for the full result set and then throwing most of the rows away.
- Avoid overhead from pretty-printing the result set and displaying it on the screen. When you retrieve the results through `impala-shell`, use `impala-shell` options such as -B and --output_delimiter to produce results without special formatting, and redirect output to a file rather than printing to the screen. Consider using INSE RT ... SELECT to write the results directly to new files in HDFS.

Verify that your queries are planned in an efficient logical manner

Examine the EXPLAIN plan for a query before actually running it.

Verify performance characteristics of queries

Verify that the low-level aspects of I/O, memory usage, network bandwidth, CPU utilization, and so on are within expected ranges by examining the query profile for a query after running it.

Hotspot analysis

In the context of Impala, a hotspot is defined as “an Impala daemon that for a single query or a workload is spending a far greater amount of time processing data relative to its neighbours”.

Before discussing the options to tackle this issue, some background is first required to understand how this problem can occur.

By default, the scheduling of scan based plan fragments is deterministic. This means that for multiple queries needing to read the same block of data, the same node will be picked to host the scan. The default scheduling logic does not take into account node workload from prior queries. The complexity of materializing a tuple depends on a few factors, namely: decoding and decompression. If the tuples are densely packed into data pages due to good encoding/ compression ratios, there will be more work required when reconstructing the data. Each compression codec offers different performance tradeoffs and should be considered before writing the data. Due to the deterministic nature of the scheduler, single nodes can become bottlenecks for highly concurrent queries that use the same tables.

If, for example, a Parquet based dataset is tiny, e.g. a small dimension table, such that it fits into a single HDFS block (Impala by default will create 256 MB blocks when Parquet is used, each containing a single row group) then there are a number of options that can be considered to resolve the potential scheduling hotspots when querying this data:

- Te scheduler’s deterministic behaviour can be changed using the following query options: REPLICA_PREFEREN CE and RANDOM_REPLICA.
- HDFS caching can be used to cache block replicas. This will cause the Impala scheduler to randomly pick a node that is hosting a cached block replica for the scan. Note, although HDFS caching has benefits, it serves only to help with the reading of raw block data and not cached tuple data, but with the right number of cached replicas (by default, HDFS only caches one replica), even load distribution can be achieved for smaller datasets.
- Do not compress the table data. The uncompressed table data spans more nodes and eliminates skew caused by compression.
- Reduce the Parquet file size via the PARQUET_FILE_SIZE query option when writing the table data. Using this approach the data will span more nodes. However it’s not recommended to drop the size below 32 MB.

Query Join Performance

Joins are the main class of queries that you can tune at the SQL level.

Queries involving join operations often require more tuning than queries that refer to only one table. The maximum size of the result set from a join query is the product of the number of rows in all the joined tables. When joining

several tables with millions or billions of rows, any missed opportunity to filter the result set, or other inefficiency in the query, could lead to an operation that does not finish in a practical time and has to be cancelled.

The simplest technique for tuning an Impala join query is to collect statistics on each table involved in the join using the COMPUTE STATS statement, and then to let Impala automatically optimize the query based on the size of each table, number of distinct values of each column, and so on. For accurate statistics about each table, issue the COMP UTE STATS statement after loading the data into that table, and again if the amount of data changes substantially due to operations, such as INSERT, LOAD DATA, or adding a partition.

If statistics are not available for all the tables in the join query, or if Impala chooses a join order that is not the most efficient, you can override the automatic join order optimization by specifying the STRAIGHT_JOIN keyword immediately after the SELECT and any DISTINCT or ALL keywords. In this case, Impala uses the order the tables appear in the query to guide how the joins are processed.

When you use the STRAIGHT_JOIN technique, you must order the tables in the join query manually instead of relying on the Impala optimizer. The optimizer uses sophisticated techniques to estimate the size of the result set at each stage of the join. For manual ordering, use this heuristic approach to start with, and then experiment to fine-tune the order:

- Specify the largest table first. This table is read from disk by each Impala node and so its size is not significant in terms of memory usage during the query.
- Next, specify the smallest table. The contents of the second, third, and so on tables are all transmitted across the network. You want to minimize the size of the result set from each subsequent stage of the join query. The most likely approach involves joining a small table first, so that the result set remains small even as subsequent larger tables are processed.
- Join the next smallest table, then the next smallest, and so on.

For example, if you had tables BIG, MEDIUM, SMALL, and TINY, the logical join order to try would be BIG, TINY, SMALL, MEDIUM.

The terms “largest” and “smallest” refers to the size of the intermediate result set based on the number of rows and columns from each table that are part of the result set. For example, if you join one table sales with another table cust omers, a query might find results from 100 different customers who made a total of 5000 purchases. In that case, you would specify SELECT ... FROM sales JOIN customers ..., putting customers on the right side because it is smaller in the context of this query.

The Impala query planner chooses between different techniques for performing join queries, depending on the absolute and relative sizes of the tables. Broadcast joins are the default, where the right-hand table is considered to be smaller than the left-hand table, and its contents are sent to all the other nodes involved in the query. The alternative technique is known as a partitioned join (not related to a partitioned table), which is more suitable for large tables of roughly equal size. With this technique, portions of each table are sent to appropriate other nodes where those subsets of rows can be processed in parallel. The choice of broadcast or partitioned join also depends on statistics being available for all tables in the join, gathered by the COMPUTE STATS statement.

To see which join strategy is used for a particular query, issue an EXPLAIN statement for the query. If you find that a query uses a broadcast join when you know through benchmarking that a partitioned join would be more efficient, or vice versa, add a hint to the query to specify the precise join mechanism to use.

How Joins Are Processed when Statistics Are Unavailable

If table or column statistics are not available for some tables in a join, Impala still reorders the tables using the information that is available. Tables with statistics are placed on the left side of the join order, in descending order of cost based on overall size and cardinality. Tables without statistics are treated as zero-size, that is, they are always placed on the right side of the join order.

Overriding Join Reordering with STRAIGHT_JOIN

If an Impala join query is inefficient because of outdated statistics or unexpected data distribution, you can keep Impala from reordering the joined tables by using the STRAIGHT_JOIN keyword immediately after the SELECT

and any DISTINCT or ALL keywords. The STRAIGHT_JOIN keyword turns off the reordering of join clauses that Impala does internally, and produces a plan that relies on the join clauses being ordered optimally in the query text.

Note: The STRAIGHT_JOIN hint affects the join order of table references in the query block containing the hint. It does not affect the join order of nested queries, such as views, inline views, or WHERE-clause subqueries. To use this hint for performance tuning of complex queries, apply the hint to all query blocks that need a fixed join order.

Table and Column Statistics

Impala can do better optimization for complex or multi-table queries when it has access to statistics about the volume of data and how the values are distributed. Impala uses this information to help parallelize and distribute the work for a query. For example, optimizing join queries requires a way of determining if one table is “bigger” than another, which is a function of the number of rows and the average row size for each table. This topic described the categories of statistics Impala can work with, and how to produce them and keep them up to date.

Overview of table statistics

The Impala query planner can make use of statistics about entire tables and partitions. This information includes physical characteristics such as the number of rows, number of data files, the total size of the data files, and the file format. For partitioned tables, the numbers are calculated per partition, and as totals for the whole table. This metadata is stored in the Metastore database, and can be updated by either Impala or Hive.

If a number is not available, the value -1 is used as a placeholder. Some numbers, such as number and total sizes of data files, are always kept up to date because they can be calculated cheaply, as part of gathering HDFS block metadata.

The following example shows table stats for an unpartitioned Parquet table. The values for the number and sizes of files are always available. Initially, the number of rows is not known, because it requires a potentially expensive scan through the entire table, and so that value is displayed as -1. The COMPUTE STATS statement fills in any unknown table stats values.

```
show table stats parquet_snappy;
+-------+--------+---------+--------------+-------------------+---------+-
------------------+...
| #Rows | #Files | Size    | Bytes Cached | Cache Replication | Format  |
 Incremental stats |...
+-------+--------+---------+--------------+-------------------+---------+---
----------------+...
| -1    | 96     | 23.35GB | NOT CACHED   | NOT CACHED        | PARQUET | f
alse            |...
+-------+--------+---------+--------------+-------------------+---------+-
------------------+...

compute stats parquet_snappy;
+-----------------------------------------+
| summary                                 |
+-----------------------------------------+
| Updated 1 partition(s) and 6 column(s). |
+-----------------------------------------+


show table stats parquet_snappy;
+------------+--------+---------+--------------+-------------------+-------
--+-------------------+...
| #Rows      | #Files | Size    | Bytes Cached | Cache Replication | Format
  | Incremental stats |...
+------------+--------+---------+--------------+-------------------+-------
--+-------------------+...
```

```
| 1000000000 | 96     | 23.35GB | NOT CACHED   | NOT CACHED        | PARQ
UET | false             |...
+------------+--------+---------+--------------+-------------------+-------
--+-------------------+...
```

To check that table statistics are available for a table, and see the details of those statistics, use the statement SHOW TABLE STATS *TABLE_NAME*.

Overview of Column Statistics

The Impala query planner can make use of statistics about individual columns when that metadata is available in the metastore database. This technique is most valuable for columns compared across tables in join queries, to help estimate how many rows the query will retrieve from each table. These statistics are also important for correlated subqueries using the EXISTS or IN operators, which are processed internally the same way as join queries.

The following example shows column stats for an unpartitioned Parquet table. The values for the maximum and average sizes of some types are always available, because those figures are constant for numeric and other fixed-size types. Initially, the number of distinct values is not known, because it requires a potentially expensive scan through the entire table, and so that value is displayed as -1. The same applies to maximum and average sizes of variable-sized types, such as STRING. The COMPUTE STATS statement fills in most unknown column stats values. (It does not record the number of NULL values, because currently Impala does not use that figure for query optimization.)

```
show column stats parquet_snappy;
+-------------+----------+------------------+--------+----------+----------+
| Column      | Type     | #Distinct Values | #Nulls | Max Size | Avg Size |
+-------------+----------+------------------+--------+----------+----------+
| id          | BIGINT   | -1               | -1     | 8        | 8        |
| val         | INT      | -1               | -1     | 4        | 4        |
| zerofill    | STRING   | -1               | -1     | -1       | -1       |
| name        | STRING   | -1               | -1     | -1       | -1       |
| assertion   | BOOLEAN  | -1               | -1     | 1        | 1        |
| location_id | SMALLINT | -1               | -1     | 2        | 2        |
+-------------+----------+------------------+--------+----------+----------+

compute stats parquet_snappy;
+-----------------------------------------+
| summary                                 |
+-----------------------------------------+
| Updated 1 partition(s) and 6 column(s). |
+-----------------------------------------+
show column stats parquet_snappy;
+-------------+----------+------------------+--------+----------+--------
-----------+
| Column      | Type     | #Distinct Values | #Nulls | Max Size | Avg Size
        |
+-------------+----------+------------------+--------+----------+--------
-----------+
| id          | BIGINT   | 183861280        | -1     | 8        | 8
        |
| val         | INT      | 139017           | -1     | 4        | 4
         |
| zerofill    | STRING   | 101761           | -1     | 6        | 6
        |
| name        | STRING   | 145636240        | -1     | 22       | 13.0002
0027160645 |
| assertion   | BOOLEAN  | 2                | -1     | 1        | 1
        |
| location_id | SMALLINT | 339              | -1     | 2        | 2
         |
```

```
+-------------+----------+------------------+--------+----------+----------
---------+
```

Note:

For column statistics to be effective in Impala, you also need to have table statistics for the applicable tables. When you use the Impala COMPUTE STATS statement, both table and column statistics are automatically gathered at the same time, for all columns in the table.

Note: Because Impala does not currently use the NULL count during query planning, Impala speeds up the COMPUTE STATS statement by skipping this NULL counting.

To check whether column statistics are available for a particular set of columns, use the SHOW COLUMN STATS *TABLE_NAME* statement, or check the extended EXPLAIN output for a query against that table that refers to those columns.

How Table and Column Statistics Work for Partitioned Tables

When you use Impala for “big data”, you are highly likely to use partitioning for your biggest tables, the ones representing data that can be logically divided based on dates, geographic regions, or similar criteria. The table and column statistics are especially useful for optimizing queries on such tables. For example, a query involving one year might involve substantially more or less data than a query involving a different year, or a range of several years. Each query might be optimized differently as a result.

The following examples show how table and column stats work with a partitioned table. The table for this example is partitioned by year, month, and day. For simplicity, the sample data consists of 5 partitions, all from the same year and month. Table stats are collected independently for each partition. (In fact, the SHOW PARTITIONS statement displays exactly the same information as SHOW TABLE STATS for a partitioned table.) Column stats apply to the entire table, not to individual partitions. Because the partition key column values are represented as HDFS directories, their characteristics are typically known in advance, even when the values for non-key columns are shown as -1.

```
show partitions year_month_day;
+-------+-------+-----+-------+--------+---------+--------------+----------
---------+---------+...
| year  | month | day | #Rows | #Files | Size    | Bytes Cached | Cache Repl
ication | Format  |...
+-------+-------+-----+-------+--------+---------+--------------+--------
-----------+---------+...
| 2013  | 12    | 1   | -1    | 1      | 2.51MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| 2013  | 12    | 2   | -1    | 1      | 2.53MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| 2013  | 12    | 3   | -1    | 1      | 2.52MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| 2013  | 12    | 4   | -1    | 1      | 2.51MB  | NOT CACHED   | NOT CAC
HED         | PARQUET |...
| 2013  | 12    | 5   | -1    | 1      | 2.52MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| Total |       |     | -1    | 5      | 12.58MB | 0B           |
        |         |...
+-------+-------+-----+-------+--------+---------+--------------+-----------
--------+---------+...

show table stats year_month_day;
+-------+-------+-----+-------+--------+---------+--------------+----------
---------+---------+...
| year  | month | day | #Rows | #Files | Size    | Bytes Cached | Cache Repl
ication | Format  |...
+-------+-------+-----+-------+--------+---------+--------------+--------
-----------+---------+...
```

```
| 2013  | 12    | 1   | -1    | 1      | 2.51MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| 2013  | 12    | 2   | -1    | 1      | 2.53MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| 2013  | 12    | 3   | -1    | 1      | 2.52MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| 2013  | 12    | 4   | -1    | 1      | 2.51MB  | NOT CACHED   | NOT CAC
HED       | PARQUET |...
| 2013  | 12    | 5   | -1    | 1      | 2.52MB  | NOT CACHED   | NOT CACHED
        | PARQUET |...
| Total |       |     | -1    | 5      | 12.58MB | 0B           |
        |       |...
+-------+-------+-----+-------+--------+---------+--------------+-----------
--------+---------+...

show column stats year_month_day;
+-----------+---------+------------------+--------+----------+----------+
| Column    | Type    | #Distinct Values | #Nulls | Max Size | Avg Size |
+-----------+---------+------------------+--------+----------+----------+
| id        | INT     | -1               | -1     | 4        | 4        |
| val       | INT     | -1               | -1     | 4        | 4        |
| zfill     | STRING  | -1               | -1     | -1       | -1       |
| name      | STRING  | -1               | -1     | -1       | -1       |
| assertion | BOOLEAN | -1               | -1     | 1        | 1        |
| year      | INT     | 1                | 0      | 4        | 4        |
| month     | INT     | 1                | 0      | 4        | 4        |
| day       | INT     | 5                | 0      | 4        | 4        |
+-----------+---------+------------------+--------+----------+----------+

compute stats year_month_day;
+-----------------------------------------+
| summary                                 |
+-----------------------------------------+
| Updated 5 partition(s) and 5 column(s). |
+-----------------------------------------+

show table stats year_month_day;
+-------+-------+-----+--------+--------+---------+--------------+-------
------------+---------+...
| year  | month | day | #Rows  | #Files | Size    | Bytes Cached | Cache
Replication | Format  |...
+-------+-------+-----+--------+--------+---------+--------------+-------
------------+---------+...
| 2013  | 12    | 1   | 93606  | 1      | 2.51MB  | NOT CACHED   | NOT CA
CHED        | PARQUET |...
| 2013  | 12    | 2   | 94158  | 1      | 2.53MB  | NOT CACHED   | NOT CA
CHED        | PARQUET |...
| 2013  | 12    | 3   | 94122  | 1      | 2.52MB  | NOT CACHED   | NOT CA
CHED        | PARQUET |...
| 2013  | 12    | 4   | 93559  | 1      | 2.51MB  | NOT CACHED   | NOT CA
CHED        | PARQUET |...
| 2013  | 12    | 5   | 93845  | 1      | 2.52MB  | NOT CACHED   | NOT CA
CHED        | PARQUET |...
| Total |       |     | 469290 | 5      | 12.58MB | 0B           |
          |         |...
+-------+-------+-----+--------+--------+---------+--------------+-------
------------+---------+...

show column stats year_month_day;
+-----------+---------+------------------+--------+----------+------------
-------+
| Column    | Type    | #Distinct Values | #Nulls | Max Size | Avg Size
     |
```

```
+-----------+---------+------------------+--------+----------+-------------
-----+
| id        | INT     | 511129           | -1     | 4        | 4
     |
| val       | INT     | 364853           | -1     | 4        | 4
     |
| zfill     | STRING  | 311430           | -1     | 6        | 6
     |
| name      | STRING  | 471975           | -1     | 22       | 13.0016002655
0293 |
| assertion | BOOLEAN | 2                | -1     | 1        | 1
     |
| year      | INT     | 1                | 0      | 4        | 4
     |
| month     | INT     | 1                | 0      | 4        | 4
     |
| day       | INT     | 5                | 0      | 4        | 4
     |
+-----------+---------+------------------+--------+----------+-----------
--------+
```

If you run the Hive statement ANALYZE TABLE COMPUTE STATISTICS FOR COLUMNS, Impala can only use the resulting column statistics if the table is unpartitioned. Impala cannot use Hive-generated column statistics for a partitioned table.

Detecting Missing Statistics

You can check whether a specific table has statistics using the SHOW TABLE STATS statement (for any table) or the SHOW PARTITIONS statement (for a partitioned table). Both statements display the same information. If a table or a partition does not have any statistics, the #Rows field contains -1. Once you compute statistics for the table or partition, the #Rows field changes to an accurate value.

The following example shows a table that initially does not have any statistics. The SHOW TABLE STATS statement displays different values for #Rows before and after the COMPUTE STATS operation.

```
[localhost:21000] > create table no_stats (x int);
[localhost:21000] > show table stats no_stats;
+-------+--------+------+--------------+--------+-------------------+
| #Rows | #Files | Size | Bytes Cached | Format | Incremental stats |
+-------+--------+------+--------------+--------+-------------------+
| -1    | 0      | 0B   | NOT CACHED   | TEXT   | false             |
+-------+--------+------+--------------+--------+-------------------+
[localhost:21000] > compute stats no_stats;
+-----------------------------------------+
| summary                                 |
+-----------------------------------------+
| Updated 1 partition(s) and 1 column(s). |
+-----------------------------------------+
[localhost:21000] > show table stats no_stats;
+-------+--------+------+--------------+--------+-------------------+
| #Rows | #Files | Size | Bytes Cached | Format | Incremental stats |
+-------+--------+------+--------------+--------+-------------------+
| 0     | 0      | 0B   | NOT CACHED   | TEXT   | false             |
+-------+--------+------+--------------+--------+-------------------+
```

The following example shows a similar progression with a partitioned table. Initially, #Rows is -1. After a COMP UTE STATS operation, #Rows changes to an accurate value. Any newly added partition starts with no statistics, meaning that you must collect statistics after adding a new partition.

```
[localhost:21000] > create table no_stats_partitioned (x int) partitioned by
 (year smallint);
[localhost:21000] > show table stats no_stats_partitioned;
```

```
+-------+-------+--------+------+--------------+--------+-------------------
+
| year  | #Rows | #Files | Size | Bytes Cached | Format | Incremental stats
 |
+-------+-------+--------+------+--------------+--------+-------------------
+
| Total | -1    | 0      | 0B   | 0B           |        |
 |
+-------+-------+--------+------+--------------+--------+-------------------
+
[localhost:21000] > show partitions no_stats_partitioned;
+-------+-------+--------+------+--------------+--------+-------------------
+
| year  | #Rows | #Files | Size | Bytes Cached | Format | Incremental stats
 |
+-------+-------+--------+------+--------------+--------+-------------------
+
| Total | -1    | 0      | 0B   | 0B           |        |
 |
+-------+-------+--------+------+--------------+--------+-------------------
+
[localhost:21000] > alter table no_stats_partitioned add partition (year=
2013);
[localhost:21000] > compute stats no_stats_partitioned;
+-----------------------------------------+
| summary                                 |
+-----------------------------------------+
| Updated 1 partition(s) and 1 column(s). |
+-----------------------------------------+
[localhost:21000] > alter table no_stats_partitioned add partition (year=
2014);
[localhost:21000] > show partitions no_stats_partitioned;
+-------+-------+--------+------+--------------+--------+-------------------
+
| year  | #Rows | #Files | Size | Bytes Cached | Format | Incremental stats
 |
+-------+-------+--------+------+--------------+--------+-------------------
+
| 2013  | 0     | 0      | 0B   | NOT CACHED   | TEXT   | false
 |
| 2014  | -1    | 0      | 0B   | NOT CACHED   | TEXT   | false
 |
| Total | 0     | 0      | 0B   | 0B           |        |
 |
+-------+-------+--------+------+--------------+--------+-------------------
+
```



Note: Because the default COMPUTE STATS statement creates and updates statistics for all partitions in a table, if you expect to frequently add new partitions, use the COMPUTE INCREMENTAL STATS syntax instead, which lets you compute stats for a single specified partition, or only for those partitions that do not already have incremental stats.

If checking each individual table is impractical, due to a large number of tables or views that hide the underlying base tables, you can also check for missing statistics for a particular query. Use the EXPLAIN statement to preview query efficiency before actually running the query. Use the query profile output available through the PROFILE command in `impala-shell` or the web UI to verify query execution and timing after running the query. Both the EXPLAIN plan and the PROFILE output display a warning if any tables or partitions involved in the query do not have statistics.

```
[localhost:21000] > create table no_stats (x int);
[localhost:21000] > explain select count(*) from no_stats;
+-----------------------------------------------------------------------------
-----------+
```

```
| Explain String
          |
+-------------------------------------------------------------------------
---------+
| Estimated Per-Host Requirements: Memory=10.00MB VCores=1
           |
| WARNING: The following tables are missing relevant table and/or column s
tatistics. |
| incremental_stats.no_stats
          |
|
          |
| 03:AGGREGATE [FINALIZE]
           |
| |  output: count:merge(*)
          |
| |
          |
| 02:EXCHANGE [UNPARTITIONED]
          |
| |
           |
| 01:AGGREGATE
          |
| |  output: count(*)
          |
| |
          |
| 00:SCAN HDFS [incremental_stats.no_stats]
           |
|    partitions=1/1 files=0 size=0B
          |
+-------------------------------------------------------------------------
----------+
```

Because Impala uses the *partition pruning* technique when possible to only evaluate certain partitions, if you have a partitioned table with statistics for some partitions and not others, whether or not the EXPLAIN statement shows the warning depends on the actual partitions used by the query. For example, you might see warnings or not for different queries against the same table:

```
-- No warning because all the partitions for the year 2012 have stats.
EXPLAIN SELECT ... FROM t1 WHERE year = 2012;

-- Missing stats warning because one or more partitions in this range
-- do not have stats.
EXPLAIN SELECT ... FROM t1 WHERE year BETWEEN 2006 AND 2009;
```

To confirm if any partitions at all in the table are missing statistics, you might explain a query that scans the entire table, such as SELECT COUNT(*) FROM *TABLE_NAME*.

Manually Setting Table Statistics with ALTER TABLE

The most crucial piece of data in all the statistics is the number of rows in the table (for an unpartitioned or partitioned table) and for each partition (for a partitioned table). The COMPUTE STATS statement always gathers statistics about all columns, as well as overall table statistics. If it is not practical to do a full COMPUTE STATS or COMPUTE INCREMENTAL STATS operation after adding a partition or inserting data, or if you can see that Impala would produce a more efficient plan if the number of rows was different, you can manually set the number of rows through an ALTER TABLE statement:

```
```

```
-- Set total number of rows. Applies to both unpartitioned and partitioned
tables.
alter table TABLE_NAME set tblproperties('numRows'='NEW_VALUE', 'STATS_GE
NERATED_VIA_STATS_TASK'='true');

-- Set total number of rows for a specific partition. Applies to partitioned
 tables only.
-- You must specify all the partition key columns in the PARTITION clause.
alter table TABLE_NAME partition (KEYCOL1=VAL1,KEYCOL2=VAL2...) set tblprop
erties('numRows'='NEW_VALUE', 'STATS_GENERATED_VIA_STATS_TASK'='true');
```

This statement avoids re-scanning any data files. For example:

```
alter table analysis_data set tblproperties('numRows'='1001000000', 'STATS_
GENERATED_VIA_STATS_TASK'='true');
```

For a partitioned table, update both the per-partition number of rows and the number of rows for the whole table. For example:

```
alter table partitioned_data partition(year=2009, month=4) set tblproperties
 ('numRows'='30000', 'STATS_GENERATED_VIA_STATS_TASK'='true');
alter table partitioned_data set tblproperties ('numRows'='1030000', 'STA
TS_GENERATED_VIA_STATS_TASK'='true');
```

In practice, the COMPUTE STATS statement, or COMPUTE INCREMENTAL STATS for a partitioned table, should be fast and convenient enough that this technique is only useful for the very largest partitioned tables. Because the column statistics might be left in a stale state, do not use this technique as a replacement for COMPUTE STATS. Only use this technique if all other means of collecting statistics are impractical, or as a low-overhead operation that you run in between periodic COMPUTE STATS or COMPUTE INCREMENTAL STATS operations.

Manually Setting Column Statistics with ALTER TABLE

You can use the SET COLUMN STATS clause of ALTER TABLE to manually set or change column statistics. Only use this technique in cases where it is impractical to run COMPUTE STATS or COMPUTE INCREMENTAL ST ATS frequently enough to keep up with data changes for a huge table.

You specify a case-insensitive symbolic name for the kind of statistics: numDVs, numNulls, avgSize, maxSize. The key names and values are both quoted. This operation applies to an entire table, not a specific partition.

For example:

```
alter table t1 set column stats x ('numDVs'='2','numNulls'='0');
```

Examples of Using Table and Column Statistics with Impala

The following examples walk through a sequence of SHOW TABLE STATS, SHOW COLUMN STATS, ALTE R TABLE, and SELECT and INSERT statements to illustrate various aspects of how Impala uses statistics to help optimize queries.

This example shows table and column statistics for the STORE column used in the TPC-DS benchmarks for decision support systems. It is a tiny table holding data for 12 stores. Initially, before any statistics are gathered by a COMP UTE STATS statement, most of the numeric fields show placeholder values of -1, indicating that the figures are unknown. The figures that are filled in are values that are easily countable or deducible at the physical level, such as the number of files, total data size of the files, and the maximum and average sizes for data types that have a constant size such as INT, FLOAT, and TIMESTAMP.

```
[localhost:21000] > show table stats store;
+-------+--------+--------+--------+
```

```
| #Rows | #Files | Size   | Format |
+-------+--------+--------+--------+
| -1    | 1      | 3.08KB | TEXT   |
+-------+--------+--------+--------+
Returned 1 row(s) in 0.03s
[localhost:21000] > show column stats store;
+--------------------+-----------+------------------+--------+----------+--
--------+
| Column             | Type      | #Distinct Values | #Nulls | Max Size |
Avg Size |
+--------------------+-----------+------------------+--------+----------+
----------+
| s_store_sk         | INT       | -1               | -1     | 4        | 4
         |
| s_store_id         | STRING    | -1               | -1     | -1       | -1
         |
| s_rec_start_date   | TIMESTAMP | -1               | -1     | 16       | 16
         |
| s_rec_end_date     | TIMESTAMP | -1               | -1     | 16       |
 16        |
| s_closed_date_sk   | INT       | -1               | -1     | 4        | 4
         |
| s_store_name       | STRING    | -1               | -1     | -1       | -1
         |
| s_number_employees | INT       | -1               | -1     | 4        | 4
         |
| s_floor_space      | INT       | -1               | -1     | 4        | 4
          |
| s_hours            | STRING    | -1               | -1     | -1       | -1
         |
| s_manager          | STRING    | -1               | -1     | -1       | -1
         |
| s_market_id        | INT       | -1               | -1     | 4        | 4
         |
| s_geography_class  | STRING    | -1               | -1     | -1       |
 -1        |
| s_market_desc      | STRING    | -1               | -1     | -1       | -1
         |
| s_market_manager   | STRING    | -1               | -1     | -1       | -1
         |
| s_division_id      | INT       | -1               | -1     | 4        | 4
         |
| s_division_name    | STRING    | -1               | -1     | -1       |
 -1        |
| s_company_id       | INT       | -1               | -1     | 4        | 4
         |
| s_company_name     | STRING    | -1               | -1     | -1       | -1
         |
| s_street_number    | STRING    | -1               | -1     | -1       | -1
         |
| s_street_name      | STRING    | -1               | -1     | -1       |
 -1        |
| s_street_type      | STRING    | -1               | -1     | -1       | -1
         |
| s_suite_number     | STRING    | -1               | -1     | -1       | -1
         |
| s_city             | STRING    | -1               | -1     | -1       | -1
         |
| s_county           | STRING    | -1               | -1     | -1       |
 -1        |
| s_state            | STRING    | -1               | -1     | -1       | -1
         |
| s_zip              | STRING    | -1               | -1     | -1       | -1
         |
```

```
| s_country          | STRING    | -1               | -1     | -1       | -1
      |
| s_gmt_offset       | FLOAT     | -1               | -1     | 4        | 4
      |
| s_tax_percentage   | FLOAT     | -1               | -1     | 4        | 4
      |
+--------------------+-----------+------------------+--------+----------+--
--------+
Returned 29 row(s) in 0.04s
```

With the Hive ANALYZE TABLE statement for column statistics, you had to specify each column for which to gather statistics. The Impala COMPUTE STATS statement automatically gathers statistics for all columns, because it reads through the entire table relatively quickly and can efficiently compute the values for all the columns. This example shows how after running the COMPUTE STATS statement, statistics are filled in for both the table and all its columns:

```
[localhost:21000] > compute stats store;
+------------------------------------------+
| summary                                  |
+------------------------------------------+
| Updated 1 partition(s) and 29 column(s). |
+------------------------------------------+
Returned 1 row(s) in 1.88s
[localhost:21000] > show table stats store;
+-------+--------+--------+--------+
| #Rows | #Files | Size   | Format |
+-------+--------+--------+--------+
| 12    | 1      | 3.08KB | TEXT   |
+-------+--------+--------+--------+
Returned 1 row(s) in 0.02s
[localhost:21000] > show column stats store;
+--------------------+-----------+------------------+--------+----------+-
------------------+
| Column             | Type      | #Distinct Values | #Nulls | Max Size | Av
g Size          |
+--------------------+-----------+------------------+--------+----------+-
------------------+
| s_store_sk         | INT       | 12               | -1     | 4        | 4
                |
| s_store_id         | STRING    | 6                | -1     | 16       | 16
                |
| s_rec_start_date   | TIMESTAMP | 4                | -1     | 16       | 16
                |
| s_rec_end_date     | TIMESTAMP | 3                | -1     | 16       | 16
                |
| s_closed_date_sk   | INT       | 3                | -1     | 4        | 4
                |
| s_store_name       | STRING    | 8                | -1     | 5        |
4.25                |
| s_number_employees | INT       | 9                | -1     | 4        | 4
                |
| s_floor_space      | INT       | 10               | -1     | 4        | 4
                |
| s_hours            | STRING    | 2                | -1     | 8        | 7.
083300113677979 |
| s_manager          | STRING    | 7                | -1     | 15       | 12
                |
| s_market_id        | INT       | 7                | -1     | 4        | 4
                |
| s_geography_class  | STRING    | 1                | -1     | 7        | 7
                |
| s_market_desc      | STRING    | 10               | -1     | 94       | 55
.5              |
```

```
| s_market_manager   | STRING    | 7                | -1     | 16        | 14
              |
| s_division_id      | INT       | 1                | -1     | 4         | 4
              |
| s_division_name    | STRING    | 1                | -1     | 7         | 7
              |
| s_company_id       | INT       | 1                | -1     | 4         | 4
              |
| s_company_name     | STRING    | 1                | -1     | 7         | 7
              |
| s_street_number    | STRING    | 9                | -1     | 3         | 2.
833300113677979 |
| s_street_name      | STRING    | 12               | -1     | 11        |
6.583300113677979 |
| s_street_type      | STRING    | 8                | -1     | 9         | 4.
833300113677979 |
| s_suite_number     | STRING    | 11               | -1     | 9         |
8.25               |
| s_city             | STRING    | 2                | -1     | 8         | 6.
5             |
| s_county           | STRING    | 1                | -1     | 17        | 17
              |
| s_state            | STRING    | 1                | -1     | 2         | 2
              |
| s_zip              | STRING    | 2                | -1     | 5         | 5
              |
| s_country          | STRING    | 1                | -1     | 13        | 13
              |
| s_gmt_offset       | FLOAT     | 1                | -1     | 4         | 4
              |
| s_tax_percentage   | FLOAT     | 5                | -1     | 4         | 4
              |
+--------------------+-----------+------------------+--------+-----------+-
------------------+
Returned 29 row(s) in 0.04s
```

The following example shows how statistics are represented for a partitioned table. In this case, we have set up a table to hold the world's most trivial census data, a single STRING field, partitioned by a YEAR column. The table statistics include a separate entry for each partition, plus final totals for the numeric fields. The column statistics include some easily deducible facts for the partitioning column, such as the number of distinct values (the number of partition subdirectories).

```
localhost:21000] > describe census;
+------+----------+---------+
| name | type     | comment |
+------+----------+---------+
| name | string   |         |
| year | smallint |         |
+------+----------+---------+
Returned 2 row(s) in 0.02s
[localhost:21000] > show table stats census;
+-------+-------+--------+------+---------+
| year  | #Rows | #Files | Size | Format  |
+-------+-------+--------+------+---------+
| 2000  | -1    | 0      | 0B   | TEXT    |
| 2004  | -1    | 0      | 0B   | TEXT    |
| 2008  | -1    | 0      | 0B   | TEXT    |
| 2010  | -1    | 0      | 0B   | TEXT    |
| 2011  | 0     | 1      | 22B  | TEXT    |
| 2012  | -1    | 1      | 22B  | TEXT    |
| 2013  | -1    | 1      | 231B | PARQUET |
| Total | 0     | 3      | 275B |         |
+-------+-------+--------+------+---------+
```

```
Returned 8 row(s) in 0.02s
[localhost:21000] > show column stats census;
+--------+----------+------------------+--------+----------+----------+
| Column | Type     | #Distinct Values | #Nulls | Max Size | Avg Size |
+--------+----------+------------------+--------+----------+----------+
| name   | STRING   | -1               | -1     | -1       | -1       |
| year   | SMALLINT | 7                | -1     | 2        | 2        |
+--------+----------+------------------+--------+----------+----------+
Returned 2 row(s) in 0.02s
```

The following example shows how the statistics are filled in by a COMPUTE STATS statement in Impala.

```
[localhost:21000] > compute stats census;
+-----------------------------------------+
| summary                                 |
+-----------------------------------------+
| Updated 3 partition(s) and 1 column(s). |
+-----------------------------------------+
Returned 1 row(s) in 2.16s
[localhost:21000] > show table stats census;
+-------+-------+--------+------+---------+
| year  | #Rows | #Files | Size | Format  |
+-------+-------+--------+------+---------+
| 2000  | -1    | 0      | 0B   | TEXT    |
| 2004  | -1    | 0      | 0B   | TEXT    |
| 2008  | -1    | 0      | 0B   | TEXT    |
| 2010  | -1    | 0      | 0B   | TEXT    |
| 2011  | 4     | 1      | 22B  | TEXT    |
| 2012  | 4     | 1      | 22B  | TEXT    |
| 2013  | 1     | 1      | 231B | PARQUET |
| Total | 9     | 3      | 275B |         |
+-------+-------+--------+------+---------+
Returned 8 row(s) in 0.02s
[localhost:21000] > show column stats census;
+--------+----------+------------------+--------+----------+----------+
| Column | Type     | #Distinct Values | #Nulls | Max Size | Avg Size |
+--------+----------+------------------+--------+----------+----------+
| name   | STRING   | 4                | -1     | 5        | 4.5      |
| year   | SMALLINT | 7                | -1     | 2        | 2        |
+--------+----------+------------------+--------+----------+----------+
Returned 2 row(s) in 0.02s
```

You can see how Impala executes a query differently in each case by observing the EXPLAIN output before and after collecting statistics. Measure the before and after query times, and examine the throughput numbers in before and after SUMMARY or PROFILE output, to verify how much the improved plan speeds up performance.

History based optimization [Preview]

History-based optimization uses runtime statistics from previous query executions to improve execution plans for similar queries without requiring table statistics.

Overview

History Based Optimization (HBO) accelerates query processing by reusing historical runtime statistics collected from previous query runs. Impala collects runtime statistics during query processing and stores them in the HBO cache. By default, the in-memory cache is maintained per coordinator and is lost when the coordinator restarts. During subsequent query planning, the planner reads the stored cardinality statistics from the HBO cache and applies them to generate an optimized query plan.

Note: This feature is in Technical Preview and is not ready for production deployments. recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

Observability updates

You can verify whether a query uses HBO statistics by reviewing the query profile or EXPLAIN output. If cardinality statistics are read from the HBO cache, the EXPLAIN output or the query plan in the profile displays the "from HBO" annotation. Extended profile updates provide additional details regarding the applied HBO statistics

EXPLAIN output examples

When HBO is enabled and statistics are found in the cache, the EXPLAIN output shows "from HBO" annotations in the cardinality estimates. The following example shows a join query where HBO provides cardinality for both scan nodes and the join operation:

```
PLAN-ROOT SINK
|
02:HASH JOIN [INNER JOIN, BROADCAST]
|  hash predicates: b.id = a.id
|  row-size=29B cardinality=31 (from HBO)
|
|--04:EXCHANGE [BROADCAST]
|  |
|  00:SCAN HDFS [functional.alltypes a]
|     partition predicates: a.`year` = 2010, a.`month` = 1
|     row-size=8B cardinality=31 (from HBO)
|
01:SCAN HDFS [functional.alltypes b]
   partition predicates: b.`year` = 2010
   row-size=21B cardinality=365 (from HBO)
```

In this example, HBO provides cardinality estimates of 31 rows for the first scan and 365 rows for the second scan.

Enabling history based optimization

Configure History Based Optimization query options in Impala to store runtime statistics and apply historical cardinality data during query planning.

Prerequisites: You must have access to the Impala shell or a supported query editor.

- Enable the STORE_HBO_STATS query option to populate the HBO cache with runtime query statistics. The statistics are stored asynchronously after the query finishes, so enabling this option does not add significant query latency: SET STORE_HBO_STATS=true;
- Enable the USE_HBO_STATS query option to instruct the planner to read and apply stored statistics from the HBO cache: SET USE_HBO_STATS=true;

History based optimization flags

History Based Optimization startup flags specify the similarity threshold, maximum stored runs, and memory cache size for the Impala daemon.

The following table describes the History Based Optimization startup flags for the Impala daemon (impalad):

Table 1: HBO flags

Flag	Default Value	Description
hbo_similarity_threshold	0.1 (10%)	Specifies the similarity threshold value used to match queries against historical entries in the HBO cache. The threshold compares scan input statistics, and hash keys identify plan-node subtrees rather than whole queries.
hbo_max_runs_per_key	100	Specifies the maximum number of historical runs stored per hash key in the HBO cache. When this limit is exceeded, the oldest run is evicted.
hbo_in_memory_backend_cache_size_bytes	1 GB	Specifies the maximum memory size, in bytes, allocated for the in-memory backend HBO cache.

Runtime Filtering

Runtime filtering is a wide-ranging optimization feature available in Impala. When only a fraction of the data in a table is needed for a query against a partitioned table or to evaluate a join condition, Impala determines the appropriate conditions while the query is running, and broadcasts that information to all the `impalad` nodes that are reading the table so that they can avoid unnecessary I/O to read partition data, and avoid unnecessary network transmission by sending only the subset of rows that match the join keys across the network.

Runtime filtering is primarily used:

- To optimize queries against large partitioned tables (under the name *dynamic partition pruning*)
- To optimize joins of large tables

The following terms are used in this topic to describe runtime filtering.

plan fragment

Impala decomposes each query into smaller units of work that are distributed across the cluster. Wherever possible, a data block is read, filtered, and aggregated by plan fragments executing on the same host. For some operations, such as joins and combining intermediate results into a final result set, data is transmitted across the network from one Impala daemon to another.

SCAN and HASH JOIN plan nodes

In the Impala query plan, a *scan node* performs the I/O to read from the underlying data files. Although this is an expensive operation from the traditional database perspective, Hadoop clusters and Impala are optimized to do this kind of I/O in a highly parallel fashion. The major potential cost savings come from using the columnar Parquet format (where Impala can avoid reading data for unneeded columns) and partitioned tables (where Impala can avoid reading data for unneeded partitions).

Most Impala joins use the *hash join* mechanism. (It is only fairly recently that Impala started using the nested-loop join technique, for certain kinds of non-equijoin queries.) In a hash join, when evaluating join conditions from two tables, Impala constructs a hash table in memory with all the different column values from the table on one side of the join. Then, for each row from the table on the other side of the join, Impala tests whether the relevant column values are in this hash table or not.

hash join

In a hash join, when evaluating join conditions from two tables, Impala constructs a hash table in memory with all the different column values from the table on one side of the join. Then, for each row from the table on the other side of the join, Impala tests whether the relevant column values are in this hash table or not.

- A *hash join node* constructs such an in-memory hash table, then performs the comparisons to identify which rows match the relevant join conditions and should be included in the result set (or at least sent on to the subsequent intermediate stage of query processing). Because some of the input for a hash join might be transmitted across the network from another host, it is

especially important from a performance perspective to prune out ahead of time any data that is known to be irrelevant.

The more distinct values are in the columns used as join keys, the larger the in-memory hash table and thus the more memory required to process the query.

broadcast join vs shuffle join

In a broadcast join, the table from one side of the join (typically the smaller table) is sent in its entirety to all the hosts involved in the query. Then each host can compare its portion of the data from the other (larger) table against the full set of possible join keys. In a shuffle join, there is no obvious "smaller" table, and so the contents of both tables are divided up, and corresponding portions of the data are transmitted to each host involved in the query.

A shuffle join is sometimes referred to in Impala as a *partitioned join*.

build and probe phases

When Impala processes a join query, the *build phase* is where the rows containing the join key columns, typically for the smaller table, are transmitted across the network and built into an in-memory hash table data structure on one or more destination nodes. The *probe phase* is where data is read locally (typically from the larger table) and the join key columns are compared to the values in the in-memory hash table. The corresponding input sources (tables, subqueries, and so on) for these phases are referred to as the *build side* and the *probe side*.

Runtime Filters

The *filter* that is transmitted between plan fragments is essentially a list of values for join key columns. When this list of values is transmitted in time to a scan node, Impala can filter out non-matching values immediately after reading them, rather than transmitting the raw data to another host to compare against the in-memory hash table on that host.

Impala supports the following types of filters based on the payload:

- *Bloom filter*: For HDFS-based tables, the Bloom filter uses a probability-based algorithm to determine all possible matching values. The probability-based aspects means that the filter might include some non-matching values, but if so, that does not cause any inaccuracy in the final results.
- *Min-max filter*: The filter is a data structure representing a minimum and maximum value. These filters are passed to Kudu to reduce the number of rows returned to Impala when scanning the probe side of the join. This filter currently only applies to Kudu tables.

Based on how filters from all join instances are aggregated, each of the above filters can be categorized as one of the following:

- *Broadcast filter*: A broadcast filter reflects the complete list of relevant values and can be immediately evaluated by a scan node.

  Broadcast filters are also classified as local or global. With a local broadcast filter, the information in the filter is used by a subsequent query fragment that is running on the same host that produced the filter. A non-local broadcast filter must be transmitted across the network to a query fragment that is running on a different host. Impala designates 3 hosts to each produce non-local broadcast filters, to guard against the possibility of a single slow host taking too long. Depending on the setting of the RUNTIME_FILTER_MODE query option (LOCAL or GLOBAL), Impala either uses a conservative optimization strategy where filters are only consumed on the same host that produced them, or a more aggressive strategy where filters are eligible to be transmitted across the network. The default for runtime filtering is the GLOBAL setting.
- *Partitioned filter*: A partitioned filter reflects only the values processed by one host in the cluster; all the partitioned filters must be combined into one (by the coordinator node) before the scan nodes can use the results to accurately filter the data as it is read from storage.

File Format Considerations for Runtime Filtering

Parquet tables get the most benefit from the runtime filtering optimizations. Runtime filtering can speed up join queries against partitioned or unpartitioned Parquet tables, and single-table queries against partitioned Parquet tables.

For other file formats (text, Avro, RCFile, and SequenceFile), runtime filtering speeds up queries against partitioned tables only. Because partitioned tables can use a mixture of formats, Impala produces the filters in all cases, even if they are not ultimately used to optimize the query.

Wait Intervals for Runtime Filters

Because it takes time to produce runtime filters, especially for partitioned filters that must be combined by the coordinator node, there is a time interval above which it is more efficient for the scan nodes to go ahead and construct their intermediate result sets, even if that intermediate data is larger than optimal. If it only takes a few seconds to produce the filters, it is worth the extra time if pruning the unnecessary data can save minutes in the overall query time. You can specify the maximum wait time in milliseconds using the RUNTIME_FILTER_WAIT_TIME_MS query option.

By default, each scan node waits for up to 1 second (1000 milliseconds) for filters to arrive. If all filters have not arrived within the specified interval, the scan node proceeds, using whatever filters did arrive to help avoid reading unnecessary data. If a filter arrives after the scan node begins reading data, the scan node applies that filter to the data that is read after the filter arrives, but not to the data that was already read.

If the cluster is relatively busy and your workload contains many resource-intensive or long-running queries, consider increasing the wait time so that complicated queries do not miss opportunities for optimization. If the cluster is lightly loaded and your workload contains many small queries taking only a few seconds, consider decreasing the wait time to avoid the 1 second delay for each query.

Query Options for Runtime Filtering

The following query options control runtime filtering.

- RUNTIME_FILTER_MODE

  This query option controls how extensively the filters are transmitted between hosts. By default, it is set to the highest level (GLOBAL).
- The other query options are tuning knobs that you typically only adjust after doing performance testing, and that you might want to change only for the duration of a single expensive query.
  - MAX_NUM_RUNTIME_FILTERS
  - DISABLE_ROW_RUNTIME_FILTERING
  - RUNTIME_FILTER_MAX_SIZE
  - RUNTIME_FILTER_MIN_SIZE
  - RUNTIME_BLOOM_FILTER_SIZE

Runtime Filtering and Query Plans

In the same way the query plan displayed by the EXPLAIN statement includes information about predicates used by each plan fragment, it also includes annotations showing whether a plan fragment produces or consumes a runtime filter.

- A plan fragment that produces a filter includes an annotation such as runtime filters: *FILTER_ID* <- *TABLE.COLUMN*
- A plan fragment that consumes a filter includes an annotation such as runtime filters: *FILTER_ID* -> *TABLE.COLUMN*

Setting the query option EXPLAIN_LEVEL=2 adds additional annotations showing the type of the filter:

- *FILTER_ID*[bloom] (for HDFS-based tables)
- *FILTER_ID*[min_max] (for Kudu tables)

The query profile (displayed by the PROFILE command in `impala-shell`) contains both the EXPLAIN plan and more detailed information about the internal workings of the query. The profile output includes the Filter routing table section with information about each filter based on its ID.

Tuning and Troubleshooting Queries that Use Runtime Filtering

These tuning and troubleshooting procedures apply to queries that are resource-intensive enough, long-running enough, and frequent enough that you can devote special attention to optimizing them individually.

Use the EXPLAIN statement and examine the runtime filters: lines to determine whether runtime filters are being applied to the WHERE predicates and join clauses that you expect. For example, runtime filtering does not apply to queries that use the nested loop join mechanism due to non-equijoin operators.

Make sure statistics are up-to-date for all tables involved in the queries. Use the COMPUTE STATS statement after loading data into non-partitioned tables, and COMPUTE INCREMENTAL STATS after adding new partitions to partitioned tables.

If join queries involving large tables use unique columns as the join keys, for example joining a primary key column with a foreign key column, the overhead of producing and transmitting the filter might outweigh the performance benefit because not much data could be pruned during the early stages of the query. For such queries, consider setting the query option RUNTIME_FILTER_MODE=OFF.

Distribute Runtime Filter Aggregation

The process of aggregating runtime filters at runtime can exert a substantial memory load on the coordinator. In response to this challenge, we initially introduced local aggregation of runtime filters within a single executor node, aiming to reduce the strain on the coordinator by transmitting filter updates only after local aggregation.

The Aggregating runtime filters feature is available from Cloudera Runtime 7.3.1.500 SP3

As Impala clusters scale up, the limitations of local filter aggregation become apparent, particularly in scenarios involving numerous nodes. This situation places significant memory stress on the coordinator node.

To mitigate this challenge, we have implemented a solution that distributes the runtime filter aggregation across specific Impala backends. The final aggregation takes place after this distributed aggregation, and the results are then published to the coordinator. This strategic distribution effectively alleviates the memory pressure on the coordinator, resulting in a more streamlined and accelerated runtime filter aggregation process.

The control mechanism for this enhancement lies in the query option MAX_NUM_FILTERS_AGGREGATED_PER_HOST. This option determines the number of executor nodes denoted as M, which are randomly selected as intermediate aggregators for the runtime filter. This approach optimizes the runtime filter aggregation process and contributes to overall system performance.

```
M = ceil(N / MAX_NUM_FILTERS_AGGREGATED_PER_HOST)
```

Where N is the number of all backend executors running the query, excluding the coordinator.

For instance, in a cluster of 400 nodes, each coordinator and intermediate aggregator would be configured to handle 20 filter updates. By aligning the MAX_NUM_FILTERS_AGGREGATED_PER_HOST with the square root of the cluster size, this approach aims to strike a balance and improve the runtime filter aggregation process, particularly in larger-scale deployments.

Min/Max Filtering

In addition to filtering at the partition level, the Parquet file format supports filtering at three levels: row group, page and row levels based on the minimum and maximum values of row group or page, and the value of the row in the file. These minimum/maximum column values are stored in the footer of the file. If the range between the minimum and maximum value in the file does not overlap with the range of data specified by the query, then the system skips the row group, page or row during scans.

Min/max filters are now enabled by default for equi-joins on lexical (leading column) or Z-Order (all columns) sorted columns in a Parquet table created by Impala. When such a sorted column is also the equi- join column, the min/max values in the filter are computed from the hash table build side and applied to the sorted column. In this way, the user can take advantage of Impala sorting the min/max values in column index in each data file for the table

by quickly eliminating non-matching row groups, pages or rows. The query option minmax_filter_sorted_columns which defaults to true can be used to control the feature. When minmax_filter_sorted_columns is true, the patch will generate min/max filters only for the sort columns. The normal control knobs minmax_filter_threshold (for threshold) and minmax_filtering_level (for filtering level) will also continue to work. When the threshold is 0, the patch automatically assigns a reasonable value for the threshold, and selects PAGE to be the filtering level.

In the backend, the skipped pages are quickly found by taking a fast code path to identify the corresponding lower and the upper bounds in the sorted min and max value arrays, given a range in the filter. The skipped pages are expressed as page ranges which are translated into row ranges later on.

A new query option minmax_filter_fast_code_path is added to control the work of the fast code path. This option can be assigned one of these three values ON (default), OFF, or VERIFICATION. The last value helps verify that the results from both the fast and the regular code path are the same.

The following example shows how min/max filtering on leading sort-by columns improves the performance of scan operators greatly.

```
select straight_join a.l_orderkey from
    simpflified_lineitem a join [SHUFFLE] tpch_parquet.lineitem b
    where a.l_orderkey = b.l_orderkey and b.l_receiptdate = "1998-12-31"
```

When this query is run on pages containing no more than 24000 rows, the results with different filtering options are as shown here:

- 84.62ms (page level filtering)
- 115.27ms (row group level filtering)
- 137.14ms (no filtering)

Support reading and writing Parquet bloom filters

Bloom filter is a performance optimization feature now available in Impala. This filter tells you, rapidly and memory-efficiently, whether the data you are looking for is present in a file.

Impala determines the appropriate conditions while the query is running. Impala can now read and write Parquet bloom filters. However, bloom filters can also provide false positives. If a bloom filter evaluates to:

- true: data might be present in the data file or might not be present.
- false: data is not present.

Currently, bloom filters are per column chunk entries which implies that you can skip entire row groups based on the filter. Writing a bloom filter is not useful for dictionary encoded columns, as all distinct values are included in the dictionary and the dictionary can give exact results in filtering with the predicate. If no value passes the predicate the whole row group can be skipped.

For example, if there is a predicate 'WHERE col = some_value' and some_value is not in the bloom filter the row group will be discarded.

Bloom filters support reading and writing columns with the following data types: integers, float, double, and Impala strings. Reading does not need any intervention from Impala, however, writing can be controlled by a new query option parquet_bloom_filter_write and the table property parquet.bloom.filter.columns.

The following table contains the list of the supported data types.

Parquet type	Impala type
INT32	TINYINT, SMALLINT, INT
INT64	BIGINT
FLOAT	FLOAT
DOUBLE	DOUBLE

BYTE_ARRAY	STRING

Query option

The query option for writing Parquet bloom filters (parquet_bloom_filter_write) accepts any of the following values:

- NEVER - never write Parquet bloom filters.
- IF_NO_DICT - write Parquet bloom filters if specified in the table properties AND if the row group is not fully dictionary encoded (the number of distinct values exceeds the maximum dictionary size); the row group may still be partially dictionary encoded, in which case the bloom filter contains all values from the whole row group, including those that are present in the dictionary.
- ALWAYS - always write Parquet bloom filters if specified in the table properties, even if the row group is fully dictionary encoded.

Table setting

The parquet.bloom.filter.columns table property is a comma separated list of 'col_name:bytes' pairs.

Where:

col_name is the name of the column for which a bloom filter should be written;

bytes represents the size (in bytes) of the bitset of the bloom filter, and is optional. If you do not provide the size, it will default to the maximal bloom filter size (ParquetBloomFilter::MAX_BYTES).

Example: “col1:1024,col2,col4:100"

Limitations

The following table contains the data types that are not supported currently. Support for these data types may be added in a future release.

Impala type	Reason for not supporting
VARCHAR(N)	truncation can change hash
CHAR(N)	padding / truncation can change hash
DECIMAL	multiple encodings supported
TIMESTAMP	multiple encodings supported, timezone conversion
DATE	not considered yet

Late Materialization of Columns

This is an optimization feature available in Impala and is implemented to optimize queries that reference a large number of columns and filter out a significant portion of the rows. This currently works only for Parquet tables.

When a query is processed, any fields referenced in the select list are decoded from file storage and materialized into row values in memory to be processed for joins, predicates, and so on. Prior to this optimization, the materialization is applied to all the row values, regardless of whether they are eliminated by predicates. However with late materialization, as the name implies, only fields that are referenced by predicates and rows that are not filtered out by predicates are fully materialized.

When a select * query is run over a 4 billion row table that returns a single row, it may take ~30 seconds to execute, however it may take less than 3 seconds if the query is replaced by select field1, field2 .... This is because the select * query is materializing all the fields for rows, regardless of whether they match or not. And in the case of the select f ield1, field2 query only the columns that are required for filtering the data need to be materialized.

Use the configuration parameter (query option) parquet_late_materialization_threshold to provide the minimum number of consecutive rows that are filtered out to avoid materialization. If set to less than 0, it disables the late materialization. In most cases the default setting of 20 will provide optimal performance.

The behavior of parquet_late_materialization_threshold has been updated. If the query option is set to a value greater than or equal to 0 and a collection value can be skipped, the late materialization threshold is now set to 1. Otherwise, it defaults to the value specified by the query option. This change was introduced with Apache Jira: IMPALA-3841.

Partitioning

Use partitioning to speed up queries that retrieve data based on values from one or more columns.

Partitioning is a technique for physically dividing the data during loading based on values from one or more columns. For example, with a school_records table partitioned on a year column, there is a separate data directory for each different year value, and all the data for that year is stored in a data file in that directory. A query that includes a WHERE condition such as YEAR=1966, YEAR IN (1989,1999), or YEAR BETWEEN 1984 AND 1989 can examine only the data files from the appropriate directory or directories, greatly reducing the amount of data to read and test.

Parquet is a popular format for partitioned Impala tables because it is well suited to handle huge data volumes.

You can add, drop, set the expected file format, or set the HDFS location of the data files for individual partitions within an Impala table.

When to Use Partitioned Tables

Partitioning is typically appropriate for:

- Tables that are very large, where reading the entire data set takes an impractical amount of time.
- Tables that are always or almost always queried with conditions on the partitioning columns. In our example of a table partitioned by year, SELECT COUNT(*) FROM school_records WHERE year = 1985 is efficient, only examining a small fraction of the data; but SELECT COUNT(*) FROM school_records has to process a separate data file for each year, resulting in more overall work than in an unpartitioned table. You would probably not partition this way if you frequently queried the table based on last name, student ID, and so on without testing the year.
- Columns that have reasonable cardinality (number of different values). If a column only has a small number of values, for example Male or Female, you do not gain much efficiency by eliminating only about 50% of the data to read for each query. If a column has only a few rows matching each value, the number of directories to process can become a limiting factor, and the data file in each directory could be too small to take advantage of the Hadoop mechanism for transmitting data in multi-megabyte blocks. For example, you might partition census data by year, store sales data by year and month, and web traffic data by year, month, and day. (Some users with high volumes of incoming data might even partition down to the individual hour and minute.)
- Data that already passes through an extract, transform, and load (ETL) pipeline. The values of the partitioning columns are stripped from the original data files and represented by directory names, so loading data into a partitioned table involves some sort of transformation or preprocessing.

SQL Statement for Partitioned Tables

In terms of Impala SQL syntax, partitioning affects these statements:

- CREATE TABLE: You specify a PARTITIONED BY clause when creating the table to identify names and data types of the partitioning columns. These columns are not included in the main list of columns for the table.
- CREATE TABLE AS SELECT: Use the PARTITIONED BY clause to create a partitioned table, copy data into it, and create new partitions based on the values in the inserted data.
- ALTER TABLE: Add or drop partitions, to work with different portions of a huge data set. You can designate the HDFS directory that holds the data files for a specific partition. With data partitioned by date values, you might “age out” data that is no longer relevant.

  If you are creating a partition for the first time and specifying its location, for maximum efficiency, use a single ALTER TABLE statement including both the ADD PARTITION and LOCATIONclauses, rather than separate statements with ADD PARTITION and SET LOCATION clauses.

- INSERT: When you insert data into a partitioned table, you identify the partitioning columns. One or more values from each inserted row are not stored in data files, but instead determine the directory where that row value is stored. You can also specify which partition to load a set of data into, with INSERT OVERWRITE statements; you can replace the contents of a specific partition but you cannot append data to a specific partition.

  By default, if an INSERT statement creates any new subdirectories underneath a partitioned table, those subdirectories are assigned default HDFS permissions for the impala user. To make each subdirectory have the same permissions as its parent directory in HDFS, specify the ##insert_inherit_permissions startup option for the Impala Daemon.
- Although the syntax of the SELECT statement is the same whether or not the table is partitioned, the way queries interact with partitioned tables can have a dramatic impact on performance and scalability. The mechanism that lets queries skip certain partitions during a query is known as partition pruning.
- SHOW PARTITIONS: Displays information about each partition in a table.

Static and Dynamic Partitioning Clauses

Specifying all the partition columns in a SQL statement is called *static partitioning*, because the statement affects a single predictable partition. For example, you use static partitioning with an ALTER TABLE statement that affects only one partition, or with an INSERT statement that inserts all values into the same partition:

```
INSERT INTO t1 PARTITION(x=10, y='a') SELECT c1 FROM some_other_table;
```

When you specify some partition key columns in an INSERT statement, but leave out the values, Impala determines which partition to insert. This technique is called *dynamic partitioning*.

```
INTSERT INTO t1 PARTITION(x, y='b') SELECT c1, c2 FROM some_other_table;
-- Create new partition if necessary based on variable year, month, and day;
 insert a single value.
INSERT INTO weather PARTITION (year, month, day) SELECT 'cloudy',2014,4,21;
-- Create new partition if necessary for specified year and month but variab
le day; insert a single value.
INSERT INTO weather PARTITION (year=2014, month=04, day) SELECT 'sunny',22;
```

The more key columns you specify in the PARTITION clause, the fewer columns you need in the SELECT list. The trailing columns in the SELECT list are substituted in order for the partition key columns with no specified value.

Refreshing a Single Partition

The REFRESH statement is typically used with partitioned tables when new data files are loaded into a partition by some non-Impala mechanism, such as a Hive or Spark job. The REFRESH statement makes Impala aware of the new data files so that they can be used in Impala queries. Because partitioned tables typically contain a high volume of data, the REFRESH operation for a full partitioned table can take significant time.

You can include a PARTITION (*PARTITION_SPEC*) clause in the REFRESH statement so that only a single partition is refreshed. For example, REFRESH big_table PARTITION (year=2017, month=9, day=30). The partition spec must include all the partition key columns.

Partition Key Columns

The columns you choose as the partition keys should be ones that are frequently used to filter query results in important, large-scale queries. Popular examples are some combination of year, month, and day when the data has associated time values, and geographic region when the data is associated with some place.

- For time-based data, split out the separate parts into their own columns, because Impala cannot partition based on a TIMESTAMP column.
- The data type of the partition columns does not have a significant effect on the storage required, because the values from those columns are not stored in the data files, rather they are represented as strings inside HDFS directory names.

- You can enable the OPTIMIZE_PARTITION_KEY_SCANS query option to speed up queries that only refer to partition key columns, such as SELECT MAX(year). This setting is not enabled by default because the query behavior is slightly different if the table contains partition directories without actual data inside.
- All the partition key columns must be scalar types.
- When Impala queries data stored in HDFS, it is most efficient to use multi-megabyte files to take advantage of the HDFS block size. For Parquet tables, the block size (and ideal size of the data files) is 256 MB.. Therefore, avoid specifying too many partition key columns, which could result in individual partitions containing only small amounts of data. For example, if you receive 1 GB of data per day, you might partition by year, month, and day; while if you receive 5 GB of data per minute, you might partition by year, month, day, hour, and minute. If you have data with a geographic component, you might partition based on postal code if you have many megabytes of data for each postal code, but if not, you might partition by some larger region such as city, state, or country. state

If you frequently run aggregate functions such as MIN(), MAX(), and COUNT(DISTINCT) on partition key columns, consider enabling the OPTIMIZE_PARTITION_KEY_SCANS query option, which optimizes such queries.

Setting Different File Formats for Partitions in a Table

Partitioned tables have the flexibility to use different file formats for different partitions. For example, if you originally received data in text format, then received new data in RCFile format, and eventually began receiving data in Parquet format, all that data could reside in the same table for queries. You just need to ensure that the table is structured so that the data files that use different file formats reside in separate partitions.

For example, here is how you might switch from text to Parquet data as you receive data for different years:

```
[localhost:21000] > CREATE TABLE census (name STRING) PARTITIONED BY (year S
MALLINT);
[localhost:21000] > ALTER TABLE census ADD PARTITION (year=2012); -- Text
format;
[localhost:21000] > ALTER TABLE census ADD PARTITION (year=2013); -- Text f
ormat switches to Parquet before data loaded;
[localhost:21000] > ALTER TABLE census PARTITION (year=2013) SET FILEFORMAT
 PARQUET;

[localhost:21000] > INSERT INTO census PARTITION (year=2012) VALUES ('Smith'
),('Jones'),('Lee'),('Singh');
[localhost:21000] > INSERT INTO census PARTITION (year=2013) VALUES ('Flo
res'),('Bogomolov'),('Cooper'),('Appiah');
```

At this point, the HDFS directory for year=2012 contains a text-format data file, while the HDFS directory for year =2013 contains a Parquet data file. As always, when loading non-trivial data, you would use INSERT ... SELECT or LOAD DATA to import data in large batches, rather than INSERT ... VALUES which produces small files that are inefficient for real-world queries.

For other file types that Impala cannot create natively, you can switch into Hive and issue the ALTER TABLE ... SET FILEFORMAT statements and INSERT or LOAD DATA statements there. After switching back to Impala, issue a REFRESH *TABLE_NAME* statement so that Impala recognizes any partitions or new data added through Hive.

Dropping Partitions

What happens to the data files when a partition is dropped depends on whether the partitioned table is designated as internal or external. For an internal (managed) table, the data files are deleted. For example, if data in the partitioned table is a copy of raw data files stored elsewhere, you might save disk space by dropping older partitions that are no longer required for reporting, knowing that the original data is still available if needed later. For an external table, the data files are left alone. For example, dropping a partition without deleting the associated files lets Impala consider a smaller set of partitions, improving query efficiency and reducing overhead for DDL operations on the table; if the data is needed again later, you can add the partition again.

Using Partitioning with Kudu Tables

Kudu tables use a more fine-grained partitioning scheme than tables containing HDFS data files. You specify a PART ITION BY clause with the CREATE TABLE statement to identify how to divide the values from the partition key columns.

Keeping Statistics Up to Date for Partitioned Tables

Because the COMPUTE STATS statement can be resource-intensive to run on a partitioned table as new partitions are added, Impala includes a variation of this statement that allows computing statistics on a per-partition basis such that stats can be incrementally updated when new partitions are added.

The COMPUTE INCREMENTAL STATS variation computes statistics only for partitions that were added or changed since the last COMPUTE INCREMENTAL STATS statement, rather than the entire table. It is typically used for tables where a full COMPUTE STATS operation takes too long to be practical each time a partition is added or dropped.

Partition Pruning for Queries

Partition pruning refers to the mechanism where a query can skip reading the data files corresponding to one or more partitions.

If you can arrange for queries to prune large numbers of unnecessary partitions from the query execution plan, the queries use fewer resources and are thus proportionally faster and more scalable.

For example, if a table is partitioned by columns YEAR, MONTH, and DAY, then WHERE clauses such as WHER E year = 2013, WHERE year < 2010, or WHERE year BETWEEN 1995 AND 1998 allow Impala to skip the data files in all partitions outside the specified range. Likewise, WHERE year = 2013 AND month BETWEEN 1 AND 3 could prune even more partitions, reading the data files for only a portion of one year.

If a view applies to a partitioned table, any partition pruning considers the clauses on both the original query and any additional WHERE predicates in the query that refers to the view.

In queries involving both analytic functions and partitioned tables, partition pruning only occurs for columns named in the PARTITION BY clause of the analytic function call. For example, if an analytic function query has a clause such as WHERE year=2016, the way to make the query prune all other YEAR partitions is to include PART ITION BY year in the analytic function call; for example, OVER (PARTITION BY year,*OTHER_COLUMNS OTHER_ANALYTIC_CLAUSES*).

Checking if Partition Pruning Happens for a Query

To check the effectiveness of partition pruning for a query, check the EXPLAIN output for the query before running it. For example, this example shows a table with 3 partitions, where the query only reads 1 of them. The notation #par titions=1/3 in the EXPLAIN plan confirms that Impala can do the appropriate partition pruning.

```
[localhost:21000] > INSERT INTO census PARTITION (year=2010) VALUES ('Smith'
),('Jones');
[localhost:21000] > INSERT INTO census PARTITION (year=2011) VALUES ('Smith
'),('Jones'),('Doe');
[localhost:21000] > INSERT INTO census PARTITION (year=2012) VALUES ('Smit
h'),('Doe');
[localhost:21000] > EXPLAIN select name from census where year=2010;
+----------------------------------------------------------------------+
| Explain String                                                       |
+----------------------------------------------------------------------+
| PLAN FRAGMENT 1                                                      |
|   PARTITION: RANDOM                                                  |
|                                                                      |
|   STREAM DATA SINK                                                   |
|     EXCHANGE ID: 1                                                   |
|     UNPARTITIONED                                                    |
|                                                                      |
```

```
|   0:SCAN HDFS                                                        |
|      table=predicate_propagation.census #partitions=1/3 size=12B |
+----------------------------------------------------------------------+
```

For a report of the volume of data that was actually read and processed at each stage of the query, check the output of the SUMMARY command immediately after running the query. For a more detailed analysis, look at the output of the PROFILE command; it includes this same summary report near the start of the profile output.

Predicate Propagation

Impala can do partition pruning in cases where the partition key column is not directly compared to a constant. Using the predicate propagation technique, Impala applies the transitive property to other parts of the WHERE clause.

In this example, the census table includes another column indicating when the data was collected, which happens in 10-year intervals. Even though the query does not compare the partition key column (YEAR) to a constant value, Impala can deduce that only the partition YEAR=2010 is required, and again only reads 1 out of 3 partitions.

```
[localhost:21000] > CREATE TABLE census (name STRING, census_year INT) PARTI
TIONED BY (year INT);
[localhost:21000] > INSERT INTO census PARTITION (year=2010) VALUES  ('Smi
th',2010),('Jones',2010);
[localhost:21000] > INSERT INTO census PARTITION (year=2011) VALUES  ('Smi
th',2020),('Jones',2020),('Doe',2020);
[localhost:21000] > INSERT INTO census PARTITION (year=2012) VALUES  ('Sm
ith',2020),('Doe',2020);
[localhost:21000] >
[localhost:21000] > EXPLAIN select name from census where year = census_year
 and census_year=2010;
+----------------------------------------------------------------------+
| Explain String                                                       |
+----------------------------------------------------------------------+
| PLAN FRAGMENT 1                                                      |
|   PARTITION: RANDOM                                                  |
|                                                                      |
|   STREAM DATA SINK                                                   |
|     EXCHANGE ID: 1                                                   |
|     UNPARTITIONED                                                    |
|                                                                      |
|   0:SCAN HDFS                                                        |
|      table=predicate_propagation.census #partitions=1/3 size=22B |
|      predicates: census_year = 2010, year = census_year              |
+----------------------------------------------------------------------+
```

Dynamic Partition Pruning

Impala supports two types of partition pruning.

- *Static partition pruning*: The conditions in the WHERE clause are analyzed to determine in advance which partitions can be safely skipped.
- *Dynamic partition pruning*: Information about the partitions is collected during the query execution, and Impala prunes unnecessary partitions. The information is not available in advance before runtime.

For example, if partition key columns are compared to literal values in a WHERE clause, Impala can perform static partition pruning during the planning phase to only read the relevant partitions:

```
-- The query only needs to read 3 partitions whose key values are known ahe
ad of time.
-- That's static partition pruning.
SELECT COUNT(*) FROM sales_table WHERE year IN (2005, 2010, 2015);
```

Dynamic partition pruning involves using information only available at run time, such as the result of a subquery. The following example shows a simple dynamic partition pruning.

```
CREATE TABLE yy (s STRING) PARTITIONED BY (year INT);
INSERT INTO yy PARTITION (year) VALUES ('1999', 1999), ('2000', 2000),
  ('2001', 2001), ('2010', 2010), ('2018', 2018);
COMPUTE STATS yy;

CREATE TABLE yy2 (s STRING, year INT);
INSERT INTO yy2 VALUES ('1999', 1999), ('2000', 2000), ('2001', 2001);
COMPUTE STATS yy2;
-- The following query reads an unknown number of partitions, whose key val
ues
-- are only known at run time. The runtime filters line shows the
-- information used in query fragment 02 to decide which partitions to skip.

EXPLAIN SELECT s FROM yy WHERE year IN (SELECT year FROM yy2);
+-----------------------------------------------------------------------------+
| PLAN-ROOT SINK                                                              |
| |                                                                           |
| 04:EXCHANGE [UNPARTITIONED]                                                 |
| |                                                                           |
| 02:HASH JOIN [LEFT SEMI JOIN, BROADCAST]                                    |
| |  hash predicates: year = year                                             |
| |  runtime filters: RF000 <- year                                     |     |
| |                                                                           |
| |--03:EXCHANGE [BROADCAST]                                                  |
| |  |                                                                        |
| |  01:SCAN HDFS [default.yy2]                                               |
| |     partitions=1/1 files=1 size=620B                                      |
| |                                                                           |
| 00:SCAN HDFS [default.yy]                                                   |
|    partitions=5/5 files=5 size=1.71KB                               |     |
|    runtime filters: RF000 -> year                                           |
+-----------------------------------------------------------------------------+

SELECT s FROM yy WHERE year IN (SELECT year FROM yy2); -- Returns 3 rows fro
m yy
PROFILE;
```

In the above example, Impala evaluates the subquery, sends the subquery results to all Impala nodes participating in the query, and then each `impalad` daemon uses the dynamic partition pruning optimization to read only the partitions with the relevant key values.

The output query plan from the EXPLAIN statement shows that runtime filters are enabled. The plan also shows that it expects to read all 5 partitions of the yy table, indicating that static partition pruning will not happen.

The Filter summary in the PROFILE output shows that the scan node filtered out based on a runtime filter of dynamic partition pruning.

```
Filter 0 (1.00 MB):
 - Files processed: 3
 - Files rejected: 1 (1)
 - Files total: 3 (3)
```

Dynamic partition pruning is especially effective for queries involving joins of several large partitioned tables. Evaluating the ON clauses of the join predicates might normally require reading data from all partitions of certain tables. If the WHERE clauses of the query refer to the partition key columns, Impala can now often skip reading many of the partitions while evaluating the ON clauses. The dynamic partition pruning optimization reduces the amount of I/O and the amount of intermediate data stored and transmitted across the network during the query.

Dynamic partition pruning is part of the runtime filtering feature, which applies to other kinds of queries in addition to queries against partitioned tables.

HDFS Caching

Impala can use the HDFS caching feature to make more effective use of RAM so that repeated queries can take advantage of data “pinned” in memory regardless of how much data is processed overall.

The HDFS caching feature lets you designate a subset of frequently accessed data to be pinned permanently in memory, remaining in the cache across multiple queries and never being evicted. This technique is suitable for tables or partitions that are frequently accessed and are small enough to fit entirely within the HDFS memory cache. For example, you might designate several dimension tables to be pinned in the cache, to speed up many different join queries that reference them. Or in a partitioned table, you might pin a partition holding data from the most recent time period because that data will be queried intensively; then when the next set of data arrives, you could unpin the previous partition and pin the partition holding the new data.

The performance gain comes from two aspects:

- Reading from RAM instead of disk
- Accessing the data straight from the cache area instead of copying from one RAM area to another

  This yields further performance improvement over the standard OS caching mechanism, which still results in memory-to-memory copying of cached data. Because accessing HDFS cached data avoids a memory-to-memory copy operation, queries involving cached data require less memory on the Impala side than the equivalent queries on uncached data.

Due to a limitation of HDFS, zero-copy reads are not supported with encryption. Cloudera recommends not using HDFS caching for Impala data files in encryption zones. The queries fall back to the normal read path during query execution, which might cause some performance overhead.

Because this Impala performance feature relies on HDFS infrastructure, it only applies to Impala tables that use HDFS data files. HDFS caching for Impala does not apply to HBase tables, S3 tables, Kudu tables, or Isilon tables.

Using HDFS Caching for Impala Tables and Partitions

Begin by choosing which tables or partitions to cache. For example, these might be lookup tables that are accessed by many different join queries, or partitions corresponding to the most recent time period that are analyzed by different reports or ad-hoc queries.

In your SQL statements, you specify logical divisions such as tables and partitions to be cached. Impala translates these requests into HDFS-level directives that apply to particular directories and files. For example, given a partitioned table CENSUS with a partition key column YEAR, you could choose to cache all or part of the data as follows.

```
-- Cache the entire table (all partitions).
ALTER TABLE census SET CACHED IN 'POOL_NAME';

-- Remove the entire table from the cache.
ALTER TABLE census SET UNCACHED;
-- Cache a portion of the table (a single partition).
-- If the table is partitioned by multiple columns (such as year, month,
day),
-- the ALTER TABLE command must specify values for all those columns.
ALTER TABLE census PARTITION (year=1960) SET CACHED IN 'POOL_NAME';

-- Cache the data from one partition on up to 4 hosts, to minimize CPU load
 on any
-- single host when the same data block is processed multiple times.
ALTER TABLE census PARTITION (year=1970)
  SET CACHED IN 'POOL_NAME' WITH REPLICATION = 4;
```

```
-- At each stage, check the volume of cached data.
-- For large tables or partitions, the background loading might take some t
ime,
-- so you might have to wait and reissue the statement until all the data
-- has finished being loaded into the cache.
SHOW TABLE STATS census;
+-------+-------+--------+------+--------------+--------+
| year  | #Rows | #Files | Size | Bytes Cached | Format |
+-------+-------+--------+------+--------------+--------+
| 1900  | -1    | 1      | 11B  | NOT CACHED   | TEXT   |
| 1940  | -1    | 1      | 11B  | NOT CACHED   | TEXT   |
| 1960  | -1    | 1      | 11B  | 11B          | TEXT   |
| 1970  | -1    | 1      | 11B  | NOT CACHED   | TEXT   |
| Total | -1    | 4      | 44B  | 11B          |        |
+-------+-------+--------+------+--------------+--------+
```

CREATE TABLE and ALTER TABLE considerations:

The HDFS caching feature affects the Impala CREATE TABLE statement as follows:

- You can put a CACHED IN '*POOL_NAME*' clause and optionally a WITH REPLICATION = *NUMBER_OF_HOSTS* clause at the end of a CREATE TABLE statement to automatically cache the entire contents of the table, including any partitions added later.

  The *POOL_NAME* is a pool that you previously set up with the `hdfs cacheadmin` command in HDFS.
- Once a table is designated for HDFS caching through the CREATE TABLE statement, if new partitions are added later through ALTER TABLE ... ADD PARTITION statements, the data in those new partitions is automatically cached in the same pool.
- If you want to perform repetitive queries on a subset of data from a large table, and it is not practical to designate the entire table or specific partitions for HDFS caching, you can create a new cached table with just a subset of the data by using CREATE TABLE ... CACHED IN '*POOL_NAME*' AS SELECT ... WHERE .... When you are finished with generating reports from this subset of data, drop the table and both the data files and the data cached in RAM are automatically deleted.
- If you have designated a table or partition as cached through the CREATE TABLE or ALTER TABLE statements, subsequent attempts to relocate the table or partition through an ALTER TABLE ... SET LOCATION statement will fail. You must issue an ALTER TABLE ... SET UNCACHED statement for the table or partition first. Otherwise, Impala would lose track of some cached data files and have no way to uncache them later.
- The optional WITH REPLICATION clause for CREATE TABLE and ALTER TABLE lets you specify a *replication factor*, the number of hosts on which to cache the same data blocks. When Impala processes a cached data block, where the cache replication factor is greater than 1, Impala randomly selects a host that has a cached copy of that data block. This optimization avoids excessive CPU usage on a single host when the same cached data block is processed multiple times. Cloudera recommends specifying a value greater than or equal to the HDFS block replication factor.

INSERT and LOAD DATA considerations:

- When HDFS caching is enabled for a table or partition, new data files are cached automatically when they are added to the appropriate directory in HDFS, without the need for a REFRESH statement in Impala. Impala automatically performs a REFRESH once the new data is loaded into the HDFS cache.
- If you perform an INSERT or LOAD DATA through Hive, Impala only recognizes the new data files after a REFRESH *TABLE_NAME* statement in Impala.
- If the cache pool is entirely full, or becomes full before all the requested data can be cached, the Impala DDL statement returns an error. This is to avoid situations where only some of the requested data could be cached.

DROP TABLE and ALTER TABLE DROP PARTITION considerations:

The HDFS caching feature interacts with the Impala DROP TABLE and ALTER TABLE ... DROP PARTITION statements as follows:

- When you issue a DROP TABLE or ALTER TABLE ... DROP PARTITION for a table that is entirely cached, or has some partitions cached, the DROP TABLE succeeds and all the cache directives Impala submitted for that table are removed from the HDFS cache system.
- The underlying data files are removed if the dropped table is an internal table, or the dropped partition is in its default location underneath an internal table. The data files are left alone if the dropped table is an external table, or if the dropped partition is in a non-default location.
- If you drop an HDFS cache pool through the `hdfs cacheadmin` command, all the Impala data files are preserved, just no longer cached. After a subsequent REFRESH, SHOW TABLE STATS reports 0 bytes cached for each associated Impala table or partition.
- If you designated the data files as cached through the `hdfs cacheadmin` command, and the data files are left behind as described in the previous item, the data files remain cached.

  Impala only removes the cache directives submitted by Impala through the CREATE TABLE or ALTER TABLE statements.
- One file can have multiple redundant cache directives. The directives all have unique IDs, and owners so that the system can tell them apart.

SHOW TABLE STATS and SHOW PARTITIONS considerations:

For each table or partition, the SHOW TABLE STATS or SHOW PARTITIONS statement displays the number of bytes currently cached by the HDFS caching feature.

A value of 0, or a smaller number than the overall size of the table or partition, indicates that the cache request has been submitted but the data has not been entirely loaded into memory yet.

If there are no cache directives in place for that table or partition, the result set displays NOT CACHED.

SELECT considerations:

The Impala HDFS caching feature interacts with the SELECT statement and query performance as follows:

- Impala automatically reads from memory any data that has been designated as cached and actually loaded into the HDFS cache. (It could take some time after the initial request to fully populate the cache for a table with large size or many partitions.)
- Impala queries take advantage of HDFS cached data regardless of whether the cache directive was issued by Impala or externally through the `hdfs cacheadmin` command, for example for an external table where the cached data files might be accessed by several different Hadoop components.
- If your query returns a large result set, the time reported for the query could be dominated by the time needed to print the results on the screen. To measure the time for the underlying query processing, query the COUNT() of the big result set, which does all the same processing but only prints a single line to the screen.
- Impala automatically randomizes which host processes a cached HDFS block, to avoid CPU hotspots. For tables where HDFS caching is not applied, Impala designates which host to process a data block using an algorithm that estimates the load on each host. If CPU hotspots still arise during queries, you can enable additional randomization for the scheduling algorithm for non-HDFS cached data by setting the SCHEDULE_RANDOM_ REPLICA query option.
- If you drop a cache pool with the `hdfs cacheadmin` command, Impala queries against the associated data files will still work, by falling back to reading the files from disk. After performing a REFRESH on the table, Impala reports the number of bytes cached as 0 for all associated tables and partitions.

Performance Considerations for HDFS Caching with Impala

Impala supports efficient reads from data that is pinned in memory through HDFS caching. Impala takes advantage of the HDFS API and reads the data from memory rather than from disk whether the data files are pinned using Impala DDL statements, or using the command-line mechanism where you specify HDFS paths.

When you examine the output of the `impala-shell SUMMARY` command, or look in the metrics report for the `impalad` daemon, you see how many bytes are read from the HDFS cache. For example, this excerpt from a query

profile illustrates that all the data read during a particular phase of the query came from the HDFS cache, because the BytesRead and BytesReadDataNodeCache values are identical.

```
HDFS_SCAN_NODE (id=0):(Total: 11s114ms, non-child: 11s114ms, % non-child: 10
0.00%)
        - AverageHdfsReadThreadConcurrency: 0.00
        - AverageScannerThreadConcurrency: 32.75
        - BytesRead: 10.47 GB (11240756479)
        - BytesReadDataNodeCache: 10.47 GB (11240756479)
        - BytesReadLocal: 10.47 GB (11240756479)
        - BytesReadShortCircuit: 10.47 GB (11240756479)
        - DecompressionTime: 27s572ms
```

For queries involving smaller amounts of data, or in single-user workloads, you might not notice a significant difference in query response time with or without HDFS caching. Even with HDFS caching turned off, the data for the query might still be in the Linux OS buffer cache. The benefits become clearer as data volume increases, and especially as the system processes more concurrent queries. HDFS caching improves the scalability of the overall system. That is, it prevents query performance from declining when the workload outstrips the capacity of the Linux OS cache.

- For small amounts of data, the query speedup might not be noticeable in terms of wall clock time. The performance might be roughly the same with HDFS caching turned on or off, due to recently used data being held in the OS cache. The difference is more pronounced with:
  - Data volumes (for all queries running concurrently) that exceed the size of the OS cache.
  - A busy cluster running many concurrent queries, where the reduction in memory-to-memory copying and overall memory usage during queries results in greater scalability and throughput.

When data is requested to be pinned in memory, that process happens in the background without blocking access to the data while the caching is in progress. Loading the data from disk could take some time. Impala reads each HDFS data block from memory if it has been pinned already, or from disk if it has not been pinned yet.

Memory Considerations

The amount of data that you can pin on each node through the HDFS caching mechanism is subject to a quota that is enforced by the underlying HDFS service. Before requesting to pin an Impala table or partition in memory, check that its size does not exceed this quota.



Note: Because the HDFS cache consists of combined memory from all the DataNodes in the cluster, cached tables or partitions can be bigger than the amount of HDFS cache memory on any single host.

The Impala HDFS caching feature interacts with the Impala memory limits as follows:

- The maximum size of each HDFS cache pool is specified externally to Impala, through the `hdfs cacheadmin` command.
- All the memory used for HDFS caching is separate from the `impalad` daemon address space and does not count towards the limits of the --mem_limit startup option, MEM_LIMIT query option, or further limits imposed through YARN resource management or the Linux cgroups mechanism.

HDFS Block Skew

For best performance of Impala parallel queries, the work is divided equally across hosts in the cluster, and all hosts take approximately equal time to finish their work. If one host takes substantially longer than others, the extra time needed for the slow host can become the dominant factor in query performance. Therefore, one of the first steps in performance tuning for Impala is to detect and correct such conditions.

The main cause of uneven performance that you can correct within Impala is *skew* in the number of HDFS data blocks processed by each host, where some hosts process substantially more data blocks than others. This condition can occur because of uneven distribution of the data values themselves, for example causing certain data files or

partitions to be large while others are very small. (Although it is possible to have unevenly distributed data without any problems with the distribution of HDFS blocks.) Block skew could also be due to the underlying block allocation policies within HDFS, the replication factor of the data files, and the way that Impala chooses the host to process each data block.

The most convenient way to detect block skew, or slow-host issues in general, is to examine the “executive summary” information from the query profile after running a query:

- In `impala-shell`, issue the SUMMARY command immediately after the query is complete, to see just the summary information. If you detect issues involving skew, you might switch to issuing the PROFILE command, which displays the summary information followed by a detailed performance analysis.
- In the Cloudera Manager interface or the Impala debug web UI, click on the Profile link associated with the query after it is complete. The executive summary information is displayed early in the profile output.

For each phase of the query, you see an Avg Time and a Max Time value, along with #Hosts indicating how many hosts are involved in that query phase. For all the phases with #Hosts greater than one, look for cases where the maximum time is substantially greater than the average time. Focus on the phases that took the longest, for example, those taking multiple seconds rather than milliseconds or microseconds.

If you detect that some hosts take longer than others, first rule out non-Impala causes. One reason that some hosts could be slower than others is if those hosts have less capacity than the others, or if they are substantially busier due to unevenly distributed non-Impala workloads:

- For clusters running Impala, keep the relative capacities of all hosts roughly equal. Any cost savings from including some underpowered hosts in the cluster will likely be outweighed by poor or uneven performance, and the time spent diagnosing performance issues.
- If non-Impala workloads cause slowdowns on some hosts but not others, use the appropriate load-balancing techniques for the non-Impala components to smooth out the load across the cluster.

If the hosts on your cluster are evenly powered and evenly loaded, examine the detailed profile output to determine which host is taking longer than others for the query phase in question. Examine how many bytes are processed during that phase on that host, how much memory is used, and how many bytes are transmitted across the network.

The most common symptom is a higher number of bytes read on one host than others, due to one host being requested to process a higher number of HDFS data blocks. This condition is more likely to occur when the number of blocks accessed by the query is relatively small. For example, if you have a 10-node cluster and the query processes 10 HDFS blocks, each node might not process exactly one block. If one node sits idle while another node processes two blocks, the query could take twice as long as if the data was perfectly distributed.

Possible solutions in this case include:

- If the query is artificially small, perhaps for benchmarking purposes, scale it up to process a larger data set. For example, if some nodes read 10 HDFS data blocks while others read 11, the overall effect of the uneven distribution is much lower than when some nodes did twice as much work as others. As a guideline, aim for a “sweet spot” where each node reads 2 GB or more from HDFS per query. Queries that process lower volumes than that could experience inconsistent performance that smooths out as queries become more data-intensive.
- If the query processes only a few large blocks, so that many nodes sit idle and cannot help to parallelize the query, consider reducing the overall block size. For example, you might adjust the PARQUET_FILE_SIZE query option before copying or converting data into a Parquet table. Or you might adjust the granularity of data files produced earlier in the ETL pipeline by non-Impala components. In Impala 2.0 and later, the default Parquet block size is 256 MB, reduced from 1 GB, to improve parallelism for common cluster sizes and data volumes.
- Reduce the amount of compression applied to the data. For text data files, the highest degree of compression (gzip) produces unsplittable files that are more difficult for Impala to process in parallel, and require extra memory during processing to hold the compressed and uncompressed data simultaneously. For binary formats such as Parquet and Avro, compression can result in fewer data blocks overall, but remember that when queries process relatively few blocks, there is less opportunity for parallel execution and many nodes in the cluster might sit idle. Note that when Impala writes Parquet data with the query option COMPRESSION_CODEC=NONE enabled, the data is still typically compact due to the encoding schemes used by Parquet, independent of the final compression step.

Understanding Performance using EXPLAIN Plan

To understand the high-level performance considerations for Impala queries, read the output of the EXPLAIN statement for the query. You can get the EXPLAIN plan without actually running the query itself.

The EXPLAIN statement gives you an outline of the logical steps that a query will perform, such as how the work will be distributed among the nodes and how intermediate results will be combined to produce the final result set. You can see these details before actually running the query. You can use this information to check that the query will not operate in some very unexpected or inefficient way.

Read the EXPLAIN plan from bottom to top:

- The last part of the plan shows the low-level details such as the expected amount of data that will be read, where you can judge the effectiveness of your partitioning strategy and estimate how long it will take to scan a table based on total data size and the size of the cluster.
- As you work your way up, next you see the operations that will be parallelized and performed on each Impala node.
- At the higher levels, you see how data flows when intermediate result sets are combined and transmitted from one node to another.
- The EXPLAIN_LEVEL query option lets you customize how much detail to show in the EXPLAIN plan depending on whether you are doing high-level or low-level tuning, dealing with logical or physical aspects of the query.

Read the EXPLAIN plan from bottom to top:

- The last part of the plan shows the low-level details such as the expected amount of data that will be read, where you can judge the effectiveness of your partitioning strategy and estimate how long it will take to scan a table based on total data size and the size of the cluster.
- As you work your way up, next you see the operations that will be parallelized and performed on each Impala node.
- At the higher levels, you see how data flows when intermediate result sets are combined and transmitted from one node to another.
- The EXPLAIN_LEVEL query option lets you customize how much detail to show in the EXPLAIN plan depending on whether you are doing high-level or low-level tuning, dealing with logical or physical aspects of the query.

The example below shows how the standard EXPLAIN output moves from the lowest (physical) level to the higher (logical) levels. The query begins by scanning a certain amount of data; each node performs an aggregation operation (evaluating COUNT(*)) on some subset of data that is local to that node; the intermediate results are transmitted back to the coordinator node (labelled here as the EXCHANGE node); lastly, the intermediate results are summed to display the final result.

```
[impalad-host:21000] > EXPLAIN select count(*) from customer_address;
+------------------------------------------------------------+
| Explain String                                             |
+------------------------------------------------------------+
| Estimated Per-Host Requirements: Memory=42.00MB VCores=1 |
|                                                            |
| 03:AGGREGATE [MERGE FINALIZE]                              |
| |  output: sum(count(*))                                   |
| |                                                          |
| 02:EXCHANGE [PARTITION=UNPARTITIONED]                      |
| |                                                          |
| 01:AGGREGATE                                               |
| |  output: count(*)                                        |
| |                                                          |
| 00:SCAN HDFS [default.customer_address]                    |
|    partitions=1/1 size=5.25MB                              |
+------------------------------------------------------------+
```

The EXPLAIN plan is also printed at the beginning of the query PROFILE report to provide convenience in examining both the logical and physical aspects of the query side-by-side.

The amount of detail displayed in the EXPLAIN output is controlled by the EXPLAIN_LEVEL query option. You customize how much detail to show in the EXPLAIN plan depending on whether you are doing high-level or low-level tuning, dealing with logical or physical aspects of the query. You typically increase this setting from standard to extended when double-checking the presence of table and column statistics during performance tuning, or when estimating query resource usage in conjunction with the resource management features.

Understanding Performance using SUMMARY Report

For an overview of the physical performance characteristics for a query, issue the SUMMARY command in impala-shell immediately after executing a query. This condensed information shows which phases of execution took the most time, and how the estimates for memory usage and number of rows at each phase compare to the actual values.

Like the EXPLAIN plan, it is easy to see potential performance bottlenecks in the SUMMARY report. Like the PROFILE output, the SUMMARY report is available after the query is run, and it displays actual timing statistics.

The SUMMARY report is also printed at the beginning of the query profile report described in the PROFILE output, for convenience in examining high-level and low-level aspects of the query side-by-side.

When the MT_DOP query option is set to a value larger than 0, the #Inst column in the output shows the number of fragment instances. Impala decomposes each query into smaller units of work that are distributed across the cluster, and these units are referred as fragments.

When the MT_DOP query option is set to 0, the #Inst column in the output shows the same value as the #Hosts column, since there is exactly one fragment for each host.

For example, here is a query involving an aggregate function on a single-node. The different stages of the query and their timings are shown (rolled up for all nodes), along with estimated and actual values used in planning the query. In this case, the AVG() function is computed for a subset of data on each node (stage 01) and then the aggregated results from all nodes are combined at the end (stage 03). You can see which stages took the most time, and whether any estimates were substantially different than the actual data distribution. (When examining the time values, be sure to consider the suffixes such as us for microseconds and ms for milliseconds, rather than just looking for the largest numbers.)

```
> SELECT AVG(ss_sales_price) FROM store_sales WHERE ss_coupon_amt = 0;
> SUMMARY;
+--------------+--------+--------+----------+----------+-------+------------
+----------+--------------+-----------------+
| Operator     | #Hosts | #Inst  | Avg Time | Max Time | #Rows | Est. #Rows
 | Peak Mem | Est. Peak Mem | Detail          |
+--------------+--------+--------+----------+----------+-------+------------
+----------+--------------+-----------------+
| 03:AGGREGATE | 1      | 1      | 1.03ms   | 1.03ms   | 1     | 1
 | 48.00 KB | -1 B          | MERGE FINALIZE  |
| 02:EXCHANGE  | 1      | 1      | 0ns      | 0ns      | 1     | 1
 | 0 B      | -1 B          | UNPARTITIONED   |
| 01:AGGREGATE | 1      | 1      |30.79ms   | 30.79ms  | 1     | 1
 | 80.00 KB | 10.00 MB      |                 |
| 00:SCAN HDFS | 1      | 1      | 5.45s    | 5.45s    | 2.21M | -1
| 64.05 MB | 432.00 MB     | tpc.store_sales |
+--------------+--------+--------+----------+----------+-------+------------
+----------+--------------+-----------------+
```

Notice how the longest initial phase of the query is measured in seconds (s), while later phases working on smaller intermediate results are measured in milliseconds (ms) or even nanoseconds (ns).

Understanding Performance using Query Profile

To understand the detailed performance characteristics for a query, issue the PROFILE command in `impala-shell` immediately after executing a query. This low-level information includes physical details about memory, CPU, I/O, and network usage, and thus is only available after the query is actually run.

The PROFILE command, available in `impala-shell`, produces a detailed low-level report showing how a query was executed. Unlike the EXPLAIN plan, this information is only available after the query has finished. It shows physical details such as the number of bytes read, maximum memory usage, and so on for each node. You can use this information to determine if the query is I/O-bound or CPU-bound, whether some network condition is imposing a bottleneck, whether a slowdown is affecting some nodes but not others, and to check that recommended configuration settings such as short-circuit local reads are in effect.

By default, time values in the profile output reflect the wall-clock time taken by an operation. For values denoting system time or user time, the measurement unit is reflected in the metric name, such as ScannerThreadsSysTime or ScannerThreadsUserTime. For example, a multi-threaded I/O operation might show a small figure for wall-clock time, while the corresponding system time is larger, representing the sum of the CPU time taken by each thread. Or a wall-clock time figure might be larger because it counts time spent waiting, while the corresponding system and user time figures only measure the time while the operation is actively using CPU cycles.

The EXPLAIN plan is also printed at the beginning of the query profile report, for convenience in examining both the logical and physical aspects of the query side-by-side. The EXPLAIN_LEVEL query option also controls the verbosity of the EXPLAIN output printed by the PROFILE command.

The Per Node Profiles section in the profile output includes the following metrics that can be controlled by the RESO URCE_TRACE_RATIO query option.

- CpuIoWaitPercentage
- CpuSysPercentage
- CpuUserPercentage
- HostDiskReadThroughput: All data read by the host as part of the execution of this query (spilling), by the HDFS data node, and by other processes running on the same system.
- HostDiskWriteThroughput: All data written by the host as part of the execution of this query (spilling), by the HDFS data node, and by other processes running on the same system.
- HostNetworkRx: All data received by the host as part of the execution of this query, other queries, and other processes running on the same system.
- HostNetworkTx: All data transmitted by the host as part of the execution of this query, other queries, and other processes running on the same system.

The ImpalaServer section of a query profile includes counters that measure client-side latency and internal execution states:

- ClientFetchWaitTimer property indicates the cumulative time that the Impala server spends waiting for the client application to retrieve rows and complete client-side processing.
- ClientFetchWaitTimeStats property provides summary statistics for the client fetch wait metrics.
- ClientFetchLockWaitTimer property measures the cumulative time that client fetch requests wait to acquire internal locks.
- GetInFlightProfileTimeStats property provides summary statistics of the time spent dumping profile records while the query execution is in flight.



Note: Dumping the profile of an in-flight query blocks client requests that fetch results. You can use this counter to monitor and identify potential delays in result retrieval.

A ClientFetchWaitTimer property value that is close to the total query duration indicates a performance bottleneck on the client side. In this scenario, the Impala server remains idle because the system must wait for the external client to fetch the query results.

The query state continues in the FINISHED state during this transfer process. The FINISHED state indicates that the client can begin fetching results from the coordinator node, rather than indicating that the data transfer is complete.

Because Impala uses distributed pipeline execution, the query engine cannot produce or compute subsequent results if the client application does not fetch the already generated rows.

Impala AI query profile analyzer

The Impala AI Query Profile Analyzer evaluates query performance in the Coordinator Web UI, surfacing bottlenecks and root causes such as join skew, memory pressure, and data spills.

Note: This feature is available in Cloudera Runtime 7.3.2.10000 and higher versions.

The Impala AI Query Profile Analyzer processes query profiles directly in the Coordinator Web user interface (UI) to identify performance issues, summarize runtime data, and provide optimization suggestions. Impala query profiles track query planning, execution phases, and resource consumption. However, these profiles can grow very large and highly technical. Impala AI Query Profile Analyzer evaluates complex performance metrics such as scanner wait times, join skew, memory pressure, cardinality issues, or data spills.

Note: Because AI output is inherently probabilistic, correctness is not guaranteed. You must verify conclusions against the raw query profile data.

To keep your experience smooth and secure, the system uses a multi-round tool-calling architecture consisting of the following components:

- Profile redaction component — Replaces sensitive identifiers, such as user names, IP addresses, hostnames, table names, and columns, with secure, consistent aliases before sending data to the large language model (LLM).
- Profile parsing and tool call interface — Breaks raw profile JSON into small segments, allowing the AI agent to request specific pieces, such as timeline details or scan node summaries, as needed.
- AI endpoint communication layer — Connects to your AI models using your existing Impala AI Functions setup. The selected LLM endpoint must support the OpenAI API format for requests and OpenAI-style tool calling. Before displaying the final analysis, this layer restores your original names and identifiers.

Analyzing query execution profiles

Analyze runtime profiles to isolate performance degradation reasons and reduce mean time to resolution (MTTR).

Before you begin

1. Your administrator must enable the required endpoint target flags within your existing Impala AI Functions configuration.
2. You must configure Impala AI Functions. For instructions, see *Impala AI functions*.

Procedure

1. Go to Clusters Impala Service Web UI .
2. Go to the Coordinator Web UI Query Details page for your query of interest.
3. Open the AI Analysis tab.
4. Click the Generate AI Analysis button.

   Review the generated performance summary text in the UI panel.

Results

The AI Analysis tab provides a clear, plain-English summary that shows exactly when your query slowed down and how you can fix it. The automatic unredaction process displays the original table names, columns, and host identifiers directly within the report.

Related Information

Impala AI functions

Planner changes for CPU usage

This release brings some changes to the query planner to improve parallel sizing and resource estimation. This is done for workload-aware autoscaling and will be available as query options. These additional query options are added for tuning purposes. This new functionality will allow additional customers to enable multi-threaded queries globally for improved performance.

Before you begin

- You must obtain the entitlement to use this feature.

Overview

Until this release, estimating the peak memory was the scaling factor in selecting an executor group to run large queries. This estimation was calculated during the query compilation time. When large queries were identified, they were enabled to run on larger executor groups mapped to different resource groups.

This release adds a new CPU costing algorithm as the scaling factor by establishing an infrastructure to allow the weighted total amount of data to process to be used as a new factor in the selection of an executor group. The new costing algorithm returns a number representing an ideal CPU core count required for a query to run efficiently. This number will be compared against the total CPU count of an Impala executor group to determine if it fits to run in that executor group or not. If the CPU requirement is higher than the max-query-cpu-core-per-node-limit, even in the largest exec group, the query may be rejected.

The following query options are added to control this CPU costing algorithm.

- COMPUTE_PROCESSING_COST

  Option to enable this CPU costing algorithm. Set MT_DOP > 0 to enable this query option.

  Note: If entitlement is enabled, COMPUTE_PROCESSING_COST will be set to true by default. If you use the REQUEST_POOL query option to direct a query to a specific pool or queue and you want to maintain the original MT_DOP behavior then you must explicitly set the flag COMPUTE_PROCESSING_COST to False.
- PROCESSING_COST_MIN_THREADS

  Option to control the minimum number of fragment instances (threads) that the costing algorithm is allowed to adjust. The costing algorithm is in charge of increasing the fragment's instance count beyond this minimum number through producer-consumer rate comparison. The maximum number of fragments is max between PROCESSING_COST_MIN_THREADS, MT_DOP, and number of cores per executor.

You can also use the following backend flags to tune the algorithm.

- query_cpu_count_divisor

  Use this flag to divide the CPU requirement of a query to fit the total available CPU in the executor group. For example, setting the value to 2 will fit the query with CPU requirement 2X to an executor group with total available CPU X. Note that setting with a fractional value less than 1 effectively multiplies the query CPU requirement. A valid value is > 0.0. The default value is 1.
- processing_cost_use_equal_expr_weight

  The default value of this flag is True. When the default value is retained, all expression evaluations are weighted equally to 1 during the plan node's processing cost calculation.

  Note: This flag is experimental and the default value should not be changed.
- min_processing_per_thread

  This flag defines the minimum processing load (in processing cost unit) that a fragment instance needs to work on before the planner considers increasing the instance count based on the processing cost rather than the MT_DOP setting. The decision is per fragment. If you set this flag to a higher number, it will reduce parallelism of a fragment (more workload per fragment), and when set to a lower number, it will increase parallelism (less

workload per fragment). Actual parallelism might still be constrained by the total number of cores in the selected executor groups, MT_DOP, or the PROCESSING_COST_MIN_THREADS query option. The value of this flag must be a positive integer and it defaults to 10M.

- skip_resource_checking_on_last_executor_group_set

  This flag defines if the resource checking will be skipped on the last executor group set. If this backend flag is set to true, memory and cpu resource checking will be skipped when a query is being planned against the last (largest) executor group set. Setting true will ensure that query will always get admitted into the last executor group set if it does not fit in any other group set.

Calcite planner in Impala [Preview]

The Apache Calcite planner is a cost-based query optimizer for Impala that evaluates alternative plans, such as different join orders, to choose a more CPU- and memory-efficient execution path.

Note: This feature is in Technical Preview and is not ready for production deployments. recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

This feature is available in Cloudera Runtime 7.3.2.10000 and higher. Impala compiles each query with either its original planner or the Apache Calcite planner. The Calcite planner applies cost-based optimization to select an execution plan. It is disabled by default, and you enable it per session with a query option.

How the Calcite planner works

When you set the PLANNER query option to calcite, Impala compiles queries with the Calcite planner. By default, both the PLANNER and FALLBACK_PLANNER query options are set to original, so queries run on the original planner. When you change PLANNER to calcite and leave FALLBACK_PLANNER at its default, Impala falls back to the original planner whenever the Calcite planner cannot compile a query.

If the Calcite planner cannot compile a query, the FALLBACK_PLANNER query option and the cause of the failure together determine whether Impala falls back to the original planner.

Impala always falls back to the original planner in the following cases, regardless of the FALLBACK_PLANNER value:

- Data Definition Language (DDL) statements and other statements that are not `SELECT` queries.
- Queries that use features the Calcite planner does not yet support, such as complex types.

For all other compilation failures, the FALLBACK_PLANNER value determines the behavior. The following table describes the supported values:

Value	Description
original	(default) **Impala falls back to the original planner when the Calcite planner cannot compile a query.**
none	Impala does not fall back to a different planner. This value behaves as though FALLBACK_PLANNER matches PLANNER. When PLANNER is calcite, the none and calcite values produce the same result.
calcite	Impala does not fall back for unexpected compilation failures, and the affected query fails instead.

Important: Cloudera recommends keeping the FALLBACK_PLANNER query option set to original in production environments. Set FALLBACK_PLANNER to calcite only for testing, because Impala then does not fall back to the original planner for unexpected compilation failures and the affected queries fail.

When to use the Calcite planner

Consider enabling it in the following cases:

- Complex queries, such as those with many joins, large multi-table joins, or deeply nested subqueries, where cost-based join ordering is most likely to improve performance.
- Queries that use constructs the original planner does not support, such as certain correlated subqueries in the `SELECT` list.
- Short or simple queries might not benefit from the Calcite planner and can continue to use the original planner, which is the default.

Enabling the Calcite planner [Preview]

Enable the Apache Calcite planner for a session by setting the PLANNER query option, so that Impala compiles queries with cost-based optimization.

About this task

The Calcite planner is disabled by default. Set the PLANNER query option to CALCITE to enable it for the current session.

Procedure

1. Set the PLANNER query option to CALCITE.

   ```
   SET PLANNER=CALCITE;
   ```

2. Set the FALLBACK_PLANNER query option to specify the planner Impala uses if the Calcite planner fails to compile a query.

   ```
   SET FALLBACK_PLANNER=ORIGINAL;
   ```

3. Run your query.

   You can also set the query option globally for all queries by configuring the default query options property in Cloudera Manager.

   a. Go to Clusters Impala Configuration .
   b. Search for Default Query Options.
   c. Enter the query option and value by using the following format:

      ```
      OPTION_NAME=VALUE
      ```

      For multiple options, separate each key-value pair with a comma.
   d. Save the changes.
   e. Restart the Impala service to apply the updated configuration.

   Impala compiles the query with the Calcite planner. If compilation fails and a different fallback planner is set, Impala compiles the query with the fallback planner.

Results

The session uses the Calcite planner for subsequent queries until you change the PLANNER query option.

Planner changes to improve cardinality estimation

Changes have been implemented in the query planner to improve cardinality estimation, constituting a pivotal element of workload-aware autoscaling.

The Planner changes to improve cardinality estimation feature is available from Cloudera Runtime 7.3.1.500 SP3

In earlier releases, Impala would generate a plan initially and then search for runtime filters based on the entire plan. In this release, selective runtime filters have been integrated. These filters serve to diminish the cardinality estimates

of scan nodes and specific join nodes situated above them. This adjustment occurs after the generation of runtime filters and before the computation of resource requirements.

The cardinality reduction is implemented across all execution modes, but its effectiveness is currently contingent on the COMPUTE_PROCESSING_COST option being set to True. This condition exists because the optimal benefits of diminished scan cardinality are realized in scenarios where multiple executor group setups are employed. Reduced scan cardinality estimation has the potential to decrease Processing Costs, diminish scan fragment parallelism, and enhance the likelihood of assigning queries to smaller executor group sets. It's worth noting that other execution modes, such as MT_DOP>0 or the legacy scanner parallelism mode, will not experience any changes in their execution plans.

Before you begin

- You must obtain the entitlement to use this feature.

Example

The following example demonstrates the execution of TPC-DS Query 3 at a scale factor of 1, utilizing 3 executor nodes, with the COMPUTE_PROCESSING_COST parameter set to True.

```
 Query:
   use tpcds_partitioned_parquet_snap;
   set COMPUTE_PROCESSING_COST=true;
   set PROCESSING_COST_MIN_THREADS=2;
   set MAX_FRAGMENT_INSTANCES_PER_NODE=16;

   -- TPCDS-Q3
   select dt.d_year,
   item.i_brand_id brand_id,
   item.i_brand brand,
   sum(ss_ext_sales_price) sum_agg
   from
   date_dim dt,
   store_sales,
   item
   where
   dt.d_date_sk = store_sales.ss_sold_date_sk
   and store_sales.ss_item_sk = item.i_item_sk
   and item.i_manufact_id = 436
   and dt.d_moy = 12
   group by
   dt.d_year,
   item.i_brand,
   item.i_brand_id
   order by
   dt.d_year,
   sum_agg desc,
   brand_id
   limit 100

Plan:
Max Per-Host Resource Reservation: Memory=33.25MB Threads=14
Per-Host Resource Estimates: Memory=169MB
Analyzed query: SELECT dt.d_year, item.i_brand_id brand_id, item.i_brand
brand,
sum(ss_ext_sales_price) sum_agg FROM tpcds_partitioned_parquet_snap.date_
dim dt,
tpcds_partitioned_parquet_snap.store_sales, tpcds_partitioned_parquet_snap.i
tem
WHERE dt.d_date_sk = store_sales.ss_sold_date_sk AND store_sales.ss_item_sk
=
item.i_item_sk AND item.i_manufact_id = CAST(436 AS INT) AND dt.d_moy = CA
ST(12
AS INT) GROUP BY dt.d_year, item.i_brand, item.i_brand_id ORDER BY dt.d_year
```

```
ASC, sum_agg DESC, brand_id ASC LIMIT CAST(100 AS TINYINT)

F05:PLAN FRAGMENT [UNPARTITIONED] hosts=1 instances=1
|  Per-Instance Resources: mem-estimate=4.03MB mem-reservation=4.00MB threa
d-reservation=1
|  max-parallelism=1 segment-costs=[406] cpu-comparison-result=18 [max(1 (se
lf) vs 18 (sum children))]
PLAN-ROOT SINK
|  output exprs: dt.d_year, item.i_brand_id, item.i_brand, sum(ss_ext_sale
s_price)
|  mem-estimate=4.00MB mem-reservation=4.00MB spill-buffer=2.00MB thread-re
servation=0 cost=400
|
12:MERGING-EXCHANGE [UNPARTITIONED]
|  order by: dt.d_year ASC, sum(ss_ext_sales_price) DESC, item.i_brand_id
 ASC
|  limit: 100
|  mem-estimate=34.63KB mem-reservation=0B thread-reservation=0
|  tuple-ids=4 row-size=52B cardinality=100 cost=6
|  in pipelines: 06(GETNEXT)
|
F04:PLAN FRAGMENT [HASH(dt.d_year,item.i_brand,item.i_brand_id)] hosts=3 in
stances=6 (adjusted from 48)
Per-Instance Resources: mem-estimate=10.38MB mem-reservation=1.94MB thread-
reservation=1
max-parallelism=6 segment-costs=[12331, 300, 6] cpu-comparison-result=18 [
max(6 (self) vs 18 (sum children))]
06:TOP-N [LIMIT=100]
|  order by: dt.d_year ASC, sum(ss_ext_sales_price) DESC, item.i_brand_id A
SC
|  mem-estimate=5.10KB mem-reservation=0B thread-reservation=0
|  tuple-ids=4 row-size=52B cardinality=100 cost=300
|  in pipelines: 06(GETNEXT), 11(OPEN)
|
11:AGGREGATE [FINALIZE]
|  output: sum:merge(ss_ext_sales_price)
|  group by: dt.d_year, item.i_brand, item.i_brand_id
|  mem-estimate=10.00MB mem-reservation=1.94MB spill-buffer=64.00KB thread-
reservation=0
|  tuple-ids=3 row-size=52B cardinality=3.04K cost=12164
|  in pipelines: 11(GETNEXT), 01(OPEN)
|
10:EXCHANGE [HASH(dt.d_year,item.i_brand,item.i_brand_id)]
|  mem-estimate=388.85KB mem-reservation=0B thread-reservation=0
|  tuple-ids=3 row-size=52B cardinality=3.04K cost=167
|  in pipelines: 01(GETNEXT)
|
F03:PLAN FRAGMENT [HASH(store_sales.ss_sold_date_sk)] hosts=3 instances=6 (a
djusted from 48)
Per-Instance Resources: mem-estimate=12.17MB mem-reservation=2.00MB thread-
reservation=1
max-parallelism=6 segment-costs=[15408, 167] cpu-comparison-result=18 [max
(18 (self) vs 11 (sum children))]
05:AGGREGATE [STREAMING]
|  output: sum(ss_ext_sales_price)
|  group by: dt.d_year, item.i_brand, item.i_brand_id
|  mem-estimate=10.00MB mem-reservation=2.00MB spill-buffer=64.00KB thread-r
eservation=0
|  tuple-ids=3 row-size=52B cardinality=3.04K cost=12164
|  in pipelines: 01(GETNEXT)
|
04:HASH JOIN [INNER JOIN, PARTITIONED]
|  hash-table-id=00
|  hash predicates: store_sales.ss_sold_date_sk = dt.d_date_sk
```

```
|  fk/pk conjuncts: store_sales.ss_sold_date_sk = dt.d_date_sk
|  mem-estimate=0B mem-reservation=0B spill-buffer=64.00KB thread-reservat
ion=0
|  tuple-ids=1,2,0 row-size=72B cardinality=3.04K cost=3041
|  in pipelines: 01(GETNEXT), 00(OPEN)
|
|--F06:PLAN FRAGMENT [HASH(store_sales.ss_sold_date_sk)] hosts=3 instance
s=6 (adjusted from 48)
|  |  Per-Instance Resources: mem-estimate=3.02MB mem-reservation=2.94MB
thread-reservation=1 runtime-filters-memory=1.00MB
|  |  max-parallelism=6 segment-costs=[6183]
|  JOIN BUILD
|  |  join-table-id=00 plan-id=01 cohort-id=01
|  |  build expressions: dt.d_date_sk
|  |  runtime filters: RF000[bloom] <- dt.d_date_sk, RF001[min_max] <- dt
.d_date_sk
|  |  mem-estimate=1.94MB mem-reservation=1.94MB spill-buffer=64.00KB thr
ead-reservation=0 cost=6087
|  |
|  09:EXCHANGE [HASH(dt.d_date_sk)]
|  |  mem-estimate=87.33KB mem-reservation=0B thread-reservation=0
|  |  tuple-ids=0 row-size=12B cardinality=6.09K cost=96
|  |  in pipelines: 00(GETNEXT)
|  |
|  F02:PLAN FRAGMENT [RANDOM] hosts=1 instances=1
|  Per-Instance Resources: mem-estimate=16.38MB mem-reservation=512.00KB thr
ead-reservation=1
|  max-parallelism=1 segment-costs=[124002]
|  00:SCAN HDFS [tpcds_partitioned_parquet_snap.date_dim dt, RANDOM]
|     HDFS partitions=1/1 files=1 size=2.15MB
|     predicates: dt.d_moy = CAST(12 AS INT)
|     stored statistics:
|       table: rows=73.05K size=2.15MB
|       columns: all
|     extrapolated-rows=disabled max-scan-range-rows=73.05K
|     parquet statistics predicates: dt.d_moy = CAST(12 AS INT)
|     parquet dictionary predicates: dt.d_moy = CAST(12 AS INT)
|     mem-estimate=16.00MB mem-reservation=512.00KB thread-reservation=0
|     tuple-ids=0 row-size=12B cardinality=6.09K cost=123906
|     in pipelines: 00(GETNEXT)
|
08:EXCHANGE [HASH(store_sales.ss_sold_date_sk)]
|  mem-estimate=877.96KB mem-reservation=0B thread-reservation=0
|  tuple-ids=1,2 row-size=60B cardinality=3.04K cost=203
|  in pipelines: 01(GETNEXT)
|
F00:PLAN FRAGMENT [RANDOM] hosts=3 instances=12 (adjusted from 48)
Per-Host Shared Resources: mem-estimate=2.00MB mem-reservation=2.00MB thr
ead-reservation=0 runtime-filters-memory=2.00MB
Per-Instance Resources: mem-estimate=17.60MB mem-reservation=1.00MB thread-r
eservation=1
max-parallelism=12 segment-costs=[91203296]
03:HASH JOIN [INNER JOIN, BROADCAST]
|  hash-table-id=01
|  hash predicates: store_sales.ss_item_sk = item.i_item_sk
|  fk/pk conjuncts: store_sales.ss_item_sk = item.i_item_sk
|  mem-estimate=0B mem-reservation=0B spill-buffer=64.00KB thread-reservat
ion=0
|  tuple-ids=1,2 row-size=60B cardinality=3.04K cost=3045
|  in pipelines: 01(GETNEXT), 02(OPEN)
|
|--F07:PLAN FRAGMENT [RANDOM] hosts=3 instances=3
|  |  Per-Instance Resources: mem-estimate=8.77MB mem-reservation=8.75MB
thread-reservation=1 runtime-filters-memory=1.00MB
```

```
|  |  max-parallelism=3 segment-costs=[22]
|  JOIN BUILD
|  |  join-table-id=01 plan-id=02 cohort-id=01
|  |  build expressions: item.i_item_sk
|  |  runtime filters: RF002[bloom] <- item.i_item_sk
|  |  mem-estimate=7.75MB mem-reservation=7.75MB spill-buffer=64.00KB threa
d-reservation=0 cost=19
|  |
|  07:EXCHANGE [BROADCAST]
|  |  mem-estimate=16.00KB mem-reservation=0B thread-reservation=0
|  |  tuple-ids=2 row-size=44B cardinality=19 cost=3
|  |  in pipelines: 02(GETNEXT)
|  |
|  F01:PLAN FRAGMENT [RANDOM] hosts=1 instances=1
|  Per-Instance Resources: mem-estimate=16.19MB mem-reservation=256.00KB
thread-reservation=1
|  max-parallelism=1 segment-costs=[68778]
|  02:SCAN HDFS [tpcds_partitioned_parquet_snap.item, RANDOM]
|     HDFS partitions=1/1 files=1 size=1.73MB
|     predicates: item.i_manufact_id = CAST(436 AS INT)
|     stored statistics:
|       table: rows=18.00K size=1.73MB
|       columns: all
|     extrapolated-rows=disabled max-scan-range-rows=18.00K
|     parquet statistics predicates: item.i_manufact_id = CAST(436 AS INT)
|     parquet dictionary predicates: item.i_manufact_id = CAST(436 AS INT)
|     mem-estimate=16.00MB mem-reservation=256.00KB thread-reservation=0
|     tuple-ids=2 row-size=44B cardinality=19 cost=68777
|     in pipelines: 02(GETNEXT)
|
01:SCAN HDFS [tpcds_partitioned_parquet_snap.store_sales, RANDOM]
   HDFS partitions=1824/1824 files=1824 size=199.36MB
   runtime filters: RF001[min_max] -> store_sales.ss_sold_date_sk, RF000[
bloom] -> store_sales.ss_sold_date_sk, RF002[bloom] -> store_sales.ss_item_s
k
   stored statistics:
     table: rows=2.88M size=199.36MB
     partitions: 1824/1824 rows=2.88M
     columns: all
   extrapolated-rows=disabled max-scan-range-rows=130.09K
   mem-estimate=16.00MB mem-reservation=1.00MB thread-reservation=0
   tuple-ids=1 row-size=16B cardinality=3.04K(filtered from 2.88M) cost=9
1200048
   in pipelines: 01(GETNEXT)
```

Caching Codegen Functions

In Impala, "codegen" refers to code generation, utilizing query-specific information to generate specialized machine code for each query.

The codegen capability feature is available from Cloudera Runtime 7.3.1.500 SP3

When executing a standard query, the query optimizer generates an optimized query plan, passing it to the executor for processing. The codegen capability converts query-specific information into machine code for faster execution. However, for sub-second queries, codegen may consume excessive time; for instance, the compilation time for codegen could take 300ms for a sub-second query. To mitigate this, Cloudera introduces support for memory caches for codegen functions, enhancing the performance of sub-second queries. Users can specify the total cache size for all codegen functions. The configuration parameter 'codegen_cache_capacity' in Impala daemons controls the total codegen cache size. For example, when set to "1GB," it restricts the codegen cache size for the particular Impala daemon to one gigabyte.

The cache is a singleton instance for each daemon, housing multiple cache entries. Each cache entry operates at the fragment level and retains all codegen functions associated with a fragment. Upon reaching maximum capacity, the cache employs an LRU (Least Recently Used) eviction policy. Storing codegen functions in the cache enables their reuse when appropriate, reducing the need for repetitive codegen compilation, which may otherwise consume hundreds of milliseconds.

The LLVM module bitcode serves as the cache key, generated prior to LLVM module optimization and final compilation. When codegen_cache_mode is set to NORMAL, the complete bitcode string is stored as the key by default. Conversely, if codegen_cache_mode is set to OPTIMAL, only a key containing the hash code derived from the NORMAL mode key is stored, reducing memory consumption.

The memory allocated for the codegen cache operates independently of the admission memory reserved for query execution. While the codegen cache memory is initially limited to a certain percentage of the total process memory, adjustments can be made during startup if it exceeds that percentage.

You can deactivate the codegen capability by setting the query option disable_codegen_cache to true or by setting the flag file option 'codegen_cache_capacity' to 0. It is important to note that even if disable_codegen_cache is set to 'false', caching will still be disabled if you set codegen_cache_capacity to 0.

Scalability Considerations

The size of your cluster and the volume of data influences query performance. Typically, adding more cluster capacity reduces problems due to memory limits or disk throughput. On the other hand, larger clusters are more likely to have other kinds of scalability issues, such as a single slow node that causes performance problems for queries.

Impact of Many Tables or Partitions on Impala Catalog Performance and Memory Usage

Because Hadoop I/O is optimized for reading and writing large files, Impala is optimized for tables containing relatively few, large data files. Schemas containing thousands of tables, or tables containing thousands of partitions, can encounter performance issues during startup or during DDL operations such as ALTER TABLE statements.

Limiting file metadata fetched in Catalog requests

To prevent Catalog service (Catalogd) Out-of-Memory (OOM) errors when coordinator fetching metadata for tables with millions of files, the new configuration flag catalog_partial_fetch_max_files has been introduced.

This flag limits the maximum number of file descriptors returned in a single Catalog fetch response. This response is for the GetPartialCatalogObject RPC, which is used in local catalog mode. See *Automatic Invalidation/Refresh of Metadata*.

Default: 1,000,000 files

If a request exceeds this limit, Catalogd truncates the response at the partition level. The Impala coordinator then automatically sends subsequent requests to fetch the remaining metadata, and it detects any version changes to force a query replan, ensuring metadata consistency.

Scalability Considerations for the Impala StateStore

If it takes a very long time for a cluster to start up with the message, This Impala daemon is not ready to accept user requests, the StateStore might be taking too long to send the entire catalog topic to the cluster. In this case, consider setting the Load Catalog in Background field to false in your Catalog Service configuration. This setting stops the StateStore from loading the entire catalog into memory at cluster startup. Instead, metadata for each table is loaded when the table is accessed for the first time.

Effect of Buffer Pool on Memory Usage

Most of the memory needed is reserved at the beginning of the query, avoiding cases where a query might run for a long time before failing with an out-of-memory error. The actual memory estimates and memory buffers are typically smaller than before, so that more queries can run concurrently or process larger volumes of data than previously.

Increase the MAX_ROW_SIZE query option setting when querying tables with columns containing long strings, many columns, or other combinations of factors that produce very large rows. If Impala encounters rows that are too large to process with the default query option settings, the query fails with an error message suggesting to increase the MAX_ROW_SIZE setting.

SQL Operations that Spill to Disk

Certain memory-intensive operations write temporary data to disk (known as *spilling* to disk) when Impala is close to exceeding its memory limit on a particular host.

What kinds of queries might spill to disk:

Several SQL clauses and constructs require memory allocations that could trigger spilling to disk:

- when a query uses a GROUP BY clause for columns with millions or billions of distinct values, Impala keeps a similar number of temporary results in memory, to accumulate the aggregate results for each value in the group.
- When large tables are joined together, Impala keeps the values of the join columns from one table in memory, to compare them to incoming values from the other table.
- When a large result set is sorted by the ORDER BY clause, each node sorts its portion of the result set in memory.
- The DISTINCT and UNION operators build in-memory data structures to represent all values found so far, to eliminate duplicates as the query progresses.

When the spill-to-disk feature is activated for a join node within a query, Impala does not produce any runtime filters for that join operation on that host. Other join nodes within the query are not affected.

The amount data depends on the portion of the data being handled by that host, and thus the operator may end up consuming different amounts of memory on different hosts.

How Impala handles scratch disk space for spilling:

Impala uses intermediate files during large sorts, joins, aggregations, or analytic function operations The files are removed when the operation finishes. By default, intermediate files are stored in the directory /tmp/impala-scratch.

You can specify locations of the intermediate files in one of the following ways:

- By starting the `impalad` daemon with the ##scratch_dirs="*PATH_TO_DIRECTORY*" configuration option.
- By specifying a different location in the Cloudera Manager in the Impala Daemon Scratch Directories field.

With either option above:

- You can specify a single directory or a comma-separated list of directories.
- You can specify an optional a capacity quota per scratch directory using the colon (:) as the delimiter.

  The capacity quota of -1 or 0 is the same as no quota for the directory.
- The scratch directories must be on the local filesystem, not in HDFS.
- You might specify different directory paths for different hosts, depending on the capacity and speed of the available storage devices.

If there is less than 1 GB free on the filesystem where that directory resides, Impala still runs, but writes a warning message to its log.

Impala successfully starts (with a warning written to the log) if it cannot create or read and write files in one of the scratch directories.

The following are examples for specifying scratch directories.

Config option	Description
--scratch_dirs=/dir1,/dir2	Use /dir1 and /dir2 as scratch directories with no capacity quota.

Config option	Description
--scratch_dirs=/dir1,/dir2:25G	Use /dir1 and /dir2 as scratch directories with no capacity quota on /dir1 and the 25GB quota on /dir2.
--scratch_dirs=/dir1:5MB,/dir2	Use /dir1 and /dir2 as scratch directories with the capacity quota of 5MB on /dir1 and no quota on /dir2.
--scratch_dirs=/dir1:-1,/dir2:0	Use /dir1 and /dir2 as scratch directories with no capacity quota.

Allocation from a scratch directory will fail if the specified limit for the directory is exceeded.

Tip: Ensure the scratch directory size aligns with your available disk capacity. We recommend setting the scratch directory size to 100GB or larger, with 10GB being the absolute minimum. For example, you can configure it as --scratch_dirs=/dir1:40GB,/dir2:60GB to allocate a total of 100GB. The scratch directory size is particularly important for large queries that frequently spill data to disk. Insufficient capacity can result in a SCRATCH_ALLOCATION_FAILED error, causing query failures. To avoid this, adjust the scratch directory size based on your workload requirements.

If Impala encounters an error reading or writing files in a scratch directory during a query, Impala logs the error, and the query fails.

Memory usage for SQL operators:

The memory required to spill to disk is reserved up front, and you can examine it in the EXPLAIN plan when the EXPLAIN_LEVEL query option is set to 2 or higher.

If an operator accumulates more data than can fit in the reserved memory, it can either reserve more memory to continue processing data in memory or start spilling data to temporary scratch files on disk. Thus, operators with spill-to-disk support can adapt to different memory constraints by using however much memory is available to speed up execution, yet tolerate low memory conditions by spilling data to disk.

The amount of data depends on the portion of the data being handled by that host, and thus the operator may end up consuming different amounts of memory on different hosts.

Avoiding queries that spill to disk:

Because the extra I/O can impose significant performance overhead on these types of queries, try to avoid this situation by using the following steps:

1. Detect how often queries spill to disk, and how much temporary data is written. Refer to the following sources:
   - The output of the PROFILE command in the `impala-shell` interpreter. This data shows the memory usage for each host and in total across the cluster. The WriteIoBytes counter reports how much data was written to disk for each operator during the query.
   - In Impala Queries in Cloudera Manager, you can see the peak memory usage for a query, combined across all nodes in the cluster.
   - In the Queries tab in the Impala debug web user interface, select the query to examine and click the corresponding Profile link. This data breaks down the memory usage for a single host within the cluster, the host whose web interface you are connected to.
2. Use one or more techniques to reduce the possibility of the queries spilling to disk:
   - Increase the Impala memory limit if practical. For example, using the SET MEM_LIMIT SQL statement, increase the available memory by more than the amount of temporary data written to disk on a particular node.
   - Increase the number of nodes in the cluster, to increase the aggregate memory available to Impala and reduce the amount of memory required on each node.
   - On a cluster with resources shared between Impala and other Hadoop components, use resource management features to allocate more memory for Impala.
   - If the memory pressure is due to running many concurrent queries rather than a few memory-intensive ones, consider using the Impala admission control feature to lower the limit on the number of concurrent queries. By spacing out the most resource-intensive queries, you can avoid spikes in memory usage and improve overall response times.
   - Tune the queries with the highest memory requirements, using one or more of the following techniques:

- Run the COMPUTE STATS statement for all tables involved in large-scale joins and aggregation queries.
- Minimize your use of STRING columns in join columns. Prefer numeric values instead.
- Examine the EXPLAIN plan to understand the execution strategy being used for the most resource-intensive queries.
- If Impala still chooses a suboptimal execution strategy even with statistics available, or if it is impractical to keep the statistics up to date for huge or rapidly changing tables, add hints to the most resource-intensive queries to select the right execution strategy.

- If your queries experience substantial performance overhead due to spilling, enable the DISABLE_UNSAFE_SPILLS query option. This option prevents queries whose memory usage is likely to be exorbitant from spilling to disk. As you tune problematic queries using the preceding steps, fewer and fewer will be cancelled by this option setting.

When to use DISABLE_UNSAFE_SPILLS:

The DISABLE_UNSAFE_SPILLS query option is suitable for an environment with ad hoc queries whose performance characteristics and memory usage are not known in advance. It prevents “worst-case scenario” queries that use large amounts of memory unnecessarily. Thus, you might turn this option on within a session while developing new SQL code, even though it is turned off for existing applications.

Organizations where table and column statistics are generally up-to-date might leave this option turned on all the time, again to avoid worst-case scenarios for untested queries or if a problem in the ETL pipeline results in a table with no statistics. Turning on DISABLE_UNSAFE_SPILLS lets you “fail fast” in this case and immediately gather statistics or tune the problematic queries.

Some organizations might leave this option turned off. For example, you might have tables large enough that the COMPUTE STATS takes substantial time to run, making it impractical to re-run after loading new data. If you have examined the EXPLAIN plans of your queries and know that they are operating efficiently, you might leave DISABLE_UNSAFE_SPILLS turned off. In that case, you know that any queries that spill will not go overboard with their memory consumption.

Limits on Query Size and Complexity

There are hard-coded limits on the maximum size and complexity of queries. Currently, the maximum number of expressions in a query is 2000. You might exceed the limits with large or deeply nested queries produced by business intelligence tools or other query generators.

If you have the ability to customize such queries or the query generation logic that produces them, replace sequences of repetitive expressions with single operators such as IN or BETWEEN that can represent multiple values or ranges. For example, instead of a large number of OR clauses:

```
WHERE val = 1 OR val = 2 OR val = 6 OR val = 100 ...
```

use a single IN clause:

```
WHERE val IN (1,2,6,100,...)
```

Scalability Considerations for Impala I/O

Impala parallelizes its I/O operations aggressively, therefore the more disks you can attach to each host, the better. Impala retrieves data from disk so quickly using bulk read operations on large blocks, that most queries are CPU-bound rather than I/O-bound.

Because the kind of sequential scanning typically done by Impala queries does not benefit much from the random-access capabilities of SSDs, spinning disks typically provide the most cost-effective kind of storage for Impala data, with little or no performance penalty as compared to SSDs.

Resource management features such as YARN, Llama, and admission control typically constrain the amount of memory, CPU, or overall number of queries in a high-concurrency environment. Currently, there is no throttling mechanism for Impala I/O.

Scalability Considerations for Table Layout

Due to the overhead of retrieving and updating table metadata in the Metastore database, try to limit the number of columns in a table to a maximum of approximately 2000. Although Impala can handle wider tables than this, the Metastore overhead can become significant, leading to query performance that is slower than expected based on the actual data volume.

To minimize overhead related to the Metastore database and Impala query planning, try to limit the number of partitions for any partitioned table to a few tens of thousands.

If the volume of data within a table makes it impractical to run exploratory queries, consider using the TABLESAM PLE clause to limit query processing to only a percentage of data within the table. This technique reduces the overhead for query startup, I/O to read the data, and the amount of network, CPU, and memory needed to process intermediate results during the query.

Kerberos-Related Network Overhead for Large Clusters

When Impala starts up, or after each kinit refresh, Impala sends a number of simultaneous requests to the KDC. For a cluster with 100 hosts, the KDC might be able to process all the requests within roughly 5 seconds. For a cluster with 1000 hosts, the time to process the requests would be roughly 500 seconds. Impala also makes a number of DNS requests at the same time as these Kerberos-related requests.

While these authentication requests are being processed, any submitted Impala queries will fail. During this period, the KDC and DNS may be slow to respond to requests from components other than Impala, so other secure services might be affected temporarily.

Avoiding CPU Hotspots for HDFS Cached Data

You can use the HDFS caching feature to reduce I/O and memory-to-memory copying for frequently accessed tables or partitions.

To avoid hotspots, include the WITH REPLICATION clause with the CREATE TABLE or ALTER TABLE statements for tables that use HDFS caching. This clause allows more than one host to cache the relevant data blocks, so the CPU load can be shared, reducing the load on any one host.

The work for HDFS cached data is divided better among all the hosts that have cached replicas for a particular data block. When more than one host has a cached replica for a data block, Impala assigns the work of processing that block to whichever host has done the least work (in terms of number of bytes read) for the current query. If hotspots persist even with this load-based scheduling algorithm, you can enable the query option SCHEDULE_RANDOM_ REPLICA=TRUE to further distribute the CPU load. This setting causes Impala to randomly pick a host to process a cached data block if the scheduling algorithm encounters a tie when deciding which host has done the least work.

Scalability Considerations for File Handle Caching

One scalability aspect that affects heavily loaded clusters is the load on the metadata layer from looking up the details as each file is opened. On HDFS, that can lead to increased load on the NameNode, and on S3, this can lead to an excessive number of S3 metadata requests. For example, a query that does a full table scan on a partitioned table may need to read thousands of partitions, each partition containing multiple data files. Accessing each column of a Parquet file also involves a separate “open” call, further increasing the load on the NameNode. High NameNode overhead can add startup time (that is, increase latency) to Impala queries, and reduce overall throughput for non-Impala workloads that also require accessing HDFS files.

You can reduce the number of calls made to your file system's metadata layer by enabling the file handle caching feature. Data files that are accessed by different queries, or even multiple times within the same query, can be accessed without a new “open” call and without fetching the file details multiple times.

Impala supports file handle caching for the following file systems:

- HDFS

The cache_remote_file_handles flag controls local and remote file handle caching for an impalad. It is recommended that you use the default value of true as this caching prevents your NameNode from overloading when your cluster has many remote HDFS reads.

- S3

  The cache_s3_file_handles impalad flag controls the S3 file handle caching. The feature is enabled by default with the flag set to true.

- Ozone

  The cache_ozone_file_handles flag controls the Ozone file handle caching. This feature is ebnabled by default with the flag set to true.

The feature is enabled by default with 20,000 file handles to be cached. To change the value, set the configuration option Maximum Cached File Handles (max_cached_file_handles) to a non-zero value for each Impala daemon (impalad). From the initial default value of 20000, adjust upward if NameNode request load is still significant, or downward if it is more important to reduce the extra memory usage on each host. Each cache entry consumes 6 KB, meaning that caching 20,000 file handles requires up to 120 MB on each Impala executor. The exact memory usage varies depending on how many file handles have actually been cached; memory is freed as file handles are evicted from the cache.

If a manual operation moves a file to the trashcan while the file handle is cached, Impala still accesses the contents of that file. This is a change from prior behavior. Previously, accessing a file that was in the trashcan would cause an error. This behavior only applies to non-Impala methods of removing files, not the Impala mechanisms such as TRUN CATE TABLE or DROP TABLE.

If files are removed, replaced, or appended by operations outside of Impala, the way to bring the file information up to date is to run the REFRESH statement on the table.

File handle cache entries are evicted as the cache fills up, or based on a timeout period when they have not been accessed for some time.

To evaluate the effectiveness of file handle caching for a particular workload, issue the PROFILE statement in `impala-shell` or examine query profiles in the Impala Web UI. Look for the ratio of CachedFileHandlesHitCount (ideally, should be high) to CachedFileHandlesMissCount (ideally, should be low). Before starting any evaluation, run several representative queries to “warm up” the cache because the first time each data file is accessed is always recorded as a cache miss.

To see metrics about file handle caching for each `impalad` instance, examine the following fields on the /metrics page in the Impala Web UI:

- impala-server.io.mgr.cached-file-handles-miss-count
- impala-server.io.mgr.num-cached-file-handles

Scaling Limits and Guidelines

Consider and respect the scalability limitations provided in this topic in order to achieve optimal scalability and performance in Impala. For example, while you might be able to create a table with 2000 columns, you will experience performance problems while querying the table. This topic does not cover functional limitations in Impala.

Unless noted otherwise, the limits listed in this topic were tested and certified.

The limits noted as "generally safe" are not certified, but recommended as generally safe. A safe range is not a hard limit as unforeseen errors or troubles in your particular environment can affect the range.

Deployment Limits

- Number of Impalad Executors: 500 nodes depending on use case
- Number of Impalad Coordinators: 1 coordinator for at most every 50 executors

Data Storage Limits

There are no hard limits for the following, but you will experience gradual performance degradation as you increase these numbers.

- Number of databases
- Number of tables - total, per database
- Number of partitions - total, per table
- Number of files - total, per table, per table per partition
- Number of views - total, per database
- Number of user-defined functions - total, per database
- Parquet
  - Number of columns per row group
  - Number of row groups per block
  - Number of HDFS blocks per file

Schema Design Limits

- Number of columns
  - 300 for Kudu tables
  - 1000 for other types of tables

Security Limits

- Number of roles: 10,000

Query Limits - Compile Time

- Maximum number of columns in a query, included in a SELECT list, INSERT, and in an expression: no limit
- Number of tables referenced: no limit
- Number of plan nodes: no limit
- Number of plan fragments: no limit
- Depth of expression tree: 1000 hard limit
- Width of expression tree: 10,000 hard limit

- Codegen: Very deeply nested expressions within queries can exceed internal Impala limits, leading to excessive memory usage. Setting the query option disable_codegen=true may reduce the impact, at a cost of longer query runtime.

Dedicated Coordinator

When scalability bottlenecks occur, you can assign one dedicated role to each Impala daemon host, either as a coordinator or as an executor, to address the issues.

Each host that runs the Impala Daemon acts as both a coordinator and as an executor, by default, managing metadata caching, query compilation, and query execution. In this configuration, Impala clients can connect to any Impala daemon and send query requests.

During highly concurrent workloads for large-scale queries, the dual roles can cause scalability issues because:

- The extra work required for a host to act as the coordinator could interfere with its capacity to perform other work for the later phases of the query. For example, coordinators can experience significant network and CPU overhead with queries containing a large number of query fragments. Each coordinator caches metadata for all table partitions and data files, which requires coordinators to be configured with a large JVM heap. Executor-

only Impala daemons should be configured with the default JVM heaps, which leaves more memory available to process joins, aggregations, and other operations performed by query executors.
- Having a large number of hosts act as coordinators can cause unnecessary network overhead, or even timeout errors, as each of those hosts communicates with the StateStore daemon for metadata updates.
- The "soft limits" imposed by the admission control feature are more likely to be exceeded when there are a large number of heavily loaded hosts acting as coordinators.

The following factors can further exacerbate the above issues:

- High number of concurrent query fragments due to query concurrency and/or query complexity
- Large metadata topic size related to the number of partitions/files/blocks
- High number of coordinator nodes
- High number of coordinators used in the same resource pool

If such scalability bottlenecks occur, you can assign one dedicated role to each Impala daemon host, either as a coordinator or as an executor, to address the issues.

- All explicit or load-balanced client connections must go to the coordinator hosts. These hosts perform the network communication to keep metadata up-to-date and route query results to the appropriate clients. The dedicated coordinator hosts do not participate in I/O-intensive operations such as scans, and CPU-intensive operations such as aggregations.
- The executor hosts perform the intensive I/O, CPU, and memory operations that make up the bulk of the work for each query. The executors do communicate with the Statestored daemon for membership status, but the dedicated executor hosts do not process the final result sets for queries.

Using dedicated coordinators offers the following benefits:

- Reduces memory usage by limiting the number of Impala nodes that need to cache metadata.
- Provides better concurrency by avoiding coordinator bottleneck.
- Eliminates query over-admission.
- Reduces resource, especially network, utilization on the StateStore daemon by limiting metadata broadcast to a subset of nodes.
- Improves reliability and performance for highly concurrent workloads by reducing workload stress on coordinators. Dedicated coordinators require 50% or fewer connections and threads.
- Reduces the number of explicit metadata refreshes required.
- Improves diagnosability if a bottleneck or other performance issue arises on a specific host, you can narrow down the cause more easily because each host is dedicated to specific operations within the overall Impala workload.

In this configuration with dedicated coordinators / executors, you cannot connect to the dedicated executor hosts through clients such as impala-shell or business intelligence tools as only the coordinator nodes support client connections.

Determining the Optimal Number of Dedicated Coordinators

You should have the smallest number of coordinators that will still satisfy your workload requirements in a cluster. A rough estimation is 1 coordinator for every 50 executors.

To maintain a healthy state and optimal performance, it is recommended that you keep the peak utilization of all resources used by Impala, including CPU, the number of threads, the number of connections, and RPCs, under 80%.

Consider the following factors to determine the right number of coordinators in your cluster:

- What is the number of concurrent queries?
- What percentage of the workload is DDL?
- What is the average query resource usage at the various stages (merge, runtime filter, result set size, etc.)?
- How many Impala Daemons (impalad) is in the cluster?
- Is there a high availability requirement?
- Compute/storage capacity reduction factor

Start with the below set of steps to determine the initial number of coordinators:

1. If your cluster has less than 10 nodes, we recommend that you configure one dedicated coordinator. Deploy the dedicated coordinator on a DataNode to avoid losing storage capacity. In most of the cases, one dedicated coordinator is enough to support all workloads on a cluster.
2. Add more coordinators if the dedicated coordinator CPU or network peak utilization is 80% or higher. You might need 1 coordinator for every 50 executors.
3. If the Impala service is shared by multiple workgroups with a dynamic resource pool assigned, use one coordinator per pool to avoid admission control over admission.
4. If high availability is required, double the number of coordinators. One set as an active set and the other as a backup set.

Advance Tuning

Use the following guidelines to further tune the throughput and stability.

1. The concurrency of DML statements does not typically depend on the number of coordinators or size of the cluster. Queries that return large result sets (10,000+ rows) consume more CPU and memory resources on the coordinator. Add one or two coordinators if the workload has many such queries.
2. DDL queries, excluding COMPUTE STATS and CREATE TABLE AS SELECT, are executed only on coordinators. If your workload contains many DDL queries running concurrently, you could add one coordinator.
3. The CPU contention on coordinators can slow down query executions when concurrency is high, especially for very short queries (<10s). Add more coordinators to avoid CPU contention.
4. On a large cluster with 50+ nodes, the number of network connections from a coordinator to executors can grow quickly as query complexity increases. The growth is much greater on coordinators than executors. Add a few more coordinators if workloads are complex, i.e. (an average number of fragments * number of Impalad) > 500, but with the low memory/CPU usage to share the load. Watch IMPALA-4603 and IMPALA-7213 to track the progress on fixing this issue.
5. When using multiple coordinators for DML statements, divide queries to different groups (number of groups = number of coordinators). Configure a separate dynamic resource pool for each group and direct each group of query requests to a specific coordinator. This is to avoid query over admission.
6. The front-end connection requirement is not a factor in determining the number of dedicated coordinators. Consider setting up a connection pool at the client side instead of adding coordinators. For a short-term solution, you could increase the value of fe_service_threads on coordinators to allow more client connections.
7. In general, you should have a very small number of coordinators so storage capacity reduction is not a concern. On a very small cluster (less than 10 nodes), deploy a dedicated coordinator on a DataNode to avoid storage capacity reduction.

Estimating Coordinator Resource Usage

Resource	Safe range	Notes / CM tsquery to monitor
Memory	(Max JVM heap setting + query concurrency * query mem_limit) <= 80% of Impala process memory allocation	Memory usage: SELECT mem_rss WHERE entityName = "Coordinator Instance ID" AND category = ROLE JVM heap usage (metadata cache): SELECT impala_jvm_heap_current_usage_bytes WHERE entityName = "Coordinator Instance ID" AND category = ROLE (only in release 5.15 and above)

TCP Connection	Incoming + outgoing < 16K	Incoming connection usage: SELECT thrift_server_backend_connections_in_use WHERE entityName = "Coordinator Instance ID" AND category = ROLE Outgoing connection usage: SELECT backends_client_cache_clients_in_use WHERE entityName = "Coordinator Instance ID" AND category = ROLE
Threads	< 32K	SELECT thread_manager_running_threads WHERE entityName = "Coordinator Instance ID" AND category = ROLE
CPU	Concurrency = non-DDL query concurrency <= number of virtual cores allocated to Impala per node	CPU usage estimation should be based on how many cores are allocated to Impala per node, not a sum of all cores of the cluster. It is recommended that concurrency should not be more than the number of virtual cores allocated to Impala per node. Query concurrency: SELECT total_impala_num_queries_registered_across_impalads WHERE entityName = "IMPALA-1" AND category = SERVICE

If usage of any of the above resources exceeds the safe range, add one more coordinator.

Monitoring Coordinator Resource Usage

Using Cloudera Manager, monitor the coordinator resource usage to understand your workload and adjust the number of coordinators according to the guidelines above. The available options are:

- Impala Queries tab: Monitor such attributes as DDL queries and Rows produced.
- Custom charts: Monitor activities, such as query complexity which is an average fragment count per query (total fragments / total queries).
- tsquery: Build the custom charts to monitor and estimate the amount of resource the coordinator needs.

The following are sample queries for common resource usage monitoring. Replace entityName values with your coordinator instance id.

Per coordinator tsquery

Resource Usage	Tsquery
Memory usage	SELECT impala_memory_total_used, mem_tracker_process_limit WHERE entityName = "Coordinator Instance ID" AND category = ROLE
JVM heap usage (metadata cache)	SELECT impala_jvm_heap_current_usage_bytes WHERE entityName = "Coordinator Instance ID" AND category = ROLE (only in release 5.15 and above)
CPU usage	SELECT cpu_user_rate / getHostFact(numCores, 1) * 100, cpu_system_rate / getHostFact(numCores, 1) * 100 WHERE entityName="Coordinator Instance ID"
Network usage (host level)	SELECT total_bytes_receive_rate_across_network_interfaces, total_bytes_transmit_rate_across_network_interfaces WHERE entityName="Coordinator Instance ID"
Incoming connection usage	SELECT thrift_server_backend_connections_in_use WHERE entityName = "Coordinator Instance ID" AND category = ROLE

Resource Usage	Tsquery
Outgoing connection usage	SELECT backends_client_cache_clients_in_use WHERE entityName = "Coordinator Instance ID" AND category = ROLE
Thread usage	SELECT thread_manager_running_threads WHERE entityName = "Coordinator Instance ID" AND category = ROLE

Cluster wide tsquery

Resource usage	Tsquery
Front-end connection usage	SELECT total_thrift_server_beeswax_frontend_connections_in_use_across_impalads, total_thrift_server_hiveserver2_frontend_connections_in_use_across_impalads
Query concurrency	SELECT total_impala_num_queries_registered_across_impalads WHERE entityName = "IMPALA-1" AND category = SERVICE

Related Information

Managing Metadata in Impala

Hadoop File Formats Support

Impala supports a number of file formats used in Apache Hadoop.

Impala can load and query data files produced by other Hadoop components such as Spark, and data files produced by Impala can be used by other components also. The following sections discuss the procedures, limitations, and performance considerations for using each file format with Impala.

The file format used for an Impala table has significant performance consequences. Some file formats include compression support that affects the size of data on the disk and, consequently, the amount of I/O and CPU resources required to deserialize data. The amounts of I/O and CPU resources required can be a limiting factor in query performance since querying often begins with moving and decompressing data. To reduce the potential impact of this part of the process, data is often compressed. By compressing data, a smaller total number of bytes are transferred from disk to memory. This reduces the amount of time taken to transfer the data, but a tradeoff occurs when the CPU decompresses the content.

For the file formats that Impala cannot write to, create the table from within Impala whenever possible and insert data using another component such as Hive or Spark. See the table below for specific file formats.

The following table lists the file formats that Impala supports.

File Type	Format	Compression Codecs	Impala Can CREATE?	Impala Can INSERT?
Parquet	Structured	Snappy, gzip, zstd, LZ4, ZLIB, deflate; currently Snappy by default	Yes.	Yes. CREATE TABLE, INSERT, LOAD DATA, and query.
ORC	Structured	gzip, Snappy, LZ4; currently gzip by default	Yes. By default, ORC reads are enabled in Impala 3.4.0 and higher. To disable, set--enable _orc_scanner to false when starting the cluster.	No. Import data by using LOAD DATA on data files already in the right format, or use INSERT in Hive followed by REFR ESH *TABLE_NAME* in Impala.

File Type	Format	Compression Codecs	Impala Can CREATE?	Impala Can INSERT?
Text	Unstructured	bzip2, deflate, gzip, Snappy,zstd ,ZLIB	Yes. CREATE TABLE with no STORED AS clause. The default file format is uncompressed text, with values separated by ASCII 0x01 characters (typically represented as Ctrl-A).	Yes if uncompressed. CREATE TABLE, INSERT, LOAD DATA, and query. No if compressed. If other kinds of compression are used, you must load data through LOAD DATA, Hive, or manually in HDFS.
Avro	Structured	Snappy, gzip, deflate	Yes.	No. Import data by using LOAD DATA on data files already in the right format, or use INSERT in Hive followed by REFRESH *TABLE_NAME* in Impala.
RCFile	Structured	Snappy, gzip, deflate, bzip2	Yes.	No. Import data by using LOAD DATA on data files already in the right format, or use INSERT in Hive followed by REFRESH *TABLE_NAME* in Impala.
SequenceFile	Structured	Snappy, gzip, deflate, bzip2	Yes.	No. Import data by using LOAD DATA on data files already in the right format, or use INSERT in Hive followed by REFRESH *TABLE_NAME* in Impala.

Impala supports the following compression codecs:

Snappy

Recommended for its effective balance between compression ratio and decompression speed. Snappy compression is very fast, but gzip provides greater space savings. Supported for text, RC, Sequence, and Avro files in Impala 2.0 and higher.

Gzip

Recommended when achieving the highest level of compression (and therefore greatest disk-space savings) is desired. Supported for text, RC, Sequence and Avro files in Impala 2.0 and higher.

Deflate

Supported for AVRO, RC, Sequence, and text files.

Bzip2

Supported for text, RC, and Sequence files in Impala 2.0 and higher.

Zstd

For Parquet and text files only.

Lz4

For Parquet files only.

Choosing the File Format for a Table

Different file formats and compression codecs work better for different data sets. Choosing the proper format for your data can yield performance improvements. Use the following considerations to decide which combination of file format and compression to use for a particular table.

- If you are working with existing files that are already in a supported file format, use the same format for the Impala table if performance is acceptable. If the original format does not yield acceptable query performance or

resource usage, consider creating a new Impala table with different file format or compression characteristics, and doing a one-time conversion by rewriting the data to the new table.

- Text files are convenient to produce through many different tools and are human-readable for ease of verification and debugging. Those characteristics are why text is the default format for an Impala CREATE TABLE statement. However, when performance and resource usage are the primary considerations, use one of the structured file formats that include metadata and built-in compression.

  A typical workflow might involve bringing data into an Impala table by copying CSV or TSV files into the appropriate data directory, and then using the INSERT ... SELECT syntax to rewrite the data into a table using a different, more compact file format.

Using Text Data Files

Impala supports using text files as the storage format for input and output. Text files are a convenient format to use for interchange with other applications or scripts that produce or read delimited text files, such as CSV or TSV with commas or tabs for delimiters.

Text files are flexible in their column definitions. For example, a text file could have more fields than the Impala table, and those extra fields are ignored during queries. Or it could have fewer fields than the Impala table, and those missing fields are treated as NULL values in queries.

You could have fields that were treated as numbers or timestamps in a table, then use ALTER TABLE ... REPLACE COLUMNS to switch them to strings, or the reverse.

Creating Text Tables

You can create tables with specific separator characters to import text files in familiar formats such as CSV, TSV, or pipe-separated with the FIELDS TERMINATED BY clause preceded by the ROW FORMAT DELIMITED clause. For example:

```
CREATE TABLE tsv(id INT, s STRING, n INT, t TIMESTAMP, b BOOLEAN)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
```

You can specify a delimiter character '\0' to use the ASCII 0 (nul) character for text tables.

You can also use these tables to produce output data files, by copying data into them through the INSERT ... SELECT syntax and then extracting the data files from the Impala data directory.

The data files created by any INSERT statements uses the Ctrl-A character (hex 01) as a separator between each column value.

Issue a DESCRIBE FORMATTED *TABLE_NAME* statement to see the details of how each table is represented internally in Impala.

Complex type considerations: Although you can create tables in this file format using the complex types (ARRAY, STRUCT, and MAP), currently, Impala cannot query these types in text tables.

Data Files for Text Tables

When Impala queries a table with data in text format, it consults all the data files in the data directory for that table, with some exceptions:

- Impala ignores any hidden files, that is, files whose names start with a dot or an underscore.
- Impala queries ignore files with extensions commonly used for temporary work files by Hadoop tools. Any files with extensions .tmp or .copying are not considered part of the Impala table. The suffix matching is case-insensitive, so for example Impala ignores both .copying and .COPYING suffixes.

- Impala uses suffixes to recognize when text data files are compressed text. For Impala to recognize the compressed text files, they must have the appropriate file extension corresponding to the compression codec, either .bz2, .gz, .snappy, or .zst, .deflate. The extensions can be in uppercase or lowercase.
- Otherwise, the file names are not significant. When you put files into an HDFS directory through ETL jobs, or point Impala to an existing HDFS directory with the CREATE EXTERNAL TABLE statement, or move data files under external control with the LOAD DATA statement, Impala preserves the original filenames.

An INSERT ... SELECT statement produces one data file from each node that processes the SELECT part of the statement. An INSERT ... VALUES statement produces a separate data file for each statement; because Impala is more efficient querying a small number of huge files than a large number of tiny files, the INSERT ... VALUES syntax is not recommended for loading a substantial volume of data. If you find yourself with a table that is inefficient due to too many small data files, reorganize the data into a few large files by doing INSERT ... SELECT to transfer the data to a new table.

Do not surround string values with quotation marks in text data files that you construct. If you need to include the separator character inside a field value, for example to put a string value with a comma inside a CSV-format data file, specify an escape character on the CREATE TABLE statement with the ESCAPED BY clause, and insert that character immediately before any separator characters that need escaping.

Special values within text data files:

- Impala recognizes the literal strings inf for infinity and nan for “Not a Number”, for FLOAT and DOUBLE columns.
- Impala recognizes the literal string \N to represent NULL. When using Sqoop, specify the options --null-non-string and --null-string to ensure all NULL values are represented correctly in the Sqoop output files. \N needs to be escaped as in the below example:

```
--null-string '\\N' --null-non-string '\\N'
```

- By default, Sqoop writes NULL values using the string null, which causes a conversion error when such rows are evaluated by Impala. (A workaround for existing tables and data files is to change the table properties through ALTER TABLE *NAME* SET TBLPROPERTIES("serialization.null.format"="null").)
- Impala can optionally skip an arbitrary number of header lines from text input files on HDFS based on the skip .header.line.count value in the TBLPROPERTIES field of the table metadata.

Character Encoding

In Impala 5.0 and higher, Impala supports reading from and writing to Text and Sequence files that use arbitrary character encodings beyond the default UTF-8.

This is enabled for tables that use the Hive LazySimpleSerDe and have the table property "serialization.encoding" set (e.g., tblproperties("serialization.encoding"="GBK")). When you query or insert data into such a table, Impala will correctly decode or encode the string data using the specified character set.



Note: For compatibility with existing features, Impala implements the encoding configuration at the partition level. Charsets that store line delimiters differently from ASCII are not allowed.

Loading Data into Text Tables

To load an existing text file into an Impala text table, use the LOAD DATA statement and specify the path of the file in HDFS. That file is moved into the appropriate Impala data directory.

To load multiple existing text files into an Impala text table, use the LOAD DATA statement and specify the HDFS path of the directory containing the files. All non-hidden files are moved into the appropriate Impala data directory.

Use the DESCRIBE FORMATTED statement to see the HDFS directory where the data files are stored, then use Linux commands such as hdfs dfs -ls *HDFS_DIRECTORY* and hdfs dfs -cat *HDFS_FILE* to display the contents of an Impala-created text file.

When you create a text file for use with an Impala text table, specify \N to represent a NULL value.

If a text file has fewer fields than the columns in the corresponding Impala table, all the corresponding columns are set to NULL when the data in that file is read by an Impala query.

If a text file has more fields than the columns in the corresponding Impala table, the extra fields are ignored when the data in that file is read by an Impala query.

You can also use manual HDFS operations such as hdfs dfs -put or hdfs dfs -cp to put data files in the data directory for an Impala table. When you copy or move new data files into the HDFS directory for the Impala table, issue a REFRESH *TABLE_NAME* statement in `impala-shell` before issuing the next query against that table, to make Impala recognize the newly added files.

Query Performance for Text Tables

Data stored in text format is relatively bulky, and not as efficient to query as binary formats such as Parquet. For the tables used in your most performance-critical queries, look into using more efficient alternate file formats.

For frequently queried data, you might load the original text data files into one Impala table, then use an INSERT statement to transfer the data to another table that uses the Parquet file format. The data is converted automatically as it is stored in the destination table.

For more compact data. consider using text data compressed in the gzip, bzip2, or Snappy formats. However, these compressed formats are not “splittable” so there is less opportunity for Impala to parallelize queries on them. You also have the choice to convert the data to Parquet using an INSERT ... SELECT statement to copy the original data into a Parquet table.

Using bzip2, deflate, gzip, Snappy-Compressed, or zstd Text Files

Impala supports using text data files that employ bzip2, deflate, gzip, Snappy, or zstd compression. These compression types are primarily for convenience within an existing ETL pipeline rather than maximum performance. Although it requires less I/O to read compressed text than the equivalent uncompressed text, files compressed by these codecs are not “splittable” and therefore cannot take full advantage of the Impala parallel query capability. Impala can read compressed text files written by Hive.

As each Snappy-compressed file is processed, the node doing the work reads the entire file into memory and then decompresses it. Therefore, the node must have enough memory to hold both the compressed and uncompressed data from the text file. The memory required to hold the uncompressed data is difficult to estimate in advance, potentially causing problems on systems with low memory limits or with resource management enabled. This memory overhead is reduced for bzip2-, deflate-, gzip-, and zstd-compressed text files. The compressed data is decompressed as it is read, rather than all at once.

To create a table to hold compressed text, create a text table with no special compression options. Specify the delimiter and escape character if required, using the ROW FORMAT clause.

Because Impala can query compressed text files but currently cannot write them, produce the compressed text files outside Impala and use the LOAD DATA statement, manual HDFS commands to move them to the appropriate Impala data directory. (Or, you can use CREATE EXTERNAL TABLE and point the LOCATION attribute at a directory containing existing compressed text files.)

Using Parquet Data Files

Impala allows you to create, manage, and query Parquet tables. Parquet is a column-oriented binary file format intended to be highly efficient for the types of large-scale queries.

Parquet is suitable for queries scanning particular columns within a table, for example, to query “wide” tables with many columns, or to perform aggregation operations such as SUM() and AVG() that need to process most or all of the values from a column.

Each Parquet data file written by Impala contains the values for a set of rows (referred to as the “row group”). Within a data file, the values from each column are organized so that they are all adjacent, enabling good compression for

the values from that column. Queries against a Parquet table can retrieve and analyze these values from any column quickly and with minimal I/O.

Creating Parquet Tables

To create a table in the Parquet format, use the STORED AS PARQUET clause in the CREATE TABLE statement. For example:

```
CREATE TABLE PARQUET_TABLE_NAME (x INT, y STRING) STORED AS PARQUET;
```

Or, to clone the column names and data types of an existing table, use the LIKE with the STORED AS PARQUET clause. For example:

```
CREATE TABLE PARQUET_TABLE_NAME LIKE OTHER_TABLE_NAME STORED AS PARQUET;
```

You can derive column definitions from a raw Parquet data file, even without an existing Impala table. For example, you can create an external table pointing to an HDFS directory, and base the column definitions on one of the files in that directory:

```
CREATE EXTERNAL TABLE ingest_existing_files LIKE PARQUET '/user/etl/destinat
ion/datafile1.dat'
  STORED AS PARQUET
  LOCATION '/user/etl/destination';
```

Or, you can refer to an existing data file and create a new empty table with suitable column definitions. Then you can use INSERT to create new data files or LOAD DATA to transfer existing data files into the new table.

```
CREATE TABLE columns_from_data_file LIKE PARQUET '/user/etl/destination/data
file1.dat'
  STORED AS PARQUET;
```

In this example, the new table is partitioned by year, month, and day. These partition key columns are not part of the data file, so you specify them in the CREATE TABLE statement:

```
CREATE TABLE columns_from_data_file LIKE PARQUET '/user/etl/destination/data
file1.dat'
  PARTITION (year INT, month TINYINT, day TINYINT)
  STORED AS PARQUET;
```

If the Parquet table has a different number of columns or different column names than the other table, specify the names of columns from the other table rather than * in the SELECT statement.

Data Type Considerations for Parquet Tables

The Parquet format defines a set of data types whose names differ from the names of the corresponding Impala data types. If you are preparing Parquet files using other Hadoop components such as Pig or MapReduce, you might need to work with the type names defined by Parquet. The following tables list the Parquet-defined types and the equivalent types in Impala.

Primitive types

Parquet type	Impala type
BINARY	STRING
BOOLEAN	BOOLEAN
DOUBLE	DOUBLE
FLOAT	FLOAT

Parquet type	Impala type
INT32	INT
INT64	BIGINT
INT96	TIMESTAMP

Logical types

Parquet uses type annotations to extend the types that it can store, by specifying how the primitive types should be interpreted.

Parquet primitive type and annotation	Impala type
BINARY annotated with the UTF8 OriginalType	STRING
BINARY annotated with the STRING LogicalType	STRING
BINARY annotated with the ENUM OriginalType	STRING
BINARY annotated with the DECIMAL OriginalType	DECIMAL
INT64 annotated with the TIMESTAMP_MILLIS OriginalType	TIMESTAMP or BIGINT (for backward compatibility)
INT64 annotated with the TIMESTAMP_MICROS OriginalType	TIMESTAMP or BIGINT (for backward compatibility)
INT64 annotated with the TIMESTAMP LogicalType	TIMESTAMP or BIGINT (for backward compatibility)

Complex types:

Impala only supports queries against the complex types (ARRAY, MAP, and STRUCT) in Parquet tables.

Loading Data into Parquet Tables

Choose from the following process to load data into Parquet tables based on whether the original data is already in an Impala table, or exists as raw data files outside Impala.

If you already have data in an Impala or Hive table, perhaps in a different file format or partitioning scheme:

- Transfer the data to a Parquet table using the Impala INSERT...SELECT statement.

  For example:

  ```
  INSERT OVERWRITE TABLE parquet_table_name SELECT * FROM other_table_name;
  ```

  You can convert, filter, repartition, and do other things to the data as part of this same INSERT statement.

  When inserting into partitioned tables, especially using the Parquet file format, you can include a hint in the INSERT statement to fine-tune the overall performance of the operation and its resource usage.

Any INSERT statement for a Parquet table requires enough free space in the HDFS filesystem to write one block. Because Parquet data files use a block size of 1 GB by default, an INSERT might fail (even for a very small amount of data) if your HDFS is running low on space.

Avoid the INSERT...VALUES syntax for Parquet tables, because INSERT...VALUES produces a separate tiny data file for each INSERT...VALUES statement, and the strength of Parquet is in its handling of data (compressing, parallelizing, and so on) in large chunks.

If you have one or more Parquet data files produced outside of Impala, you can quickly make the data query-able through Impala by one of the following methods:

- The LOAD DATA statement moves a single data file or a directory full of data files into the data directory for an Impala table. It does no validation or conversion of the data.

  The original data files must be somewhere in HDFS, not the local filesystem.
- The CREATE TABLE statement with the LOCATION clause creates a table where the data continues to reside outside the Impala data directory.

  The original data files must be somewhere in HDFS, not the local filesystem.

  For extra safety, if the data is intended to be long-lived and reused by other applications, you can use the CREA TE EXTERNAL TABLE syntax so that the data files are not deleted by an Impala DROP TABLE statement.
- If the Parquet table already exists, you can copy Parquet data files directly into it using the hadoop distcp -pb command, then use the REFRESH statement to make Impala recognize the newly added data.

  You must preserve the block size of the Parquet data files by using the hadoop distcp -pb command rather than a -put or -cp operation on the Parquet files.



Note:

Currently, Impala always decodes the column data in Parquet files based on the ordinal position of the columns, not by looking up the position of each column based on its name. Parquet files produced outside of Impala must write column data in the same order as the columns are declared in the Impala table. Any optional columns that are omitted from the data files must be the rightmost columns in the Impala table definition.

If you created compressed Parquet files through some tool other than Impala, make sure that any compression codecs are supported in Parquet by Impala. For example, Impala does not currently support LZO compression in Parquet files. Also doublecheck that you used any recommended compatibility settings in the other tool, such as spark.sql.parquet.binaryAsString when writing Parquet files through Spark.

Recent versions of Sqoop can produce Parquet output files using the --as-parquetfile option.

If the data exists outside Impala and is in some other format, combine both of the preceding techniques. First, use a LOAD DATA or CREATE EXTERNAL TABLE ... LOCATION statement to bring the data into an Impala table that uses the appropriate file format. Then, use an INSERT...SELECT statement to copy the data to the Parquet table, converting to Parquet format as part of the process.

Loading data into Parquet tables is a memory-intensive operation, because the incoming data is buffered until it reaches one data block in size, then that chunk of data is organized and compressed in memory before being written out. The memory consumption can be larger when inserting data into partitioned Parquet tables, because a separate data file is written for each combination of partition key column values, potentially requiring several large chunks to be manipulated in memory at once.

When inserting into a partitioned Parquet table, Impala redistributes the data among the nodes to reduce memory consumption. You might still need to temporarily increase the memory dedicated to Impala during the insert operation, or break up the load operation into several INSERT statements, or both.

Query Performance for Parquet Tables

Query performance for Parquet tables depends on the number of columns needed to process the SELECT list and WHERE clauses of the query, the way data is divided into large data files with block size equal to file size, the reduction in I/O by reading the data for each column in compressed format, which data files can be skipped (for partitioned tables), and the CPU overhead of decompressing the data for each column.

For example, the following is an efficient query for a Parquet table:

```
SELECT AVG(income) FROM census_data WHERE state = 'CA';
```

The query processes only 2 columns out of a large number of total columns. If the table is partitioned by the STATE column, it is even more efficient because the query only has to read and decode 1 column from each data file, and it can read only the data files in the partition directory for the state 'CA', skipping the data files for all the other states, which will be physically located in other directories.

The following is a relatively inefficient query for a Parquet table:

```
SELECT * FROM census_data;
```

Impala would have to read the entire contents of each large data file, and decompress the contents of each column for each row group, negating the I/O optimizations of the column-oriented format. This query might still be faster for a Parquet table than a table with some other file format, but it does not take advantage of the unique strengths of Parquet data files.

Impala can optimize queries on Parquet tables, especially join queries, better when statistics are available for all the tables. Issue the COMPUTE STATS statement for each table after substantial amounts of data are loaded into or appended to it.

The runtime filtering feature works best with Parquet tables. The per-row filtering aspect only applies to Parquet tables.

Impala queries are optimized for files stored in Amazon S3. For Impala tables that use the Parquet file formats, the PARQUET_OBJECT_STORE_SPLIT_SIZE query option determines how Impala divides the I/O work of reading the data files. By default, this value is 256 MB. Impala parallelizes S3 read operations on the files as if they were made up of 256 MB blocks to match the row group size produced by Impala.

Parquet files written by Impala include embedded metadata specifying the minimum and maximum values for each column, within each row group and each data page within the row group. Impala-written Parquet files typically contain a single row group; a row group can contain many data pages. Impala uses this information (currently, only the metadata for each row group) when reading each Parquet data file during a query, to quickly determine whether each row group within the file potentially includes any rows that match the conditions in the WHERE clause.

For example, if the column X within a particular Parquet file has a minimum value of 1 and a maximum value of 100, then a query including the clause WHERE x > 200 can quickly determine that it is safe to skip that particular file, instead of scanning all the associated column values.

This optimization technique is especially effective for tables that use the SORT BY clause for the columns most frequently checked in WHERE clauses, because any INSERT operation on such tables produces Parquet data files with relatively narrow ranges of column values within each file.

To disable Impala from writing the Parquet page index when creating Parquet files, set the PARQUET_WRITE_PAGE_INDEX query option to FALSE.

Partitioning for Parquet Tables

Partitioning is an important performance technique for Impala generally. This section explains some of the performance considerations for partitioned Parquet tables.

The Parquet file format is ideal for tables containing many columns, where most queries only refer to a small subset of the columns. The physical layout of Parquet data files lets Impala read only a small fraction of the data for many queries. The performance benefits of this approach are amplified when you use Parquet tables in combination with partitioning. Impala can skip the data files for certain partitions entirely, based on the comparisons in the WHERE clause that refer to the partition key columns. For example, queries on partitioned tables often analyze data for time intervals based on columns such as YEAR, MONTH, and/or DAY, or for geographic regions.

As Parquet data files use a large block size, when deciding how finely to partition the data, try to find a granularity where each partition contains 256 MB or more of data, rather than creating a large number of smaller files split among many partitions.

Inserting into a partitioned Parquet table can be a resource-intensive operation, because each Impala node could potentially be writing a separate data file to HDFS for each combination of different values for the partition key columns. The large number of simultaneous open files could exceed the HDFS “transceivers” limit. To avoid exceeding this limit, consider the following techniques:

- Load different subsets of data using separate INSERT statements with specific values for the PARTITION clause, such as PARTITION (year=2010).
- Increase the “transceivers” value for HDFS, sometimes spelled “xcievers” (sic). The property value in the hdfs-site.xml configuration file is dfs.datanode.max.transfer.threads.

  For example, if you were loading 12 years of data partitioned by year, month, and day, even a value of 4096 might not be high enough.
- Use the COMPUTE STATS statement to collect column statistics on the source table from which data is being copied, so that the Impala query can estimate the number of different values in the partition key columns and distribute the work accordingly.

Enabling Compression for Parquet Tables

When Impala writes Parquet data files using the INSERT statement, the underlying compression is controlled by the COMPRESSION_CODEC query option. The allowed values for this query option are snappy (the default), gzip, zstd, lz4, and none, the compression codecs that Impala supports for Parquet.

Snappy

By default, the underlying data files for a Parquet table are compressed with Snappy. The combination of fast compression and decompression makes it a good choice for many data sets.

GZip

If you need more intensive compression (at the expense of more CPU cycles for uncompressing during queries), set the COMPRESSION_CODEC query option to gzip before inserting the data.

Zstd

Zstd is a real-time compression algorithm offering a tradeoff between speed and ratio of compression. Compression levels from 1 up to 22 are supported. The lower the level, the faster the speed at the cost of compression ratio.

Lz4

Lz4 is a lossless compression algorithm providing extremely fast and scalable compression and decompression.

None

If your data compresses very poorly, or you want to avoid the CPU overhead of compression and decompression entirely, set the COMPRESSION_CODEC query option to none before inserting the data.

The actual compression ratios, and relative insert and query speeds, will vary depending on the characteristics of the actual data.

Because Parquet data files are typically large, each directory will have a different number of data files and the row groups will be arranged differently.

At the same time, the less aggressive the compression, the faster the data can be decompressed.

For example, using a table with a billion rows, switching from Snappy to GZip compression shrinks the data by an additional 40% or so, while switching from Snappy compression to no compression expands the data also by about 40%. A query that evaluates all the values for a particular column runs faster with no compression than with Snappy compression, and faster with Snappy compression than with Gzip compression.

The data files using the various compression codecs are all compatible with each other for read operations. The metadata about the compression format is written into each data file, and can be decoded during queries regardless of the COMPRESSION_CODEC setting in effect at the time.

Exchanging Parquet Data Files with Other Cloudera Components

You can read and write Parquet data files from other Cloudera components, such as Hive.

Impala supports the scalar data types that you can encode in a Parquet data file, but not composite or nested types such as maps or arrays. Impala can query Parquet data files that include composite or nested types, as long as the query only refers to columns with scalar types.

If you copy Parquet data files between nodes, or even between different directories on the same node, make sure to preserve the block size by using the command hadoop distcp -pb. To verify that the block size was preserved, issue the command hdfs fsck -blocks *HDFS_PATH_OF_IMPALA_TABLE_DIR* and check that the average block size is at or near 256 MB (or whatever other size is defined by the PARQUET_FILE_SIZE query option).. (The hadoop distcp operation typically leaves some directories behind, with names matching _distcp_logs_*, that you can delete from the destination directory afterward.)

Issue the command `hadoop distcp` for details about `distcp` command syntax.

Impala can query Parquet files that use the PLAIN, PLAIN_DICTIONARY, BIT_PACKED, and RLE encodings. Starting with this release, Impala supports decoding RLE_DICTIONARY encoded pages. This encoding is identical to the already-supported PLAIN_DICTIONARY encoding but the PLAIN enum value is used for the dictionary pages and the RLE_DICTIONARY enum value is used for the data pages. When creating files outside of Impala for use by Impala, make sure to use one of the supported encodings.

In particular, for MapReduce jobs, parquet.writer.version must not be defined (especially as PARQUET_2_0) for writing the configurations of Parquet MR jobs.

Data using the version 2.0 of Parquet writer might not be consumable by Impala, due to use of the RLE_DICTIONARY encoding.

Use the default version of the Parquet writer and refrain from overriding the default writer version by setting the parquet.writer.version property or via WriterVersion.PARQUET_2_0 in the Parquet API.

Supported Encodings in Impala. Impala can read Parquet data written with the following encodings:

- Plain (PLAIN = 0)
- Dictionary encoding (PLAIN_DICTIONARY = 2 and RLE_DICTIONARY = 8)
- Run Length Encoding / Bit-Packing Hybrid (RLE = 3)

Unsupported Encodings in Impala. The following encodings are not supported:

- Delta binary packed (DELTA_BINARY_PACKED = 5)
- Delta-length byte array (DELTA_LENGTH_BYTE_ARRAY = 6)
- Delta byte array (DELTA_BYTE_ARRAY = 7)
- Byte stream split (BYTE_STREAM_SPLIT = 9)

Deprecated Encoding. The following encoding is deprecated and not supported by Impala:

- Bit-packed (BIT_PACKED = 4)

For more information, see *Apache Parquet encoding*.

How Parquet Data Files Are Organized

Although Parquet is a column-oriented file format, Parquet keeps all the data for a row within the same data file, to ensure that the columns for a row are always available on the same node for processing. Parquet sets a large HDFS block size and a matching maximum data file size to ensure that I/O and network transfer requests apply to large batches of data.

Within that data file, the data for a set of rows is rearranged so that all the values from the first column are organized in one contiguous block, then all the values from the second column, and so on. Putting the values from the same column next to each other lets Impala use effective compression techniques on the values in that column.

Note:

Impala INSERT statements write Parquet data files using an HDFS block size that matches the data file size, to ensure that each data file is represented by a single HDFS block, and the entire file can be processed on a single node without requiring any remote reads.

If you create Parquet data files outside of Impala, such as through a MapReduce or Pig job, ensure that the HDFS block size is greater than or equal to the file size, so that the “one file per block” relationship is maintained. Set the dfs.block.size or the dfs.blocksize property large enough that each file fits within a single HDFS block, even if that size is larger than the normal HDFS block size.

If the block size is reset to a lower value during a file copy, you will see lower performance for queries involving those files, and the PROFILE statement will reveal that some I/O is being done suboptimally, through remote reads.

When Impala retrieves or tests the data for a particular column, it opens all the data files, but only reads the portion of each file containing the values for that column. The column values are stored consecutively, minimizing the I/O required to process the values within a single column. If other columns are named in the SELECT list or WHERE clauses, the data for all columns in the same row is available within that same data file.

If an INSERT statement brings in less than one Parquet block's worth of data, the resulting data file is smaller than ideal. Thus, if you do split up an ETL job to use multiple INSERT statements, try to keep the volume of data for each INSERT statement to approximately 256 MB, or a multiple of 256 MB.

RLE and Dictionary Encoding for Parquet Data Files

Parquet uses some automatic compression techniques, such as run-length encoding (RLE) and dictionary encoding, based on analysis of the actual data values. Once the data values are encoded in a compact form, the encoded data can optionally be further compressed using a compression algorithm. Parquet data files created by Impala can use Snappy, GZip, or no compression; the Parquet spec also allows LZO compression, but currently Impala does not support LZO-compressed Parquet files.

RLE and dictionary encoding are compression techniques that Impala applies automatically to groups of Parquet data values, in addition to any Snappy or GZip compression applied to the entire data files. These automatic optimizations can save you time and planning that are normally needed for a traditional data warehouse. For example, dictionary encoding reduces the need to create numeric IDs as abbreviations for longer string values.

Run-length encoding condenses sequences of repeated data values. For example, if many consecutive rows all contain the same value for a country code, those repeating values can be represented by the value followed by a count of how many times it appears consecutively.

Dictionary encoding takes the different values present in a column, and represents each one in compact 2-byte form rather than the original value, which could be several bytes. (Additional compression is applied to the compacted values, for extra space savings.) This type of encoding applies when the number of different values for a column is less than 2**16 (16,384). It does not apply to columns of data type BOOLEAN, which are already very short. TIME STAMP columns sometimes have a unique value for each row, in which case they can quickly exceed the 2**16 limit on distinct values. The 2**16 limit on different values within a column is reset for each data file, so if several different data files each contained 10,000 different city names, the city name column in each data file could still be condensed using dictionary encoding.

Compacting Data Files for Parquet Tables

If you reuse existing table structures or ETL processes for Parquet tables, you might encounter a “many small files” situation, which is suboptimal for query efficiency.

Here are techniques to help you produce large data files in Parquet INSERT operations, and to compact existing too-small data files:

- When inserting into a partitioned Parquet table, use statically partitioned INSERT statements where the partition key values are specified as constant values. Ideally, use a separate INSERT statement for each partition.

- You might set the NUM_NODES option to 1 briefly, during INSERT or CREATE TABLE AS SELECT statements. Normally, those statements produce one or more data files per data node. If the write operation involves small amounts of data, a Parquet table, and/or a partitioned table, the default behavior could produce many small files when intuitively you might expect only a single output file. SET NUM_NODES=1 turns off the “distributed” aspect of the write operation, making it more likely to produce only one or a few data files.
- Be prepared to reduce the number of partition key columns from what you are used to with traditional analytic database systems.
- Do not expect Impala-written Parquet files to fill up the entire Parquet block size. Impala estimates on the conservative side when figuring out how much data to write to each Parquet file. Typically, the uncompressed data in memory is substantially reduced on disk by the compression and encoding techniques in the Parquet file format. The final data file size varies depending on the compressibility of the data. Therefore, it is not an indication of a problem if 256 MB of text data is turned into 2 Parquet data files, each less than 256 MB.
- If you accidentally end up with a table with many small data files, consider using one or more of the preceding techniques and copying all the data into a new Parquet table, either through CREATE TABLE AS SELECT or INSERT ... SELECT statements.

  To avoid rewriting queries to change table names, you can adopt a convention of always running important queries against a view. Changing the view definition immediately switches any subsequent queries to use the new underlying tables:

Schema Evolution for Parquet Tables

Schema evolution refers to using the statement ALTER TABLE ... REPLACE COLUMNS to change the names, data type, or number of columns in a table. You can perform schema evolution for Parquet tables as follows:

- The Impala ALTER TABLE statement never changes any data files in the tables. From the Impala side, schema evolution involves interpreting the same data files in terms of a new table definition. Some types of schema changes make sense and are represented correctly. Other types of changes cannot be represented in a sensible way, and produce special result values or conversion errors during queries.
- The INSERT statement always creates data using the latest table definition. You might end up with data files with different numbers of columns or internal data representations if you do a sequence of INSERT and ALTER TA BLE ... REPLACE COLUMNS statements.
- If you use ALTER TABLE ... REPLACE COLUMNS to define additional columns at the end, when the original data files are used in a query, these final columns are considered to be all NULL values.
- If you use ALTER TABLE ... REPLACE COLUMNS to define fewer columns than before, when the original data files are used in a query, the unused columns still present in the data file are ignored.
- Parquet represents the TINYINT, SMALLINT, and INT types the same internally, all stored in 32-bit integers.
  - That means it is easy to promote a TINYINT column to SMALLINT or INT, or a SMALLINT column to INT. The numbers are represented exactly the same in the data file, and the columns being promoted would not contain any out-of-range values.
  - If you change any of these column types to a smaller type, any values that are out-of-range for the new type are returned incorrectly, typically as negative numbers.
  - You cannot change a TINYINT, SMALLINT, or INT column to BIGINT, or the other way around. Although the ALTER TABLE succeeds, any attempt to query those columns results in conversion errors.
  - Any other type conversion for columns produces a conversion error during queries. For example, INT to STRI NG, FLOAT to DOUBLE, TIMESTAMP to STRING, DECIMAL(9,0) to DECIMAL(5,2), and so on.

You might find that you have Parquet files where the columns do not line up in the same order as in your Impala table. For example, you might have a Parquet file that was part of a table with columns C1,C2,C3,C4, and now you want to reuse the same Parquet file in a table with columns C4,C2. By default, Impala expects the columns in the data file to appear in the same order as the columns defined for the table, making it impractical to do some kinds of file reuse or schema evolution.

The query option PARQUET_FALLBACK_SCHEMA_RESOLUTION=name lets Impala resolve columns by name, and therefore handle out-of-order or extra columns in the data file.

Related Information

Apache Parquet encoding

Using ORC Data Files

Impala can read ORC data files.

Disabling ORC reads

By default, ORC reads are enabled in Impala 3.4.0. To disable the support of ORC data files:

1. In Cloudera Manager, navigate to Clusters Impala .
2. In the Configuration tab, set --enable_orc_scanner=false in the Impala Command Line Argument Advanced Configuration Snippet (Safety Valve)field.

   To enable ORC reads, set the property to true.
3. Restart the cluster.

Creating ORC Tables and Loading Data

To create a table in the ORC format, use the STORED AS ORC clause in the CREATE TABLE statement.

Because Impala can query ORC tables but cannot currently write to, after creating ORC tables, use the Hive shell to load the data.

After loading data into a table through Hive or other mechanism outside of Impala, issue a REFRESH *TABLE_NAME* statement the next time you connect to the Impala node, before querying the table, to make Impala recognize the new data.

For example, here is how you might create some ORC tables in Impala (by specifying the columns explicitly, or cloning the structure of another table), load data through Hive, and query them through Impala:

```
$ impala-shell -i localhost
[localhost:21000] default> CREATE TABLE orc_table (x INT) STORED AS ORC;
[localhost:21000] default> CREATE TABLE orc_clone LIKE some_other_table STO
RED AS ORC;
[localhost:21000] default> quit;

$ hive
hive> INSERT INTO TABLE orc_table SELECT x FROM some_other_table;
3 Rows loaded to orc_table
Time taken: 4.169 seconds
hive> quit;
$ impala-shell -i localhost
[localhost:21000] default> -- Make Impala recognize the data loaded through
 Hive;
[localhost:21000] default> REFRESH orc_table;
[localhost:21000] default> SELECT * FROM orc_table;

Fetched 3 row(s) in 0.11s
```

Data Type Considerations for ORC Tables

The ORC format defines a set of data types whose names differ from the names of the corresponding Impala data types. The following figure lists the ORC-defined types and the equivalent types in Impala.

Primitive types:

ORC type	Impala type
BINARY	STRING
BOOLEAN	BOOLEAN
DOUBLE	DOUBLE
FLOAT	FLOAT
TINYINT	TINYINT
SMALLINT	SMALLINT
INT	INT
BIGINT	BIGINT
TIMESTAMP	TIMESTAMP
DATE	Not supported

Complex types:

Impala supports the complex types ARRAY, STRUCT, and MAP in ORC files. Because Impala has better performance on Parquet than ORC, if you plan to use complex types, become familiar with the performance and storage aspects of Parquet first.

Enabling Compression for ORC Tables

ORC tables are in ZLIB (Deflate in Impala) compression by default. The supported compressions for ORC tables are NONE, ZLIB, and SNAPPY.

Set the compression when you load ORC tables in Hive.

Using Avro Data Files

Impala supports creating and querying Avro tables. You need to use Hive to insert data into Avro tables.

Creating Avro Tables

To create a new table using the Avro file format, use the STORED AS AVRO clause in the CREATE TABLE statement. If you create the table through Impala, you must include column definitions that match the fields specified in the Avro schema. With Hive, you can omit the columns and just specify the Avro schema.

The following examples demonstrate creating an Avro table in Impala, using either an inline column specification or one taken from a JSON file stored in HDFS:

```
[localhost:21000] > CREATE TABLE avro_only_sql_columns
                  > (col1 BOOLEAN, col2 INT)
                  > STORED AS AVRO;

[localhost:21000] > CREATE TABLE impala_avro_table
                  > (col1 BOOLEAN, col2 INT)
                  > STORED AS AVRO
                  > TBLPROPERTIES ('avro.schema.literal'='{
                  >    "name": "my_record",
                  >    "type": "record",
                  >    "fields": [
                  >       {"name":"col1", "type":"boolean"},
                  >       {"name":"col2", "type":"int"}]}');
[localhost:21000] > CREATE TABLE avro_examples_of_all_types
```

```
                    (col1 BOOLEAN, col2 INT)
                  >   STORED AS AVRO
                  >   TBLPROPERTIES ('avro.schema.url'='hdfs://localhost:802
0/avro_schemas/alltypes.json');
```

Each field of the record becomes a column of the table. Note that any other information, such as the record name, is ignored.

Note: For nullable Avro columns, make sure to put the "null" entry before the actual type name. In Impala, all columns are nullable; Impala currently does not have a NOT NULL clause. Any non-nullable property is only enforced on the Avro side.

If you create the table through Hive, switch back to `impala-shell` and issue an INVALIDATE METADATA *TABLE_NAME* statement. Then you can run queries for that table through `impala-shell`.

In rare instances, a mismatch could occur between the Avro schema and the column definitions in the Metastore database. Impala checks for such inconsistencies during a CREATE TABLE statement and each time it loads the metadata for a table (for example, after INVALIDATE METADATA). Impala uses the following rules to determine how to treat mismatching columns, a process known as *schema reconciliation*:

- If there is a mismatch in the number of columns, Impala uses the column definitions from the Avro schema.
- If there is a mismatch in column name or type, Impala uses the column definition from the Avro schema. Because a CHAR or VARCHAR column in Impala maps to an Avro STRING, this case is not considered a mismatch and the column is preserved as CHAR or VARCHAR in the reconciled schema.
- An Impala TIMESTAMP column definition maps to an Avro STRING and is presented as a STRING in the reconciled schema, because Avro has no binary TIMESTAMP representation.

Specifying the Avro Schema through JSON:

While you can embed a schema directly in your CREATE TABLE statement, as shown above, column width restrictions in the Hive Metastore limit the length of schema you can specify. If you encounter problems with long schema literals, try storing your schema as a JSON file in HDFS instead. Specify your schema in HDFS using table properties in the following format using the avro.schema.url in TBLPROPERTIES clause.

```
TBLPROPERTIES ('avro.schema.url'='hdfs//your-name-node:port/path/to/schema.j
son');
```

Data Type Considerations for Avro Tables

The Avro format defines a set of data types whose names differ from the names of the corresponding Impala data types. If you are preparing Avro files using other Hadoop components such as Pig or MapReduce, you might need to work with the type names defined by Avro. The following figure lists the Avro-defined types and the equivalent types in Impala.

Primitive types:

Avro type	Impala type
STRING	STRING
STRING	CHAR
STRING	VARCHAR
INT	INT
BOOLEAN	BOOLEAN
LONG	BIGINT
FLOAT	FLOAT
DOUBLE	DOUBLE

The Avro specification allows string values up to 2**64 bytes in length:

- Impala queries for Avro tables use 32-bit integers to hold string lengths.
- Impala truncates CHAR and VARCHAR values in Avro tables to (2**31)-1 bytes.
- If a query encounters a STRING value longer than (2**31)-1 bytes in an Avro table, the query fails.

Logical types:

Avro type	Impala type
BYTES annotated	DECIMAL
INT32 annotated	DATE

Avro types not supported by Impala

- RECORD
- MAP
- ARRAY
- UNION
- ENUM
- FIXED
- NULL

Impala types not supported by Avro:

- TIMESTAMP

Impala issues warning messages if there are any mismatches between the types specified in the SQL column definitions and the underlying types; for example, any TINYINT or SMALLINT columns are treated as INT in the underlying Avro files, and therefore are displayed as INT in any DESCRIBE or SHOW CREATE TABLE output.



Note:

Currently, Avro tables cannot contain TIMESTAMP columns. If you need to store date and time values in Avro tables, as a workaround you can use a STRING representation of the values, convert the values to BIGINT with the UNIX_TIMESTAMP() function, or create separate numeric columns for individual date and time fields using the EXTRACT() function.

Using a Hive-Created Avro Table in Impala

If you have an Avro table created through Hive, you can use it in Impala as long as it contains only Impala-compatible data types. It cannot contain Avro types not supported by Impala, such as ENUM and FIXED Because Impala and Hive share the same metastore database, Impala can directly access the table definitions and data for tables that were created in Hive.

If you create an Avro table in Hive, issue an INVALIDATE METADATA in Impala. This is a one-time operation to make Impala aware of the new table. You can issue the statement while connected to any Impala node, and the catalog service broadcasts the change to all other Impala nodes.

If you load new data into an Avro table through Hive, either through a Hive LOAD DATA or INSERT statement, or by manually copying or moving files into the data directory for the table, issue a REFRESH *TABLE_NAME* statement the next time you connect to Impala through `impala-shell`.

If you issue the LOAD DATA statement through Impala, you do not need a REFRESH afterward.

Impala only supports fields of type BOOLEAN, INT, LONG, FLOAT, DOUBLE, and STRING, or unions of these types with null, for example, ["string", "null"]. Unions with null essentially create a nullable type.

Loading Data into Avro Tables

Currently, Impala cannot write Avro data files. Therefore, an Avro table cannot be used as the destination of an Impala INSERT statement or CREATE TABLE AS SELECT.

To copy data from another table, issue any INSERT statements through Hive.

After loading data into a table through Hive or other mechanism outside of Impala, issue a REFRESH *TABLE_NAME* statement the next time you connect to the Impala node, before querying the table, to make Impala recognize the new data.

If you already have data files in Avro format, you can also issue LOAD DATA in either Impala or Hive. Impala can move existing Avro data files into an Avro table, it just cannot create new Avro data files.

Enabling Compression for Avro Tables

To enable compression for Avro tables, specify settings in the Hive shell to enable compression and to specify a codec, then issue a CREATE TABLE statement as in the preceding examples. Impala supports the snappy and deflate codecs for Avro tables.

For example:

```
hive> set hive.exec.compress.output=true;
hive> set avro.output.codec=snappy;
```

Handling Avro Schema Evolution

Impala can handle with Avro data files that employ *schema evolution*, where different data files within the same table use slightly different type definitions. (You would perform the schema evolution operation by issuing an ALTER TABLE statement in the Hive shell.) The old and new types for any changed columns must be compatible, for example a column might start as an INT and later change to a BIGINT or FLOAT.

As with any other tables where the definitions are changed or data is added outside of the current `impalad` node, ensure that Impala loads the latest metadata for the table if the Avro schema is modified through Hive. Issue a REFRESH *TABLE_NAME* or INVALIDATE METADATA *TABLE_NAME* statement. REFRESH reloads the metadata immediately, INVALIDATE METADATA reloads the metadata the next time the table is accessed.

When Avro data files or columns are not consulted during a query, Impala does not check for consistency. Thus, if you issue SELECT c1, c2 FROM t1, Impala does not return any error if the column c3 changed in an incompatible way. If a query retrieves data from some partitions but not others, Impala does not check the data files for the unused partitions.

In the Hive DDL statements, you can specify an avro.schema.literal table property (if the schema definition is short) or an avro.schema.url property (if the schema definition is long, or to allow convenient editing for the definition).

Query Performance for Avro Tables

In general, expect query performance with Avro tables to be faster than with tables using text data, but slower than with Parquet tables.

Using RCFile Data Files

Impala supports creating and querying RCFile tables.

Creating RCFile Tables and Loading Data

To create a table in the RCFile format, use the STORED AS RCFILE clause in the CREATE TABLE statement.

Because Impala can query RCFile tables but cannot currently write to, after creating RCFile tables, use the Hive shell to load the data.

After loading data into a table through Hive or other mechanism outside of Impala, issue a REFRESH *TABLE_NAME* statement the next time you connect to the Impala node, before querying the table, to make Impala recognize the new data.

Query Performance for RCFile Tables

In general, expect query performance with RCFile tables to be faster than with tables using text data, but slower than with Parquet tables.

Enabling Compression for RCFile Tables

You may want to enable compression on existing tables. Enabling compression provides performance gains in most cases and is supported for RCFile tables.

The compression type is specified in the SET mapred.output.compression.codec command.

For example, to enable Snappy compression, you would specify the following additional settings when loading data through the Hive shell.

```
SET mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCode
c;
```

Using SequenceFile Data Files

Impala supports creating and querying SequenceFile tables.

Creating SequenceFile Tables and Loading Data

To create a table in the SequenceFile format, use the STORED AS SEQUENCEFILE clause in the CREATE T ABLE statement.

Because Impala can query SequenceFile tables but cannot currently write to, after creating SequenceFile tables, use the Hive shell to load the data.

After loading data into a table through Hive or other mechanism outside of Impala, issue a REFRESH *TABLE_NAME* statement the next time you connect to the Impala node, before querying the table, to make Impala recognize the new data.

Enabling Compression for SequenceFile Tables

You may want to enable compression on existing tables. Enabling compression provides performance gains in most cases and is supported for SequenceFile tables.

The compression type is specified in the SET mapred.output.compression.codec command.

For example, to enable Snappy compression, you would specify the following additional settings when loading data through the Hive shell.

```
SET mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCode
c;
```

Query Performance for SequenceFile Tables

In general, expect query performance with SequenceFile tables to be faster than with tables using text data, but slower than with Parquet tables.

Storage Systems Supports

This section describes the storage systems that Impala supports.

Impala with HDFS

You can use Impala to query data files that reside on HDFS.

Although Impala typically works well with many large files in an HDFS storage system, there are times when you might perform some file cleanup to reclaim space, or advise developers on techniques to minimize space consumption and file duplication.

- Use compact binary file formats where practical. Numeric and time-based data in particular can be stored in more compact form in binary data files. Depending on the file format, various compression and encoding features can reduce file size even further. You can specify the STORED AS clause as part of the CREATE TABLE statement, or ALTER TABLE with the SET FILEFORMAT clause for an existing table or partition within a partitioned table.
- You manage underlying data files differently depending on whether the corresponding Impala table is defined as an internal or external table:
  - Use the DESCRIBE FORMATTED statement to check if a particular table is internal (managed by Impala) or external, and to see the physical location of the data files in HDFS.
  - For Impala-managed ("internal") tables, use DROP TABLE statements to remove data files.
  - For tables not managed by Impala ("external") tables, use appropriate HDFS-related commands such as hadoop fs, hdfs dfs, or distcp, to create, move, copy, or delete files within HDFS directories that are accessible by the impala user. Issue a REFRESH *TABLE_NAME* statement after adding or removing any files from the data directory of an external table.
  - Use external tables to reference HDFS data files in their original location. With this technique, you avoid copying the files, and you can map more than one Impala table to the same set of data files. When you drop the Impala table, the data files are left undisturbed.
  - Use the LOAD DATA statement to move HDFS files into the data directory for an Impala table from inside Impala, without the need to specify the HDFS path of the destination directory. This technique works for both internal and external tables.
- Make sure that the HDFS trashcan is configured correctly. When you remove files from HDFS, the space might not be reclaimed for use by other files until sometime later, when the trashcan is emptied.
- Drop all tables in a database before dropping the database itself.
- If an INSERT statement encounters an error, and you see a directory named .impala_insert_staging or _impala_insert_staging left behind in the data directory for the table, it might contain temporary data files taking up space in HDFS. You might be able to salvage these data files.

  For example, delete those files through commands such as hadoop fs or hdfs dfs to reclaim space before re-trying the INSERT. Issue DESCRIBE FORMATTED *TABLE_NAME* to see the HDFS path where you can check for temporary files.

Configuring Scratch Space for Spilling to Disk

Impala uses intermediate files during large sorts, joins, aggregations, or analytic function operations The files are removed when the operation finishes. By default, intermediate files are stored in the directory /tmp/impala-scratch.

You can specify locations of the intermediate files in one of the following ways:

- By starting the `impalad` daemon with the ##scratch_dirs=*PATH_TO_DIRECTORY* configuration option.
- By specifying a different location in the Cloudera Manager in the Impala Daemon Scratch Directories field.

With either option above:

- You can specify a single directory or a comma-separated list of directories.
- You can specify an optional a capacity quota per scratch directory using the colon (:) as the delimiter.

  The capacity quota of -1 or 0 is the same as no quota for the directory.
- The scratch directories must be on the local filesystem, not in HDFS.

- You might specify different directory paths for different hosts, depending on the capacity and speed of the available storage devices.

If there is less than 1 GB free on the filesystem where that directory resides, Impala still runs, but writes a warning message to its log.

Impala successfully starts (with a warning written to the log) if it cannot create or read and write files in one of the scratch directories.

The following are examples for specifying scratch directories.

Config option	Description
--scratch_dirs=/dir1,/dir2	Use /dir1 and /dir2 as scratch directories with no capacity quota.
--scratch_dirs=/dir1,/dir2:25G	Use /dir1 and /dir2 as scratch directories with no capacity quota on /dir1 and the 25GB quota on /dir2.
--scratch_dirs=/dir1:5MB,/dir2	Use /dir1 and /dir2 as scratch directories with the capacity quota of 5MB on /dir1 and no quota on /dir2.
--scratch_dirs=/dir1:-1,/dir2:0	Use /dir1 and /dir2 as scratch directories with no capacity quota.

Allocation from a scratch directory will fail if the specified limit for the directory is exceeded.

If Impala encounters an error reading or writing files in a scratch directory during a query, Impala logs the error, and the query fails.

Priority Based Scratch Directory Selection

The location of the intermediate files are configured by starting the impalad daemon with the flag ##scratch_dirs=*PATH_TO_DIRECTORY*. Currently this startup flag uses the configured scratch directories in a round robin fashion. Automatic selection of scratch directories in a round robin fashion may not always be ideal in every situation since these directories could come from different classes of storage system volumes having different performance characteristics (SSD vs HDD, local storage vs network attached storage, etc.). To optimize your workload, you have an option to configure the priority of the scratch directories based on your storage system configuration.

The scratch directories will be selected for spilling based on how you configure the priorities of the directories and if you provide the same priority for multiple directories then the directories will be selected in a round robin fashion.

The valid formats for specifying the priority directories are as shown here:

```
<dir-path>:<limit>:<priority>
<dir-path>::<priority>
```

Example:

```
/dir1:200GB:0
/dir1::0
```

The following formats use the default priority:

```
/dir1
/dir1:200GB
/dir1:200GB:
```

In the example below, dir1 will be used as a spill target until it is full and then dir2, dir3, and dir4 will be used in a round robin fashion.

```
##scratch_dirs=/dir1:200GB:0, /dir2:1024GB:1, /dir3:1024GB:1, /dir4:1024GB:1
```

Increasing Scratch Capacity

You can compress the data spilled to disk to increase the effective scratch capacity. You typically more than double capacity using compression and reduce spilling to disk. Use the --disk_spill_compression_codec and –-disk_spill_punch_holes startup options. The --disk_spill_compression_codec takes any value supported by the COMPRESSION_CODEC query option. The value is not case-sensitive. A value of ZSTD or LZ4 is recommended (default is NONE).

For example:

```
--disk_spill_compression_codec=LZ4
--disk_spill_punch_holes=true
```

If you set --disk_spill_compression_codec to a value other than NONE, you must set --disk_spill_punch_holes to true.

The hole punching feature supported by many file systems is used to reclaim space in scratch files during execution of a query that spills to disk. This results in lower scratch space requirements in many cases, especially when combined with disk spill compression. When this option is not enabled, scratch space is still recycled by a query, but less effectively in many cases.

You can specify a compression level for ZSTD only. For example:

```
--disk_spill_compression_codec=ZSTD:10
--disk_spill_punch_holes=true
```

Compression levels from 1 up to 22 (default 3) are supported for ZSTD. The lower the compression level, the faster the speed at the cost of compression ratio.

Configure Impala Daemon to spill to HDFS

Impala occasionally needs to use persistent storage for writing intermediate files during large sorts, joins, aggregations, or analytic function operations. If your workload results in large volumes of intermediate data being written, it is recommended to configure the heavy spilling queries to use a remote storage location rather than the local one. The advantage of using remote storage for scratch space is that it is elastic and can handle any amount of spilling.

Before you begin

- Identify the HDFS scratch directory where you want your new Impala to write the temporary data.
- Identify the IP address, host name, or service identifier of HDFS.
- Identify the port number of the HDFS NameNode (if not-default).
- Configure Impala to write temporary data to disk during query processing.

Configuring the Start-up Option in Impala daemon

You can use the Impalad start option scratch_dirs to specify the locations of the intermediate files.

Use the following format for this start up option:

```
--scratch_dirs="hdfs://AUTHORITY/PATH(:MAX_BYTES), LOCAL_BUFFER_DIR
 (,LOCAL_DIR…)"
```

- Where hdfs://*AUTHORITY*/*PATH* is the remote directory.
- *AUTHORITY* may include ip_address or hostname and port, or service_id.

- *MAX_BYTES* is optional.

Using the above format:

- You can specify only one remote directory. When you configure a remote directory, you must specify a local buffer directory as the buffer. However you can use multiple local directories with the remote directory. If you specify multiple local directories, the first local directory would be used as the local buffer directory.
- If you configure both remote and local directories, the remote directory is only used when the local directories are fully utilized.
- The size of a remote intermediate file could affect the query performance, and the value can be set by --remote _tmp_file_size=*SIZE* in the start-up option. The default size of a remote intermediate file is 16MB while the maximum is 512MB.

Examples

- A HDFS scratch dir with one local buffer dir, file size 64MB. The space of HDFS scratch dir is limited to 300G.

```
--scratch_dirs=hdfs://10.0.0.49:20500/tmp:300G,/local_buffer_dir --remot
e_tmp_file_size=64M
```

- A HDFS scratch dir with one local buffer dir limited to 512MB, and one local dir limited to 10GB. The space of HDFS scratch dir is limited to 300G. The HDFS NameNode uses its default port (8020).

```
--scratch_dirs=hdfs://hdfsnn/tmp:300G,/local_buffer_dir:512M,/local_dir:
10G
```

- A HDFS scratch dir with one local buffer dir, and multiple prioritized local dirs. The space of HDFS scratch dir is unlimited. The HDFS service identifier is hdfs1.

```
--scratch_dirs=hdfs://hdfs1/tmp,/local_buffer_dir,/local_dir_1:5G:1,/loc
al_dir_2:5G:2
```

Even though max_bytes is optional, it is highly recommended to configure for spilling to HDFS because the HDFS cluster space is limited.

Impala with Kudu

You can use Impala to query tables stored by Apache Kudu. This capability allows convenient access to a storage system that is tuned for different kinds of workloads than the default with Impala.

By default, Impala tables are stored on HDFS. HDFS files are ideal for bulk loads (append operations) and queries using full-table scans, but do not support in-place updates or deletes. Kudu is an alternative storage engine used by Impala which can do both in-place updates (for mixed read/write workloads) and fast scans (for data-warehouse/ analytic operations). Using Kudu tables with Impala can simplify the ETL pipeline by avoiding extra steps to segregate and reorganize newly arrived data.

The underlying storage is managed and organized by Kudu, not represented as HDFS data files.

The combination of Kudu and Impala works best for tables where scan performance is important, but data arrives continuously, in small batches, or needs to be updated without being completely replaced. HDFS-backed tables can require substantial overhead to replace or reorganize data files as new data arrives. Impala can perform efficient lookups and scans within Kudu tables, and Impala can also perform update or delete operations efficiently. You can also use the Kudu Java, C++, and Python APIs to do ingestion or transformation operations outside of Impala, and Impala can query the current data at any time.

With HDFS-backed tables, you are typically concerned with the number of DataNodes in the cluster, how many and how large HDFS data files are read during a query, and therefore the amount of work performed by each DataNode and the network communication to combine intermediate results and produce the final result set.

Certain Impala SQL statements and clauses, such as DELETE, UPDATE, UPSERT, and PRIMARY KEY work only with Kudu tables.

Other statements and clauses, such as LOAD DATA, TRUNCATE TABLE, and INSERT OVERWRITE, are not applicable to Kudu tables.

Query performance for Kudu tables

For queries involving Kudu tables, Impala can delegate much of the work of filtering the result set to Kudu, avoiding some of the I/O involved in full table scans of tables containing HDFS data files. This type of optimization is especially effective for partitioned Kudu tables, where the Impala query WHERE clause refers to one or more primary key columns that are also used as partition key columns. For example, if a partitioned Kudu table uses a HASH clause for col1 and a RANGE clause for col2, a query using a clause such as WHERE col1 IN (1,2,3) AN D col2 > 100 can determine exactly which tablet servers contain relevant data, and therefore parallelize the query efficiently.

Impala can push down additional information to optimize join queries involving Kudu tables. If the join clause contains predicates of the form *COLUMN* = *EXPRESSION*, after Impala constructs a hash table of possible matching values for the join columns from the bigger table (either an HDFS table or a Kudu table), Impala can “push down” the minimum and maximum matching column values to Kudu, so that Kudu can more efficiently locate matching rows in the second (smaller) table. These min/max filters are affected by the RUNTIME_FILTER_MODE, RUNTIME_FILTER_WAIT_TIME_MS, and DISABLE_ROW_RUNTIME_FILTERING query options; the min/ max filters are not affected by the RUNTIME_BLOOM_FILTER_SIZE, RUNTIME_FILTER_MIN_SIZE, RUNT IME_FILTER_MAX_SIZE, and MAX_NUM_RUNTIME_FILTERS query options.

The TABLESAMPLE clause of the SELECT statement does not apply to a table reference derived from a view, a subquery, or anything other than a real base table. This clause only works for tables backed by HDFS or HDFS-like data files, therefore it does not apply to Kudu tables.

Configuring for Kudu Tables

The Kudu configuration property must be set for Impala to connect to the appropriate Kudu server.

Procedure

1. In Cloudera Manager, navigate to Clusters Impala Service .
2. In the Configuration tab, specify the Kudu service you want to use in the Kudu Service field.

   Typically, the required value for this setting is *KUDU_HOST*:7051. In a high-availability Kudu deployment, specify the names of multiple Kudu hosts separated by commas.
3. If the Kudu Service field is not set, you can still associate the appropriate value for each table by specifying a TBLPROPERTIES('kudu.master_addresses') clause in the CREATE TABLE statement or changing the TBLP ROPERTIES('kudu.master_addresses') value with an ALTER TABLE statement.

Impala DDL for Kudu

You can use the Impala CREATE TABLE and ALTER TABLE statements to create and fine-tune the characteristics of Kudu tables. Impala supports specific features and properties that only apply to Kudu tables.

Non-unique Primary Keys for Kudu Tables

Kudu now allows a user to create a non-unique primary key for a table when creating a table. The data engine handles this by appending a system generated auto-incrementing column to the non-unique primary key columns. This is done to guarantee the uniqueness of the primary key. This auto-incrementing column is named as 'auto_incrementing_id' with bigint type and this column is only system generated and cannot be explicitly created by the user. This auto_incrementing_id column is unique across a partition/tablet i.e. every partition/tablet would have this column starting from one and incrementing monotonically. The assignment to this column during insertion is automatic.

Create a Kudu Table with a non-unique PRIMARY KEY

The following example shows creating a table with a non-unique PRIMARY KEY.

```
CREATE TABLE kudu_tbl1
(
 id INT NON UNIQUE PRIMARY KEY,
 name STRING
)
 PARTITION BY HASH (id) PARTITIONS 3 STORED as KUDU;
```

The effective PRIMARY KEY in the above case will be {id, auto_increment_id}

Note: "auto_incrementing_id" column cannot be added, removed or renamed with ALTER TABLE statements.

Verify the PRIMARY KEY is non-unique

You can now check the PRIMARY KEY created is non-unique by running the following DESCRIBE command. A new property "key_unique" shows if the primary key is unique. System generated column "auto_incrementing_id" is shown in the output for the table as a non-unique primary key.

```
describe kudu_tbl1
+----------------------+--------+---------+-------------+------------+---
-------+---------------+---------------+----------------------+------------+
| name                 | type   | comment | primary_key | key_unique | nulla
ble | default_value | encoding      | compression          | block_size |
+----------------------+--------+---------+-------------+------------+-----
-----+---------------+---------------+----------------------+------------+
| id                   | int    |         | true        | false      | false
    |               | AUTO_ENCODING | DEFAULT_COMPRESSION | 0          |
| auto_incrementing_id | bigint |         | true        | false      | fa
lse    |               | AUTO_ENCODING | DEFAULT_COMPRESSION | 0          |
| name                 | string |         | false       |            | true
    |               | AUTO_ENCODING | DEFAULT_COMPRESSION | 0          |
+----------------------+--------+---------+-------------+------------+-----
-----+---------------+---------------+----------------------+------------+
Fetched 3 row(s) in 4.72s
```

Query Auto Incrementing Column

When you query a table using the SELECT statement, it will not display the system generated auto incrementing column unless the column is explicitly specified in the select list.

Create a Kudu table without a PRIMARY KEY attribute

You can create a Kudu table without specifying a PRIMARY KEY or a PARTITION KEY since they are optional, however you cannot create a Kudu table without specifying both PRIMARY KEY and PARTITION KEY. If you do not specify the primary key attribute, the partition key columns can be promoted as a non-unique primary key. This is possible only if those columns are the beginning columns of the table.

In the following example, 'a' and 'b' will be promoted as a non-unique primary key, 'auto_incrementing_id' column will be added by Kudu engine. 'a', 'b' and 'auto_incrementing_id' form the effective unique composite primary key.

```
CREATE TABLE auto_table
(
a BIGINT,
b STRING,
)
PARTITION BY HASH(a, b) PARTITIONS 2 STORED AS KUDU;
```

The effective primary key in this case would be {a, b, auto_incrementing_id}

Limitations

- UPSERT operation is not supported for Kudu tables with non-unique primary key. If you run an UPSERT statement for a Kudu table with a non-unique primary key it will fail with an error.
- Since the auto generated key for each row will be assigned after the row's data is generated and after the row lands in the tablet, you cannot use this column in the partition key.

PRIMARY KEY Attribute

To create a Kudu table, use the STORED AS KUDU clause in the CREATE TABLE statement.

The column list in a CREATE TABLE statement can include the following attributes, which only apply to Kudu tables:

```
[NON UNIQUE] PRIMARY KEY
| [NOT] NULL
| ENCODING CODEC
| COMPRESSION ALGORITHM
| DEFAULT CONSTANT_EXPRESSION
| BLOCK_SIZE NUMBER
```

The primary key for a Kudu table is a column, or set of columns, that uniquely identifies every row. The primary key value also is used as the natural sort order for the values from the table.

The notion of primary key only applies to Kudu tables. Every Kudu table requires a primary key.

Because all of the primary key columns must have non-null values, specifying a column in the PRIMARY KEY or NON-UNIQUE PRIMARY KEY clause implicitly adds the NOT NULL attribute to that column.

The primary key columns must be the first ones specified in the CREATE TABLE statement.

When the primary key is a single column, you can specify the PRIMARY KEY attribute either inline in a single column definition, or as a separate clause at the end of the column list.

For example:

```
        CREATE TABLE pk_inline
        (
        col1 BIGINT PRIMARY KEY,
        col2 STRING,
        col3 BOOLEAN
        ) PARTITION BY HASH(col1) PARTITIONS 2 STORED AS KUDU;

        CREATE TABLE pk_at_end
        (
        col1 BIGINT,
        col2 STRING,
        col3 BOOLEAN,
        PRIMARY KEY (col1)
        ) PARTITION BY HASH(col1) PARTITIONS 2 STORED AS KUDU;

        CREATE TABLE pk_inline
        (
        col1 BIGINT [NON UNIQUE] PRIMARY KEY,
        col2 STRING,
        col3 BOOLEAN
        ) PARTITION BY HASH(col1) PARTITIONS 2 STORED AS KUDU;

        CREATE TABLE pk_at_end
```

```
        (
        col1 BIGINT,
        col2 STRING,
        col3 BOOLEAN,
        [NON UNIQUE] PRIMARY KEY (col1)
        ) PARTITION BY HASH(col1) PARTITIONS 2 STORED AS KUDU;
```

The contents of the primary key columns cannot be changed by an UPDATE or UPSERT statement.

Including too many columns in the primary key (more than 5 or 6) can reduce the performance of write operations. Therefore, pick the most selective and most frequently tested non-null columns for the primary key specification. If a column must always have a value, but that value might change later, leave it out of the primary key and use a NOT NULL clause for that column instead.

NULL | NOT NULL Attribute

For Kudu tables, you can specify which columns can contain nulls or not. This constraint offers an extra level of consistency enforcement for Kudu tables. If an application requires a field to always be specified, include a NOT NULL clause in the corresponding column definition, and Kudu prevents rows from being inserted with a NULL in that column.

For example, a table containing geographic information might require the latitude and longitude coordinates to always be specified. Other attributes might be allowed to be NULL. For example, a location might not have a designated place name, its altitude might be unimportant, and its population might be initially unknown, to be filled in later.

For non-Kudu tables, Impala allows any column to contain NULL values, because it is not practical to enforce a “not null” constraint on HDFS data files that could be prepared using external tools and ETL processes.

```
CREATE TABLE required_columns
(
  id BIGINT PRIMARY KEY,
  latitude DOUBLE NOT NULL,
  longitude DOUBLE NOT NULL,
  place_name STRING,
  altitude DOUBLE,
  population BIGINT
) PARTITION BY HASH(id) PARTITIONS 2 STORED AS KUDU;
```

During performance optimization, Kudu can use the knowledge that nulls are not allowed to skip certain checks on each input row, speeding up queries and join operations. Therefore, specify NOT NULL constraints when appropriate.

The NULL clause is the default condition for all columns that are not part of the primary key. You can omit it, or specify it to clarify that you have made a conscious design decision to allow nulls in a column.

Because primary key columns cannot contain any NULL values, the NOT NULL clause is not required for the primary key columns, but you might still specify it to make your code self-describing.

DEFAULT Attribute

You can specify a default value for columns in Kudu tables. The default value can be any constant expression, for example, a combination of literal values, arithmetic and string operations. It cannot contain references to columns or non-deterministic function calls.

The following example shows different kinds of expressions for the DEFAULT clause. The requirement to use a constant value means that you can fill in a placeholder value such as NULL, empty string, 0, -1, 'N/A' and so on, but you cannot reference functions or column names. Therefore, you cannot use DEFAULT to do things such as

automatically making an uppercase copy of a string value, storing Boolean values based on tests of other columns, or add or subtract one from another column representing a sequence number.

```
CREATE TABLE default_vals
(
  id BIGINT PRIMARY KEY,
  name STRING NOT NULL DEFAULT 'unknown',
  address STRING DEFAULT upper('no fixed address'),
  age INT DEFAULT -1,
  earthling BOOLEAN DEFAULT TRUE,
  planet_of_origin STRING DEFAULT 'Earth',
  optional_col STRING DEFAULT NULL
) PARTITION BY HASH(id) PARTITIONS 2 STORED AS KUDU;
```

Note:

When designing an entirely new schema, prefer to use NULL as the placeholder for any unknown or missing values, because that is the universal convention among database systems. Null values can be stored efficiently, and easily checked with the IS NULL or IS NOT NULL operators. The DEFAULT attribute is appropriate when ingesting data that already has an established convention for representing unknown or missing values, or where the vast majority of rows have some common non-null value.

ENCODING Attribute

Each column in a Kudu table can optionally use an encoding, a low-overhead form of compression that reduces the size on disk, then requires additional CPU cycles to reconstruct the original values during queries. Typically, highly compressible data benefits from the reduced I/O to read the data back from disk.

The encoding keywords that Impala recognizes are:

- AUTO_ENCODING: Use the default encoding based on the column type, which are bitshuffle for the numeric type columns and dictionary for the string type columns.
- PLAIN_ENCODING: Leave the value in its original binary format.
- RLE: Compress repeated values (when sorted in primary key order) by including a count.
- DICT_ENCODING: When the number of different string values is low, replace the original string with a numeric ID.
- BIT_SHUFFLE: Rearrange the bits of the values to efficiently compress sequences of values that are identical or vary only slightly based on primary key order. The resulting encoded data is also compressed with LZ4.
- PREFIX_ENCODING: Compress common prefixes in string values; mainly for use internally within Kudu.

COMPRESSION Attribute

You can specify a compression algorithm to use for each column in a Kudu table. This attribute imposes more CPU overhead when retrieving the values than the ENCODING attribute does. Therefore, use it primarily for columns with long strings that do not benefit much from the less-expensive ENCODING attribute.

The choices for COMPRESSION are LZ4, SNAPPY, and ZLIB.

Note:

Columns that use the BITSHUFFLE encoding are already compressed using LZ4, and so typically do not need any additional COMPRESSION attribute.

BLOCK_SIZE Attribute

Although Kudu does not use HDFS files internally, and thus is not affected by the HDFS block size, it does have an underlying unit of I/O called the *block size*. The BLOCK_SIZE attribute lets you set the block size for any column.

The block size attribute is a relatively advanced feature. This is an unsupported feature and is considered experimental.

Kudu Replication Factor

By default, Kudu tables created through Impala use a tablet replication factor of 3. To change the replication factor for a Kudu table, specify the replication factor using the TBLPROPERTIES in the CREATE TABLE statement as below where *N* is the replication factor you want to use:

```
TBLPROPERTIES ('kudu.num_tablet_replicas' = 'n')
```

The number of replicas for a Kudu table must be odd.

Altering the kudu.num_tablet_replicas property after table creation currently has no effect.

How Impala Handles Kudu Metadata

Much of the metadata for Kudu tables is handled by the underlying storage layer. Kudu tables have less reliance on the Metastore database, and require less metadata caching on the Impala side. For example, information about partitions in Kudu tables is managed by Kudu, and Impala does not cache any block locality metadata for Kudu tables.

The REFRESH and INVALIDATE METADATA statements are needed less frequently for Kudu tables than for HDFS-backed tables. Neither statement is needed when data is added to, removed, or updated in a Kudu table, even if the changes are made directly to Kudu through a client program using the Kudu API.

Run REFRESH *TABLE_NAME* or INVALIDATE METADATA *TABLE_NAME* for a Kudu table only after making a change to the Kudu table schema, such as adding or dropping a column.

Because Kudu manages the metadata for its own tables separately from the Metastore database, there is a table name stored in the Metastore database for Impala to use, and a table name on the Kudu side, and these names can be modified independently through ALTER TABLE statements.

To avoid potential name conflicts, the prefix impala:: and the Impala database name are encoded into the underlying Kudu table name.

Partitioning for Kudu Tables

Kudu tables use special mechanisms to distribute data among the underlying tablet servers. Although referred as partitioned tables, they are distinguished from traditional Impala partitioned tables with the different syntax in CREA TE TABLE statement.

Kudu tables use PARTITION BY, HASH, RANGE, and range specification clauses rather than the PARTITIONED BY clause for HDFS-backed tables, which specifies only a column name and creates a new partition for each different value.

To see the current partitioning scheme for a Kudu table, you can use the SHOW CREATE TABLE statement or the SHOW PARTITIONS statement. The CREATE TABLE syntax displayed by this statement includes all the hash, range, or both clauses that reflect the original table structure plus any subsequent ALTER TABLE statements that changed the table structure.

To see the underlying buckets and partitions for a Kudu table, use the SHOW TABLE STATS or SHOW PARTITI ONS statement.

Hash Partitioning

Hash partitioning is the simplest type of partitioning for Kudu tables. For hash-partitioned Kudu tables, inserted rows are divided up between a fixed number of “buckets” by applying a hash function to the values of the columns specified in the HASH clause. Hashing ensures that rows with similar values are evenly distributed, instead of clumping together all in the same bucket. Spreading new rows across the buckets this way lets insertion operations work in parallel across multiple tablet servers. Separating the hashed values can impose additional overhead on queries, where queries with range-based predicates might have to read multiple tablets to retrieve all the relevant values.

```
-- 1M rows with 50 hash partitions = approximately 20,000 rows per partitio
n.
-- The values in each partition are not sequential, but rather based on a
 hash function.
-- Rows 1, 99999, and 123456 might be in the same partition.
CREATE TABLE million_rows (id string primary key, s string)
  PARTITION BY HASH(id) PARTITIONS 50
  STORED AS KUDU;

-- Because the ID values are unique, we expect the rows to be roughly
-- evenly distributed between the buckets in the destination table.
INSERT INTO million_rows SELECT * FROM billion_rows ORDER BY id LIMIT 1e6;
```

The largest number of buckets that you can create with a PARTITIONS clause varies depending on the number of tablet servers in the cluster, while the smallest is 2. For large tables, prefer to use roughly 10 partitions per server in the cluster.

Range Partitioning

Range partitioning lets you specify partitioning precisely, based on single values or ranges of values within one or more columns. You add one or more RANGE clauses to the CREATE TABLE statement, following the PARTITIO N BY clause. The RANGE clause includes a combination of constant expressions, VALUE or VALUES keywords, and comparison operators.

```
CREATE TABLE t1 (id STRING PRIMARY KEY, s STRING)
  PARTITION BY RANGE (PARTITION 'a' <= VALUES < '{', PARTITION 'A' <= VAL
UES < '[', PARTITION '00000' <= VALUES < '00001')
  STORED AS KUDU;
```



Note:

When defining ranges, be careful to avoid “fencepost errors” where values at the extreme ends might be included or omitted by accident. For example, in the tables defined in the preceding code listings, the range "a" <= VALUES < "{" ensures that any values starting with z, such as za or zzz or zzz-ZZZ, are all included, by using a less-than operator for the smallest value after all the values starting with z.

For range-partitioned Kudu tables, an appropriate range must exist before a data value can be created in the table. Any INSERT, UPDATE, or UPSERT statements fail if they try to create column values that fall outside the specified ranges. The error checking for ranges is performed on the Kudu side. Impala passes the specified range information to Kudu, and passes back any error or warning if the ranges are not valid. (A nonsensical range specification causes an error for a DDL statement, but only a warning for a DML statement.)

Partition ranges can be non-contiguous:

```
PARTITION BY RANGE (year) (PARTITION 1885 <= VALUES <= 1889, PARTITION 1893
 <= VALUES <= 1897)
```

The ALTER TABLE statement with the ADD PARTITION or DROP PARTITION clauses can be used to add or remove ranges from an existing Kudu table.

```
ALTER TABLE foo ADD PARTITION 30 <= VALUES < 50;
ALTER TABLE foo DROP RANGE PARTITION '1' <= VALUES < '5';
```

When a range is added, the new range must not overlap with any of the previous ranges; that is, it can only fill in gaps within the previous ranges.

When a range is removed, all the associated rows in the table are deleted regardless whether the table is internal or external.

Kudu tables can also use a combination of hash and range partitioning. For example:

```
PARTITION BY HASH (school) PARTITIONS 10,
  RANGE (letter_grade) (PARTITION VALUE = 'A', PARTITION VALUE = 'B',
    PARTITION VALUE = 'C', PARTITION VALUE = 'D', PARTITION VALUE = 'F')
```

Range partitioning with UTC timestamps: When you enable UTC timestamp conversion, Impala performs partitioning based on UTC values. If you add range partitions that include timestamps, such as during a CREATE T ABLE statement, you must define those ranges in UTC. Impala does not automatically convert local time to UTC for partition definitions

Custom hash partitioning for Kudu tables

You can now use the Impala CREATE TABLE and ALTER TABLE statements to allow custom hash partition specification at the range partition level. To display the hash schema specified for each partition, you can use the SHOW HASH SCHEMA statement.

HASH syntax within a partition is similar to the table-level syntax except that HASH clauses must follow the PARTITION clause and commas are not allowed within a partition. To keep the syntax consistent, commas in the table-level partition spec and between PARTITION clauses are now optional but allowed for backward compatibility.

Example:

```
CREATE TABLE t1 (id int, c2 int, PRIMARY KEY(id, c2))
PARTITION BY HASH(id) PARTITIONS 3 HASH(c2) PARTITIONS 4
RANGE (c2)
(
PARTITION 0 <= VALUES < 10
PARTITION 10 <= VALUES < 20
HASH(id) PARTITIONS 2 HASH(c2) PARTITIONS 3
PARTITION 20 <= VALUES < 30
)
STORED AS KUDU;


ALTER TABLE t1 ADD RANGE PARTITION 30 <= VALUES < 40
HASH(id) PARTITIONS 3 HASH(c2) PARTITIONS 4;
```

Creating external table pointing to an existing Kudu table using storage handler privileges

A user can now create an external Kudu table pointing to an existing Kudu table if the user is granted the RWSTORAGE privilege on the resource specified by a storage handler URI. Before this release, a user was required to have the ALL privilege on SERVER to create an external Kudu table for an existing Kudu table. This has been simplified by the introduction of a new type of resource called storage handler URI and a new access type called RWSTORAGE that will be supported by Apache Ranger.

For instance, to allow a user <usr> to create an external Kudu table based on an existing Kudu table 'impala::tpch_kudu.nation', it suffices for an administrator to execute the following command to grant the necessary privilege to the user <usr>, where "localhost" is the default address of the Kudu master host assuming there is only one single master host on the cluster.

```
GRANT RWSTORAGE ON
   STORAGEHANDLER_URI 'kudu://localhost/impala::tpch_kudu.nation'
   TO USER <usr>
```

For a user to create a new Kudu table in Impala, the RWSTORAGE privilege on a specific storage handler URI or the ALL privilege on SERVER are not required. A user is allowed to create such a Kudu table once the user is granted the privilege in Ranger's Hadoop SQL policy repository to create a table under the corresponding database.

For information on configuring a policy that allows authorized users to create data tables using storage handlers, see the link under Related Information.

Related Information

Configure a resource-based storage handler policy

Impala DML for Kudu Tables

In DML statements, Impala supports specific features and properties that only apply to Kudu tables.

The UPDATE and DELETE statements let you modify data within Kudu tables without rewriting substantial amounts of table data.

The UPSERT statement acts as a combination of INSERT and UPDATE, inserting rows where the primary key does not already exist, and updating the non-primary key columns where the primary key does already exist in the table.

The INSERT statement for Kudu tables honors the unique and NOT NULL requirements for the primary key columns.

Because Impala and Kudu do not support transactions, the effects of any INSERT, UPDATE, or DELETE statement are immediately visible. For example, you cannot do a sequence of UPDATE statements and only make the changes visible after all the statements are finished. Also, if a DML statement fails partway through, any rows that were already inserted, deleted, or changed remain in the table; there is no rollback mechanism to undo the changes.

In particular, an INSERT ... SELECT statement that refers to the table being inserted into might insert more rows than expected, because the SELECT part of the statement sees some of the new rows being inserted and processes them again.

Query hints:

The INSERT or UPSERT operations into Kudu tables automatically add an exchange and a sort node to the plan that partitions and sorts the rows according to the partitioning/primary key scheme of the target table (unless the number of rows to be inserted is small enough to trigger single node execution). Since Kudu partitions and sorts rows on write, pre-partitioning and sorting takes some of the load off of Kudu and helps large INSERT operations to complete without timing out. However, this default behavior may slow down the end-to-end performance of the INSERT or UPSERT operations.

You can use the /* +NOCLUSTERED */ and /* +NOSHUFFLE */ hints together to disable partitioning and sorting before the rows are sent to Kudu. Additionally, since sorting may consume a large amount of memory, consider setting the MEM_LIMIT query option for those queries.

Loading Data into Kudu Tables

Kudu tables are well-suited to use cases where data arrives continuously, in small or moderate volumes. To bring data into Kudu tables, use the Impala INSERT and UPSERT statements.

The LOAD DATA statement does not support Kudu tables.

Because Kudu manages its own storage layer that is optimized for smaller block sizes than HDFS, and performs its own maintenance to keep data evenly distributed, it is not subject to the “many small files” issue and does not need explicit reorganization and compaction as the data grows over time.

The partitions within a Kudu table can be specified to cover a variety of possible data distributions, instead of hardcoding a new partition for each new day, hour, and so on, which can lead to inefficient, hard-to-scale, and hard-to-manage partition schemes with HDFS tables.

Your strategy for performing ETL or bulk updates on Kudu tables should take into account the limitations on consistency for DML operations.

- The INSERT, UPDATE, and UPSERT operations should produce an identical result even when executed multiple times.
- If a bulk operation is in danger of exceeding capacity limits due to timeouts or high memory usage, split it into a series of smaller operations.

- Avoid running concurrent ETL operations where the end results depend on precise ordering. In particular, do not rely on an INSERT ... SELECT statement that selects from the same table into which it is inserting, unless you include extra conditions in the WHERE clause to avoid reading the newly inserted rows within the same statement.
- Because relationships between tables cannot be enforced by Impala and Kudu, and cannot be committed or rolled back together, do not expect transactional semantics for multi-table operations.

Consistency Considerations for Kudu Tables

Kudu tables have consistency characteristics such as uniqueness, controlled by the primary key columns and non-nullable columns. The emphasis for consistency is on preventing duplicate or incomplete data from being stored in a table.

Currently, Kudu does not enforce strong consistency for order of operations, total success or total failure of a multi-row statement, or data that is read while a write operation is in progress. Changes are applied atomically to each row, but not applied as a single unit to all rows affected by a multi-row DML statement. That is, Kudu does not currently have atomic multi-row statements or isolation between statements.

If some rows are rejected during a DML operation because of a mismatch with duplicate primary key values, NOT NULL constraints, and so on, the statement succeeds with a warning. Impala still inserts, deletes, or updates the other rows that are not affected by the constraint violation.

Consequently, the number of rows affected by a DML operation on a Kudu table might be different than you expect.

Because there is no strong consistency guarantee for information being inserted into, deleted from, or updated across multiple tables simultaneously, consider denormalizing the data where practical. That is, if you run separate INSERT statements to insert related rows into two different tables, one INSERT might fail while the other succeeds, leaving the data in an inconsistent state. Even if both inserts succeed, a join query might happen during the interval between the completion of the first and second statements, and the query would encounter incomplete inconsistent data. Denormalizing the data into a single wide table can reduce the possibility of inconsistency due to multi-table operations.

Information about the number of rows affected by a DML operation is reported in `impala-shell` output, and in the PROFILE output, but is not currently reported to HiveServer2 clients such as JDBC or ODBC applications.

Handling Date, Time, or Timestamp Data with Kudu

You can include TIMESTAMP columns in Kudu tables. The behavior of TIMESTAMP for Kudu tables has some special considerations:

- Any nanoseconds in the original 96-bit value produced by Impala are not stored, because Kudu represents date/time columns using 64-bit values. The nanosecond portion of the value is rounded, not truncated. Therefore, a TIMESTAMP value that you store in a Kudu table might not be bit-for-bit identical to the value returned by a query.
- The conversion between the Impala 96-bit representation and the Kudu 64-bit representation introduces some performance overhead when reading or writing TIMESTAMP columns. You can minimize the overhead during writes by performing inserts through the Kudu API. Because the overhead during reads applies to each query, you might continue to use a BIGINT column to represent date/time values in performance-critical applications.
- The Impala TIMESTAMP type has a narrower range for years than the underlying Kudu data type. Impala can represent years 1400-9999. If year values outside this range are written to a Kudu table by a non-Impala client, Impala returns NULL by default when reading those TIMESTAMP values during a query. Or, if the ABORT_ON_ERROR query option is enabled, the query fails when it encounters a value with an out-of-range year.

Impala with HBase

You can use Impala to query data residing in HBase tables, a key-value data store where the value consists of multiple fields. The key is mapped to one column in the Impala table, and the various fields of the value are mapped to the other columns in the Impala table.

HBase tables are the best suited in Impala in the following use cases.

- Storing rapidly incrementing counters, such as how many times a web page has been viewed, or on a social network, how many connections a user has or how many votes a post received.

  The append-only storage mechanism of HBase is efficient for writing each change to disk, and a query always returns the latest value. An application could query specific totals like these from HBase, and combine the results with a broader set of data queried from Impala.
- Storing very wide tables in HBase.

  Wide tables have many columns, possibly thousands, typically recording many attributes for an important subject such as a user of an online service. These tables are also often sparse, that is, most of the columns values are NULL, 0, false, empty string, or other blank or placeholder value. (For example, any particular web site user might have never used some site feature, filled in a certain field in their profile, visited a particular part of the site, and so on.) A typical query against this kind of table is to look up a single row to retrieve all the information about a specific subject, rather than summing, averaging, or filtering millions of rows as in typical Impala-managed tables.

HBase works out of the box with Impala. There is no mandatory configuration needed to use these two components together.

For efficient queries, use WHERE clauses to find a single key value or a range of key values wherever practical, by testing the Impala column corresponding to the HBase row key. Avoid queries that do full-table scans, which are efficient for regular Impala tables but inefficient in HBase.

To work with an HBase table from Impala, ensure that the impala user has read/write privileges for the HBase table, using the GRANT command in the HBase shell.

Creating HBase Tables for Impala

You create the tables on the Impala side using the Hive shell, because the Impala CREATE TABLE statement currently does not support custom SerDes and some other syntax needed for HBase tables.

- You create the new table as an HBase table using the STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' clause on the Hive CREATE TABLE statement.
- You map these specially created tables to corresponding tables that exist in HBase, with the clause TBLPROPERTIES("hbase.table.name" = "*TABLE_NAME_IN_HBASE*") on the Hive CREATE TABLE statement.
- You define the column corresponding to the HBase row key as a string with the #string keyword, or map it to a STRING column.
- After creating a new table through Hive, issue the INVALIDATE METADATA statement in Impala to make Impala aware of the new table.

Because Impala and Hive share the same Metastore database, once you create the table in Hive, you can query or insert into it through Impala.

Supported Data Types for HBase Columns

HBase does not enforce any typing for the key or value fields. All the type enforcement is done on the Impala side.

HBase row key

When creating the table through the Hive shell, use the STRING data type for the column that corresponds to the HBase row key.

For best performance of Impala queries against HBase tables, most queries will perform comparisons in the WHERE clause against the column that corresponds to the HBase row key.

Impala can translate predicates (through operators such as =, <, and BETWEEN) against this column into fast lookups in HBase, but this optimization (“predicate pushdown”) only works when that column is defined as STRING.

HBase numeric column

Define HBase numeric columns as the binary data type in Hive CREATE TABLE statement using the #binary (#b) keyword.

Defining numeric columns as binary can reduce the overall data volume in the HBase tables.

Impala also supports reading and writing to columns that are defined in the Hive CREATE TABLE statement using binary data types , represented in the Hive table definition using the #binary keyword, often abbreviated as #b.

Loading Data into HBase Tables

The Impala INSERT statement supports HBase tables. The INSERT ... VALUES syntax is ideally suited to HBase tables, because inserting a single row is an efficient operation for an HBase table.

When you use the INSERT ... SELECT syntax, the result in the HBase table could be fewer rows than you expect. HBase only stores the most recent version of each unique row key, so if an INSERT ... SELECT statement copies over multiple rows containing the same value for the key column, subsequent queries will only return one row with each key column value.

Successive INSERT statements using the same value for the key column achieves the same result as UPDATE.

Examples of Querying HBase Tables from Impala

The following examples create an HBase table with four column families, create a corresponding table through Hive, then insert and query the table through Impala.

In HBase, create a table. Table names are quoted in the CREATE statement in HBase.

```
hbase(main):001:0> CREATE 'hbasealltypessmall', 'boolsCF', 'intsCF', 'floats
CF', 'stringsCF'
```

Issue the following CREATE TABLE statement in the Hive shell.

This example creates an external table mapped to the HBase table, usable by both Impala and Hive. It is defined as an external table so that when dropped by Impala or Hive, the original HBase table is not touched at all.

The WITH SERDEPROPERTIES clause specifies that the first column (ID) represents the row key, and maps the remaining columns of the SQL table to HBase column families. The mapping relies on the ordinal order of the columns in the table, not the column names in the CREATE TABLE statement. The first column is defined to be the lookup key; the STRING data type produces the fastest key-based lookups for HBase tables.



Note: For Impala with HBase tables, the most important aspect to ensure good performance is to use a STRING column as the row key, as shown in this example.

```
hive> CREATE EXTERNAL TABLE hbasestringids (
    >   id string,
    >   bool_col boolean,
    >   tinyint_col tinyint,
    >   smallint_col smallint,
    >   int_col int,
    >   bigint_col bigint,
    >   float_col float,
    >   double_col double,
    >   date_string_col string,
    >   string_col string,
    >   timestamp_col timestamp)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES (
    >   "hbase.columns.mapping" =
    >   ":key,boolsCF:bool_col,intsCF:tinyint_col,intsCF:smallint_col,intsCF
:int_col,intsCF:\
    >   bigint_col,floatsCF:float_col,floatsCF:double_col,stringsCF:date_s
tring_col,\
    >   stringsCF:string_col,stringsCF:timestamp_col"
```

```
    > )
    > TBLPROPERTIES("hbase.table.name" = "hbasealltypessmall");
```

Once you have established the mapping to an HBase table, you can issue DML statements and queries from Impala. The following example shows a series of INSERT statements followed by a query. The ideal kind of query from a performance standpoint retrieves a row from the table based on a row key mapped to a string column.

The INVALIDATE METADATA statement makes the table created through Hive visible to Impala.

```
[impala] > INVALIDATE METADATA hbasestringids;

[impala] > INSERT INTO hbasestringids VALUES ('0001',true,3.141,9.94,1234567
,32768,4000,76,'2014-12-31','Hello world',NOW());

[impala] > IMSERT INTO hbasestringids VALUES ('0002',false,2.004,6.196,15
00,8000,129,127,'2014-01-01','Foo bar',NOW());

[impala] > SELECT * FROM hbasestringids WHERE id = '0001';
+------+----------+------------+-------------------+------------+---------
+--------------+-------------+-----------------+-------------+-------------
------------------+
| id   | bool_col | double_col | float_col         | bigint_col | int_col |
smallint_col | tinyint_col | date_string_col | string_col  | timestamp_col
             |
+------+----------+------------+-------------------+------------+---------+
--------------+-------------+-----------------+-------------+---------------
----------------+
| 0001 | true     | 3.141      | 9.939999580383301 | 1234567    | 32768   |
 4000         | 76          | 2014-12-31      | Hello world | 2015-02-10 16
:36:59.764838000 |
+------+----------+------------+-------------------+------------+---------
+--------------+-------------+-----------------+-------------+-------------
------------------+
```

Performance Considerations for the Impala-HBase Integration

Impala uses the HBase client API via Java Native Interface (JNI) to query data stored in HBase. This querying does not read HFiles directly. The extra communication overhead makes it important to choose what data to store in HBase or in HDFS, and construct efficient queries that can retrieve the HBase data efficiently:

- Use HBase table for queries that return a single row or a small range of rows, not queries that perform a full table scan of an entire table. (If a query has a HBase table and no WHERE clause referencing that table, that is a strong indicator that it is an inefficient query for an HBase table.)
- HBase may offer acceptable performance for storing small dimension tables where the table is small enough that executing a full table scan for every query is efficient enough. However, Kudu is almost always a superior alternative for storing dimension tables. HDFS tables are also appropriate for dimension tables that do not need to support update queries, delete queries or insert queries with small numbers of rows.

Query predicates are applied to row keys as start and stop keys, thereby limiting the scope of a particular lookup. If row keys are not mapped to string columns, then ordering is typically incorrect and comparison operations do not work. For example, if row keys are not mapped to string columns, evaluating for greater than (>) or less than (<) cannot be completed.

Predicates on non-key columns can be sent to HBase to scan as SingleColumnValueFilters, providing some performance gains. In such a case, HBase returns fewer rows than if those same predicates were applied using Impala. While there is some improvement, it is not as great when start and stop rows are used. This is because the number of rows that HBase must examine is not limited as it is when start and stop rows are used. As long as the row key predicate only applies to a single row, HBase will locate and return that row. Conversely, if a non-key predicate is used, even if it only applies to a single row, HBase must still scan the entire table to find the correct result.

Limitations and Restrictions of the Impala and HBase Integration

The Impala integration with HBase has the following limitations and restrictions, some inherited from the integration between HBase and Hive, and some unique to Impala:

- If you issue a DROP TABLE for an internal (Impala-managed) table that is mapped to an HBase table, the underlying table is not removed in HBase. The Hive DROP TABLE statement removes the HBase table in this case.
- The INSERT OVERWRITE statement is not available for HBase tables. You can insert new data, or modify an existing row by inserting a new row with the same key value, but not replace the entire contents of the table. You can do an INSERT OVERWRITE in Hive if you need this capability.
- If you issue a CREATE TABLE LIKE statement for a table mapped to an HBase table, the new table is also an HBase table, but inherits the same underlying HBase table name as the original. The new table is effectively an alias for the old one, not a new table with identical column structure. Avoid using CREATE TABLE LIKE for HBase tables, to avoid any confusion.
- Copying data into an HBase table using the Impala INSERT ... SELECT syntax might produce fewer new rows than are in the query result set. If the result set contains multiple rows with the same value for the key column, each row supercedes any previous rows with the same key value. Because the order of the inserted rows is unpredictable, you cannot rely on this technique to preserve the “latest” version of a particular key value.
- The LOAD DATA statement cannot be used with HBase tables.
- The TABLESAMPLE clause of the SELECT statement does not suppport HBase tables.

Impala with Azure Data Lake Store (ADLS)

You can use Impala to query data residing on the Azure Data Lake Store (ADLS) filesystem and Azure Blob File System (ABFS). This capability allows convenient access to a storage system that is remotely managed, accessible from anywhere, and integrated with various cloud-based services.

Impala can query files in any supported file format from ADLS. The ADLS storage location can be for an entire table or individual partitions in a partitioned table.

The default Impala tables use data files stored on HDFS, which are ideal for bulk loads and queries using full-table scans. In contrast, queries against ADLS data are less performant, making ADLS suitable for holding “cold” data that is only queried occasionally, while more frequently accessed “hot” data resides in HDFS. In a partitioned table, you can set the LOCATION attribute for individual partitions to put some partitions on HDFS and others on ADLS, typically depending on the age of the data.

Impala requires that the default filesystem for the cluster be HDFS. You cannot use ADLS as the only filesystem in the cluster.

To be able to access ADLS, first set up an Azure account, configure an ADLS store, and configure your cluster with appropriate credentials.

Creating Impala Databases, Tables, and Partitions for Data Stored on ADLS

To create a table that resides on ADLS, specify the ADLS details in the LOCATION clause of the CREATE TABLE or ALTER TABLE statement. The syntax for the LOCATION clause is:

- For ADLS Gen1:

```
LOCATION 'adl://ACCOUNT.azuredatalakestore.net/PATH/FILE'
```

- For ADLS Gen2:

```
LOCATION 'abfs://CONTAINER@ACCOUNT.dfs.core.windows.net/PATH/FILE'
```

or

```
LOCATION 'abfss://CONTAINER@ACCOUNT.dfs.core.windows.net/PATH/FILE'
```

CONTAINER denotes the parent location that holds the files and folders, which is the Containers in the Azure Storage Blobs service.

ACCOUNT is the name given for your storage account.

Any reference to an ADLS location must be fully qualified. (This rule applies when ADLS is not designated as the default filesystem.)

Once a table or partition is designated as residing on ADLS, the SELECT statement transparently accesses the data files from the appropriate storage layer.

ALTER TABLE can also set the LOCATION property for an individual partition so that some data in a table resides on ADLS and other data in the same table resides on HDFS.



Note:

By default, TLS is enabled both with abfs:// and abfss://.

When you set the fs.azure.always.use.https=false property, TLS is disabled with abfs://, and TLS is enabled with abfss://

For a partitioned table, either specify a separate LOCATION clause for each new partition, or specify a base LOCA TION for the table and set up a directory structure in ADLS to mirror the way Impala partitioned tables are structured in HDFS.

Although, strictly speaking, ADLS filenames do not have directory paths, Impala treats ADLS filenames with / characters the same as HDFS pathnames that include directories.

To point a nonpartitioned table or an individual partition at ADLS, specify a single directory path in ADLS, which could be any arbitrary directory.

To replicate the structure of an entire Impala partitioned table or database in ADLS requires more care, with directories and subdirectories nested and named to match the equivalent directory tree in HDFS. Consider setting up an empty staging area if necessary in HDFS, and recording the complete directory structure so that you can replicate it in ADLS.

For example, the following session creates a partitioned table where only a single partition resides on ADLS. The partitions for years 2013 and 2014 are located on HDFS. The partition for year 2015 includes a LOCATION attribute with an adl:// URL, and so refers to data residing on ADLS, under a specific path underneath the store impalademo.

```
CREATE TABLE mostly_on_hdfs (x INT) PARTITIONED BY (year INT);
ALTER TABLE mostly_on_hdfs ADD PARTITION (year=2013);
ALTER TABLE mostly_on_hdfs ADD PARTITION (year=2014);
ALTER TABLE mostly_on_hdfs ADD PARTITION (year=2015)
   LOCATION 'adl://impalademo.azuredatalakestore.net/dir1/dir2/dir3/t1';
```

When working with multiple tables with data files stored in ADLS, you can create a database with the LOCATION attribute pointing to an ADLS path. Specify a URL of the form as shown above. Any tables created inside that database automatically create directories underneath the one specified by the database LOCATION attribute.

Use the standard ADLS file upload methods to actually put the data files into the right locations. You can also put the directory paths and data files in place before creating the associated Impala databases or tables, and Impala automatically uses the data from the appropriate location after the associated databases and tables are created.

You can switch whether an existing table or partition points to data in HDFS or ADLS. For example, if you have an Impala table or partition pointing to data files in HDFS or ADLS, and you later transfer those data files to the other filesystem, use an ALTER TABLE statement to adjust the LOCATION attribute of the corresponding table or partition to reflect that change. This location-switching technique is not practical for entire databases that have a custom LOCATION attribute.

You cannot use the ALTER TABLE ... SET CACHED statement for tables or partitions that are located in ADLS.

Internal and External Tables Located on ADLS

Just as with tables located on HDFS storage, you can designate ADLS-based tables as either internal (managed by Impala) or external, by using the syntax CREATE TABLE or CREATE EXTERNAL TABLE respectively.

When you drop an internal table, the files associated with the table are removed, even if they are on ADLS storage. When you drop an external table, the files associated with the table are left alone, and are still available for access by other tools or components.

If the data on ADLS is intended to be long-lived and accessed by other tools in addition to Impala, create any associated ADLS tables with the CREATE EXTERNAL TABLE syntax, so that the files are not deleted from ADLS when the table is dropped.

If the data on ADLS is only needed for querying by Impala and can be safely discarded once the Impala workflow is complete, create the associated ADLS tables using the CREATE TABLE syntax so that dropping the table also deletes the corresponding data files on ADLS.

Loading Data into ADLS for Impala Queries

If your ETL pipeline involves moving data into ADLS and then querying through Impala, you can either use Impala DML statements to create, move, or copy the data, or use the same data loading techniques as you would for non-Impala data.

Using Impala DML Statements for ADLS Data:

The Impala DML statements (INSERT, LOAD DATA, and CREATE TABLE AS SELECT) can write data into a table or partition that resides in the Azure Data Lake Store (ADLS) or ADLS Gen2.

Manually Loading Data into Impala Tables on ADLS:

You can use the Microsoft-provided methods to bring data files into ADLS for querying through Impala. See the Microsoft ADLS documentation for details.

After you upload data files to a location already mapped to an Impala table or partition, or if you delete files in ADLS from such a location, issue the REFRESH statement to make Impala aware of the new set of data files.

Running and Queries for Data Stored on ADLS

Once the appropriate LOCATION attributes are set up at the table or partition level, you query data stored in ADLS the same as data stored on HDFS or in HBase:

- Queries against ADLS data support all the same file formats as for HDFS data.
- Tables can be unpartitioned or partitioned. For partitioned tables, either manually construct paths in ADLS corresponding to the HDFS directories representing partition key values, or use ALTER TABLE ... ADD PARTITION to set up the appropriate paths in ADLS.
- HDFS, Kudu, and HBase tables can be joined to ADLS tables, or ADLS tables can be joined with each other.
- Authorization to control access to databases, tables, or columns works the same whether the data is in HDFS or in ADLS.
- The `catalogd` daemon caches metadata for both HDFS and ADLS tables. Use REFRESH and INVALIDATE METADATA for ADLS tables in the same situations where you would issue those statements for HDFS tables.
- Queries against ADLS tables are subject to the same kinds of admission control and resource management as HDFS tables.
- Metadata about ADLS tables is stored in the same Metastore database as for HDFS tables.
- You can set up views referring to ADLS tables, the same as for HDFS tables.
- The COMPUTE STATS, SHOW TABLE STATS, and SHOW COLUMN STATS statements support ADLS tables.

Query Performance for ADLS Data

Although Impala queries for data stored in ADLS might be less performant than queries against the equivalent data stored in HDFS, you can still do some tuning. Here are techniques you can use to interpret explain plans and profiles for queries against ADLS data, and tips to achieve the best performance possible for such queries.

All else being equal, performance is expected to be lower for queries running against data on ADLS rather than HDFS. The actual mechanics of the SELECT statement are somewhat different when the data is in ADLS. Although the work is still distributed across the datanodes of the cluster, Impala might parallelize the work for a distributed query differently for data on HDFS and ADLS. ADLS does not have the same block notion as HDFS, so Impala uses heuristics to determine how to split up large ADLS files for processing in parallel. Because all hosts can access any ADLS data file with equal efficiency, the distribution of work might be different than for HDFS data, where the data blocks are physically read using short-circuit local reads by hosts that contain the appropriate block replicas. Although the I/O to read the ADLS data might be spread evenly across the hosts of the cluster, the fact that all data is initially retrieved across the network means that the overall query performance is likely to be lower for ADLS data than for HDFS data.

Because data files written to ADLS do not have a default block size the way HDFS data files do, any Impala INSE RT or CREATE TABLE AS SELECT statements use the PARQUET_FILE_SIZE query option setting to define the size of Parquet data files. (Using a large block size is more important for Parquet tables than for tables that use other file formats.)

When optimizing aspects of for complex queries such as the join order, Impala treats tables on HDFS and ADLS the same way.

In query profile reports, the numbers for BytesReadLocal, BytesReadShortCircuit, BytesReadDataNodeCached, and BytesReadRemoteUnexpected are blank because those metrics come from HDFS.

All the I/O for ADLS tables involves remote reads, and they will appear as “remote read” operations in the query profile.

Impala with Amazon S3

You can use Impala to query data residing on the Amazon S3 object store. This capability allows convenient access to a storage system that is remotely managed, accessible from anywhere, and integrated with various cloud-based services.

Impala can query files in any supported file format from S3. The S3 storage location can be for an entire table, or individual partitions in a partitioned table.

Best Practices for Using Impala with S3

The following guidelines summarize the best practices described in the rest of this topic:

- Any reference to an S3 location must be fully qualified when S3 is not designated as the default storage, for example, s3a:://[s3-bucket-name].
- Set the fs.s3a.connection.maximum safety valve setting, to 1500 for `impalad`.
- Set the fs.s3a.block.size safety valve setting to 134217728 (128 MB in bytes) if most Parquet files queried by Impala were written by Hive or ParquetMR jobs.
- Set the PARQUET_OBJECT_STORE_SPLIT_SIZE query option to 268435456 (256 MB in bytes) if most Parquet files queried by Impala were written by Impala.
- DROP TABLE .. PURGE is much faster than the default DROP TABLE. The same applies to ALTER TA BLE ... DROP PARTITION PURGE versus the default DROP PARTITION operation. However, due to the eventually consistent nature of S3, the files for that table or partition could remain for some unbounded time when using PURGE. The default DROP TABLE/PARTITION is slow because Impala copies the files to the S3A trashcan, and Impala waits until all the data is moved. DROP TABLE/PARTITION .. PURGE is a fast delete operation, and the Impala statement finishes quickly even though the change might not have propagated fully throughout S3.

- INSERT statements are faster than INSERT OVERWRITE for S3. The S3_SKIP_INSERT_STAGING query option, which is set to true by default, skips the staging step for regular INSERT (but not INSERT OVERWRITE). This makes the operation much faster, but consistency is not guaranteed: if a node fails during execution, the table could end up with inconsistent data. Set this option to false if stronger consistency is required, but this setting will make the INSERT operations slower.
  - For Impala-ACID tables, both INSERT and INSERT OVERWRITE tables for S3 are fast, regardless of the setting of S3_SKIP_INSERT_STAGING. Plus, consistency is guaranteed with ACID tables.
- Enable data cache for remote reads.
- Too many files in a table can make metadata loading and updating slow in S3. If too many requests are made to S3, S3 has a back-off mechanism and responds slower than usual.
  - If you have many small files due to over-granular partitioning, configure partitions with many megabytes of data so that even a query against a single partition can be parallelized effectively.
  - If you have many small files because of many small INSERT queries, use bulk INSERTs so that more data is written to fewer files.

Creating Impala Databases, Tables, and Partitions for Data Stored in S3

To create a table that resides in S3, run the CREATE TABLE or ALTER TABLE statement with the LOCATION clause.

ALTER TABLE can set the LOCATION property for an individual partition, so that some data in a table resides in S3 and other data in the same table resides on HDFS.

The syntax for the LOCATION clause is:

```
LOCATION 's3a://BUCKET_NAME/PATH/TO/FILE'
```

The file system prefix is always s3a://. Impala does not support the s3:// or s3n:// prefixes.

For a partitioned table, either specify a separate LOCATION clause for each new partition, or specify a base LOCATION for the table and set up a directory structure in S3 to mirror the way Impala partitioned tables are structured in S3.

You point a nonpartitioned table or an individual partition at S3 by specifying a single directory path in S3, which could be any arbitrary directory. To replicate the structure of an entire Impala partitioned table or database in S3 requires more care, with directories and subdirectories nested and named to match the equivalent directory tree in HDFS. Consider setting up an empty staging area if necessary in HDFS, and recording the complete directory structure so that you can replicate it in S3.

When working with multiple tables with data files stored in S3, you can create a database with a LOCATION attribute pointing to an S3 path. Specify a URL of the form s3a://*BUCKET*/*ROOT*/*PATH*/*FOR*/*DATABASE* for the LOCATION attribute of the database. Any tables created inside that database automatically create directories underneath the one specified by the database LOCATION attribute.

The following example creates a table with one partition for the year 2017 resides on HDFS and one partition for the year 2018 resides in S3.

The partition for year 2018 includes a LOCATION attribute with an s3a:// URL, and so refers to data residing in S3, under a specific path underneath the bucket impala-demo.

```
CREATE TABLE mostly_on_hdfs (x int) PARTITIONED BY (year INT);
ALTER TABLE mostly_on_hdfs ADD PARTITION (year=2017);
ALTER TABLE mostly_on_hdfs ADD PARTITION (year=2018)
   LOCATION 's3a://impala-demo/dir1/dir2/dir3/t1';
```

The following session creates a database and two partitioned tables residing entirely in S3, one partitioned by a single column and the other partitioned by multiple columns.

- Because a LOCATION attribute with an s3a:// URL is specified for the database, the tables inside that database are automatically created in S3 underneath the database directory.

- To see the names of the associated subdirectories, including the partition key values, use an S3 client tool to examine how the directory structure is organized in S3.

```
CREATE DATABASE db_on_s3 LOCATION 's3a://impala-demo/dir1/dir2/dir3';
CREATE TABLE partitioned_multiple_keys (x INT)
   PARTITIONED BY (year SMALLINT, month TINYINT, day TINYINT);
ALTER TABLE partitioned_multiple_keys
   ADD PARTITION (year=2015,month=1,day=1);
ALTER TABLE partitioned_multiple_keys
   ADD PARTITION (year=2015,month=1,day=31);

!hdfs dfs -ls -R s3a://impala-demo/dir1/dir2/dir3
2015-03-17 13:56:34          0 dir1/dir2/dir3/
2015-03-17 16:47:13          0 dir1/dir2/dir3/partitioned_multiple_keys/
2015-03-17 16:47:44          0 dir1/dir2/dir3/partitioned_multiple_keys/yea
r=2015/month=1/day=1/
2015-03-17 16:47:50          0 dir1/dir2/dir3/partitioned_multiple_keys/ye
ar=2015/month=1/day=31/
```

The CREATE DATABASE and CREATE TABLE statements create the associated directory paths if they do not already exist. You can specify multiple levels of directories, and the CREATE statement creates all appropriate levels, similar to using mkdir -p.

Use the standard S3 file upload methods to put the actual data files into the right locations. You can also put the directory paths and data files in place before creating the associated Impala databases or tables, and Impala automatically uses the data from the appropriate location after the associated databases and tables are created.

Use the ALTER TABLE statement with the LOCATION clause to switch whether an existing table or partition points to data in HDFS or S3. For example, if you have an Impala table or partition pointing to data files in HDFS or S3, and you later transfer those data files to the other filesystem, use the ALTER TABLE statement to adjust the LOCATION attribute of the corresponding table or partition to reflect that change.

Internal and External Tables Located in S3

Just as with tables located on HDFS storage, you can designate S3-based tables as either internal (managed by Impala) or external with the CREATE TABLE or CREATE EXTERNAL TABLE statement respectively.

When you drop an internal table, the files associated with the table are removed, even if they are in S3 storage. When you drop an external table, the files associated with the table are left alone, and are still available for access by other tools or components.

If the data in S3 is intended to be long-lived and accessed by other tools in addition to Impala, create any associated S3 tables with the CREATE EXTERNAL TABLE statement, so that the files are not deleted from S3 when the table is dropped.

If the data in S3 is only needed for querying by Impala and can be safely discarded once the Impala workflow is complete, create the associated S3 tables using the CREATE TABLE statement, so that dropping the table also deletes the corresponding data files in S3.

Loading Data into S3 for Impala Queries

If your ETL pipeline involves moving data into S3 and then querying through Impala, you can either use Impala DML statements to create, move, or copy the data, or use the same data loading techniques as you would for non-Impala data.

Using Impala DML Statements for S3 Data:

Impala DML statements (INSERT, LOAD DATA, and CREATE TABLE AS SELECT) can write data into a table or partition that resides in S3.

Because of differences between S3 and traditional filesystems, DML operations for S3 tables can take longer than for tables on HDFS. For example, both the LOAD DATA statement and the final stage of the INSERT and CREA

TE TABLE AS SELECT statements involve moving files from one directory to another. (In the case of INSERT and CREATE TABLE AS SELECT, the files are moved from a temporary staging directory to the final destination directory.) Because S3 does not support a “rename” operation for existing objects, in these cases, Impala copies the data files from one location to another and then removes the original files.

Manually Loading Data into Impala Tables in S3:

You can use the Amazon-provided methods to bring data files into S3 for querying through Impala.

After you upload data files to a location already mapped to an Impala table or partition, or if you delete files in S3 from such a location outside of Impala, issue the REFRESH statement to make Impala aware of the new set of data files.

Running and Tuning Impala Queries for Data Stored in S3

Once a table or partition is designated as residing in S3, the SELECT statement transparently accesses the data files from the appropriate storage layer.

- Queries against S3 data support all the same file formats as HDFS data.
- Tables can be unpartitioned or partitioned. For partitioned tables, either manually construct paths in S3 based on partition key values, or use ALTER TABLE ... ADD PARTITION to set up the appropriate paths in S3.
- HDFS and HBase tables can be joined to S3 tables, or S3 tables can be joined with each other.
- Authorization to control access to databases, tables, or columns works the same whether the data is in HDFS or in S3.
- The Catalog Server (`catalogd`) daemon caches metadata for both HDFS and S3 tables.
- Queries against S3 tables are subject to the same kinds of admission control and resource management as HDFS tables.
- Metadata about S3 tables is stored in the same Metastore database as for HDFS tables.
- You can set up views referring to S3 tables.
- The COMPUTE STATS, SHOW TABLE STATS, and SHOW COLUMN STATS statements work for S3 tables.

Query Performance for S3

All else being equal, performance is expected to be lower for queries running against data in S3 than HDFS. The actual mechanics of the SELECT statement are different when the data is in S3. Although the work is still distributed across the DataNodes of the cluster, Impala might parallelize the work for a distributed query differently for data on HDFS and S3.

S3 does not have the same block notion as HDFS, so Impala uses heuristics to determine how to split up large S3 files for processing in parallel. Because all hosts can access any S3 data file with equal efficiency, the distribution of work might be different than for HDFS data, where the data blocks are physically read using short-circuit local reads by hosts that contain the appropriate block replicas. Although the I/O to read the S3 data might be spread evenly across the hosts of the cluster, the fact that all data is initially retrieved across the network means that the overall query performance is likely to be lower for S3 data than for HDFS data.

Use the PARQUET_OBJECT_STORE_SPLIT_SIZE query option to control the Parquet-specific split size. The default value is 256 MB.

When optimizing aspects of complex queries such as the join order, Impala treats tables on HDFS and S3 the same way. Therefore, follow all the same tuning recommendations for S3 tables as for HDFS ones, such as using the COMPUTE STATS statement to help Impala construct accurate estimates of row counts and cardinality.

In query profile reports, the numbers for BytesReadLocal, BytesReadShortCircuit, BytesReadDataNodeCached, and BytesReadRemoteUnexpected are blank because those metrics come from HDFS. By definition, all the I/O for S3 tables involves remote reads.

Restrictions on Impala Support for S3

The following restrictions apply when using Impala with S3:

- Impala does not support the old s3:// block-based and s3n:// filesystem schemes, and it only supports s3a://.
- Although S3 is often used to store JSON-formatted data, the current Impala support for S3 does not include directly querying JSON data. For Impala queries, use data files in one of the file formats listed in Hadoop File Formats Support on page 66. If you have data in JSON format, you can prepare a flattened version of that data for querying by Impala as part of your ETL cycle.
- You cannot use the ALTER TABLE ... SET CACHED statement for tables or partitions that are located in S3.

Specifying Impala Credentials to Access S3

Specify an Impala credential to access data in Amazon S3.

Procedure

1. In Cloudera Manager, navigate to Administration External Accounts .
2. In the AWS Credentials tab, click Add Access Key Credentials.
3. Enter a Name of your choosing for this account.
4. Enter the AWS Access Key ID.
5. Enter the AWS Secret Key.
6. Click Add.
7. Click Save to finish adding the AWS Credential.
8. Select Cluster Access to S3.

Impala with Ozone

You can use Impala to query data files that reside on Apache Ozone distributed storage, rather than in HDFS.

The typical use case for Impala and Ozone together is to use Ozone for the default filesystem, replacing HDFS entirely. In this configuration, when you create a database, table, or partition, the data always resides on Ozone storage and you do not need to specify any special LOCATION attribute. If you do specify a LOCATION attribute, its value refers to a path within the Ozone filesystem.

For example:

If the default filesystem is Ozone, all Impala data resides there and all Impala databases and tables are also located there.

```
CREATE TABLE t1 (x INT, s STRING);
```

You can specify LOCATION for database, table, or partition, using values from the Ozone filesystem.

```
CREATE DATABASE d1 LOCATION
'/some/path/on/ozone/server/d1.db'; CREATE TABLE d1.t2 (a TINYINT, b BOOLEAN
);
```

Impala can write to, delete, and rename data files and database, table, and partition directories on Ozone storage. Therefore, Impala statements such as CREATE TABLE, DROP TABLE, CREATE DATABASE, DROP DAT ABASE, ALTER TABLE, and INSERT work the same with Ozone storage as with HDFS.

Ozone supports multiple protocols: ofs, o3fs, and s3a. Impala supports reading ofs and o3fs. Impala can also read s3a (see Impala with Amazon S3 for more information). However, ofs is their newer protocol, and the only one Impala supports as a default filesystem. We recommend using it for DDL statements to avoid access limitations, and for DML statements and SELECT statements for performance.

Because Apache Ozone storage buckets use a global value for the block size rather than a configurable value for each file, the PARQUET_FILE_SIZE query option has no effect when Impala inserts data into a table or partition residing on Ozone storage.

Impala's spill-to-disk feature may be configured to use Ozone storage by specifying a full URI (e.g. ofs://host:port/volume/bucket/key) for the spill location.

See Spill to remote storage for details on configuring remote spill-to-disk.

Configure Impala Daemon to spill to Ozone

You can use Ozone as a scratch space for writing intermediate files during large sorts, joins, aggregations, or analytic function operations.

Before you begin

- Identify the Ozone scratch directory where you want your new Impala to write the temporary data.
- Identify the IP address, host name, or service identifier of Ozone.
- Identify the port number of the Ozone Manager (if not-default).

Configuring the Start-up Option in Impala daemon

You can use the Impalad start option scratch_dirs to specify the locations of the intermediate files.

```
--scratch_dirs="ofs://AUTHORITY/PATH(:MAX_BYTES), LOCAL_BUFFER_DIR
 (,LOCAL_DIR…)"
```

- Where ofs://*AUTHORITY*/*PATH* is the remote directory.
- authority may include ip_address or hostname and port, or service_id.
- max_bytes is optional.

Using the above format:

- You can specify only one remote directory. When you configure a remote directory, you must specify a local buffer directory as the buffer. However you can use multiple local directories with the remote directory. If you specify multiple local directories, the first local directory would be used as the local buffer directory.
- If you configure both remote and local directories, the remote directory is only used when the local directories are fully utilized.
- The size of a remote intermediate file could affect the query performance, and the value can be set by --remote _tmp_file_size=*SIZE* in the start-up option. The default size of a remote intermediate file is 16MB while the maximum is 512MB.

Examples

- An Ozone scratch dir with one local buffer dir, file size 64MB. The space of Ozone scratch dir is limited to 300G.

  ```
  --scratch_dirs=ofs://10.0.0.49:29000/tmp:300G,/local_buffer_dir --remote
  _tmp_file_size=64M
  ```

- An Ozone scratch dir with one local buffer dir limited to 512MB, and one local dir limited to 10GB. The space of Ozone scratch dir is limited to 300G. The Ozone Manager uses its default port (9862).

  ```
  --scratch_dirs=ofs://ozonemgr/tmp:300G,/local_buffer_dir:512M,/local_dir
  :10G
  ```

- An Ozone scratch dir with one local buffer dir, and multiple prioritized local dirs. The space of Ozone scratch dir is unlimited. The Ozone service identifier is ozone1.

  ```
  --scratch_dirs=ofs://ozone1/tmp,/local_buffer_dir,/local_dir_1:5G:1,/loc
  al_dir_2:5G:2
  ```

Even though max_bytes is optional, it is highly recommended to configure for spilling to Ozone because the Ozone cluster space is limited.

Ports Used by Impala

Impala uses the TCP ports listed in the following table.

Before deploying Impala, ensure these ports are open on each system. You can use the corresponding settings in Cloudera Manager to overwrite the default port numbers.

Scope	Cloudera Manager Setting / Startup Flag	Default Port	Comment
Impala Daemon	Impala Daemon Frontend Port --beeswax_port	21000	Port on which Beeswax client requests are served by Impala Daemons.
Impala Daemon	Impala Daemon Frontend Port --hs2_port	21050	Port on which HiveServer2 client requests are served by Impala Daemons.
Impala Daemon	Impala Daemon Backend Port --be_port	22000	Internal use only. Impala daemons use this port to communicate with each other.
Impala Daemon	StateStoreSubscriber Service Port --state_store_subscriber_port	23000	Internal use only. Impala daemons listen on this port for updates from the StateStore daemon.
Catalog Daemon	StateStoreSubscriber Service Port --state_store_subscriber_port	23020	Internal use only. The catalog daemon listens on this port for updates from the StateStore daemon.
Impala Daemon	Impala Daemon HTTP Server Port --webserver_port	25000	Impala debug Web UI for administrators to monitor and troubleshoot.
Impala StateStore Daemon	StateStore HTTP Server Port --webserver_port	25010	StateStore debug Web UI for administrators to monitor and troubleshoot.
Impala Catalog Daemon	Catalog HTTP Server Port --webserver_port	25020	Catalog Server debug Web UI for administrators to monitor and troubleshoot.
Impala StateStore Daemon	StateStore Service Port --state_store_port	24000	Internal use only. The StateStore daemon listens on this port for registration/unregistration requests.
Impala Catalog Daemon	Catalog Service Port --catalog_service_port	26000	Internal use only. The catalog service uses this port to communicate with the Impala daemons.
Impala Daemon	KRPC Port --krpc_port	27000	Internal use only. Impala daemons use this port for KRPC based communication with each other.
Impala Daemon	HiveServer2 HTTP Port --hs2_http_port	28000	Used to transmit commands and receive results by client applications over HTTP via the HiveServer2 protocol.

Related Information

Starting and Stopping Impala

Migration Guide

This topic describes the important differences between Impala in CDH and Impala in Cloudera. These changes were made in Cloudera for the optimal interoperability between Hive and Impala for the improved user experience. Review the changes before you migrate your Impala workload from CDH to Cloudera.

Default Value Changes in Configuration Options

Configuration Option	Scope	Default in CDH 6.x	Defaul
DEFAULT_FILE_FORMAT	Query	TEXT	PARQ
hms_event_polling_interval_s	Catalogd	0	2
ENABLE_ORC_SCANNER	Query	TRUE	FALSE
use_local_catalog	Coordinator / Catalogd	false	true
catalog_topic_mode	Coordinator	full	minima

New Configuration Options

Configuration Option	Scope	Default Value
default_transactional_type	Coordinator	insert_only
DEFAULT_TRANSACTIONAL_TYPE	Query	insert_only
disable_hdfs_num_rows_estimate	Impalad	false
disconnected_session_timeout	Coordinator	900
PARQUET_OBJECT_STORE_SPLIT_SIZE	Query	256 MB
SPOOL_QUERY_RESULTS	Query	FALSE
MAX_RESULT_SPOOLING_MEM	Query	100 MB
MAX_SPILLED_RESULT_SPOOLING_MEM	Query	1 GB
FETCH_ROWS_TIMEOUT_MS	Query	10 s
DISABLE_HBASE_NUM_ROWS_ESTIMATE	Query	FALSE

Default File Format

New default behavior:

In Cloudera, the default file format of the tables is Parquet.

When you issue the CREATE TABLE statement without the STORED AS clause, Impala creates a Parquet table instead of a Text table as in CDH.

For example, if you create an external table based on a text file without providing stored as clause and then tries to issue a select query, the query will fail, in Cloudera, since Impala expects the file to be in the Parquet file format.

Steps to switch to the CDH behavior:

- Add the explicitly STORED AS clause to in the CREATE TABLE statements if the file format is not Parquet.
- Start Coordinator with the default_transactional_type flag set to NONE for all tables to be created.
- Set the DEFAULT_FILE_FORMAT query option to TEXT to revert to the default of the Text format for one or more CREATE TABLE statements.

Default Insert-only Transactional Tables

New default behavior:

By default, managed tables are transactional tables with the insert_only property in Cloudera.

- You can no longer perform file system modifications (add/remove files) on a managed table in Cloudera. The directory structure for transactional tables is different than non-transactional tables, and any out-of-band files which are added may or may not be picked up by Hive and Impala.
- The insert_only transactional tables cannot be currently altered in Impala. The ALTER TABLE statement on a transactional table currently throws an error.

- Impala does not currently support compaction on transaction tables. You should use Hive to compact the tables as needed.
- The SELECT, INSERT, INSERT OVERWRITE, TRUNCATE statements are supported on the insert-only transactional tables.

Steps to switch to the CDH behavior:

- If you do not want transactional tables, set the DEFAULT_TRANSACTIONAL_TYPE query option to NONE so that any newly created managed tables are not transactional by default.
- External tables do not drop the data files when the table is dropped. If you wish to purge the data along with table when table is dropped, add external.table.purge = true in the table properties. When external.table.purge is set to true, the data is removed when the DROP TABLE statement is executed.

Managed and External Tablespace Directories

New default behavior:

In Cloudera, there are separate HDFS directories for managed and external tables.

- The data files for managed tables are located in warehouse location specified by the Cloudera Manager configuration setting, hive_warehouse_directory.
- The data files for external tables are located in warehouse location specified by the Cloudera Manager configuration setting, hive_warehouse_external_directory.

If you perform file system level operations for adding/removing files on the table, you need to consider if its an external table or managed table to find the location of the table directory.

Steps to switch to the CDH behavior:

Check the output of the DESCRIBE FORMATTED command to find the table location.

Hive Default File Format Interoperability

New default behavior:

The managed tables created by Hive are of ORC file format, by default, and support full transactional capabilities. If you create a table without specifying the STROED AS clause and load data from Hive, then such tables are not readable or writable by Impala.

Steps to switch to the CDH behavior:

- You must use the STORED AS PARQUET clause when you create tables in Hive if you want interoperability with Impala on those tables.
- If you want to change this default file format at the system level, in the Hive_on_Tez service configuration in Cloudera Manager, set the hive_default_fileformat_managed field to parquet.

Statistics Interoperability Between Hive and Impala

New default behavior:

Statistics for tables are engine specific, namely, Hive or Impala, so that each engine could use its own statistics and not overwrite the statistics generated by the other engine.

When you issue the COMPUTE STATS statement on Impala, you need to issue the corresponding statement on Hive to ensure both Hive and Impala statistics are accurate.

Impala COMPUTE STATS command does not overwrite the Hive stats for the same table.

Steps to switch to the CDH behavior:

There is no workaround.

ORC Table Support Disabled

New default behavior:

ORC table support is disabled by default. The CDH queries against ORC tables will fail in Cloudera.

Steps to switch to the CDH behavior:

Set the query option ENABLE_ORC_SCANNER to TRUE to re-enable ORC table support.

Note that this option does not work on a full transactional ORC table, and the queries will still return an error.

Reconnect to HS2 Session

New default behavior:

Clients can disconnect from Impala while keeping the session running. Clients can reconnect to the same session by presenting the session_token. By default, disconnected sessions are terminated after 15 min.

- Your end users will not notice a difference in behaviour.
- If clients are disconnected without the driver explicitly closing the session (e.g. because of a network fault), disconnected sessions and the queries associated with them may remain open and continue consuming resources until the disconnected session is timed out. Administrators may notice these disconnected sessions and/or the associated resource consumption.

Steps to switch to the CDH behavior:

You can adjust the --disconnected_session_timeout flag to a lower value so that disconnected sessions are cleaned up more quickly.

Automatic Row Count Estimation when Statistics Are Not Present

New default behavior:

If there are no statistics available on a table, Impala will try to estimate the cardinality by estimating the size of table based on the number of rows in the table. This behavior is turned on by default and should result in better plans for most cases when stats are not present.

For some edge cases, it is possible that Impala will generate a bad plan (when compared to the same query in CDH) when the statistics are not present on that table and could negatively affect the query performance.

Steps to switch to the CDH behavior:

Set the DISABLE_HDFS_NUM_ROWS_ESTIMATE query option to TRUE to disable this optimization.

Metadata V2

All catalog metadata improvements are enabled by default in CDP, resulting in many performance and scalability improvements, such as reduced memory footprint on coordinators and automatic cache eviction.

See Impala Metadata Management for the details about catalog improvements.

Amazon EC2 Instances

Impala Data Hub clusters in Cloudera are created using the Data Mart cluster type. The recommended EC2 instance type for Impala Data Mart Cloudera Data Hub clusters is r5d.2xlarge. This instance type provides a good balance of memory vs CPU and also includes fast ephemeral SSD storage which is used by Impala data cache and spill.

When provisioning a Data Mart Cloudera Data Hub cluster with the Cluster Definition option in the UI (or the --cl uster-definition-name option in the CLI), the r5d.2xlarge instance type will be selected by default.

If you choose the Custom option to provision a cluster with the Data Mart blueprint, you should explicitly specify the r5d.2xlarge instance type since this isn't the default type for Custom clusters.

Impala DWX clusters use the r5d.2xlarge instance type for executor instances and it's not possible to override this.

Authorization Support by Ranger

In Cloudera, Ranger is the authorization provider for Impala.

New behavior:

- The CREATE ROLE, GRANT ROLE, SHOW ROLE statements are not supported as Ranger currently does not support roles.
- When a particular resource is renamed, currently, the policy is not automatically transferred to the newly renamed resource.
- SHOW GRANT with an invalid user/group does not return an error.

The following table lists the different access type requirements to run SQL statements in Hive and Impala.

SQL Statement	Impala Access Requirement	Hive Access Requirement
DESCRIBE *VIEW*	VIEW_METADATA on the underlying tables	SELECT on the view
ALTER TABLE RENAME ALTER VIEW RENAME	ALL on the target table / view ALTER on the source table / view	ALTER on the source table / view
SHOW DATABASES SHOW TABLES	VIEW_METADATA	USE

Governance Support by Atlas

Both CDH and Cloudera environments support governance functionality for Impala operations. The two environments collect similar information to describe Impala activities, including:

- Audits for Impala access requests
- Metadata describing Impala queries
- Metadata describing any new data assets created or updated by Impala operations

The services that support these operations are different in the two environment. Functionality is distributed across services as follows:

Feature	CDH	CDP
Auditing		
• Access requests	Audit tab in Navigator console	Audit page in Ranger console
• Service operations that create or update metadata catalog entries	Audit tab in Navigator console	Audit tab for each entity in Atlas dashboard
• Service operations in general	Audit tab in Navigator console	No other audits collected.
Metadata Catalog		

• Impala operations:
  • CREATE TABLE AS SELECT
  • CREATE VIEW
  • ALTER VIEW AS SELECT
  • INSERT INTO
  • INSERT
  • OVERWRITE
	Process and Process Execution entities
	Column- and table-level lineage
		Process and Process Execution entities
		Column- and table-level lineage

Related Information

On-demand Metadata
Impala Authorization
Impala metadata collection
SQL transactions in Impala

Setting up Data Cache for Remote Reads

When Impala compute nodes and its storage are not co-located, the network bandwidth requirement goes up as the network traffic includes the data fetch as well as the shuffling exchange traffic of intermediate results. To mitigate the pressure on the network, you can enable the compute nodes to cache the working set read from remote filesystems, such as, remote HDFS data node, S3, ABFS, ADLS.

About this task

To enable remote data cache as follows.

Procedure

1. In Cloudera Manager, navigate to Clusters Impala Service .
2. In the Configuration tab, select Enable Local Data Cache to enable the local Impala Daemon data cache for caching of remote reads.
3. In Impala Daemon Data Cache Directories, add the directories Impala Daemon will use for caching of remote read data.
4. In Impala Daemon Data Cache Per Directory Capacity, specify the maximum amount of local disk space Impala will use per daemon in each of the configured directrories for caching of remote read data.
5. Click Save Changes and restart the Impala service.

Managing Metadata in Impala

This section describes various knobs you can use to control how Impala manages its metadata in order to improve performance and scalability.

On-demand Metadata

With the on-demand metadata feature, the Impala coordinators pull metadata as needed from catalogd and cache it locally. The cached metadata gets evicted automatically under memory pressure.

The granularity of on-demand metadata fetches is at the partition level between the coordinator and catalogd. Common use cases like add/drop partitions do not trigger unnecessary serialization/deserialization of large metadata.

The feature can be used in either of the following modes.

Metadata on-demand mode

In this mode, all coordinators use the metadata on-demand.

Set the following on catalogd:

```
--catalog_topic_mode=minimal
```

Set the following on all impalad coordinators:

```
--use_local_catalog=true
```

Mixed mode

In this mode, only some coordinators are enabled to use the metadata on-demand.

We recommend that you use the mixed mode only for testing local catalog’s impact on heap usage.

Set the following on catalogd:

```
--catalog_topic_mode=mixed
```

Set the following on impalad coordinators with metdadata on-demand:

```
--use_local_catalog=true
```

Flags related to use_local_catalog

When use_local_catalog is enabled or set to True on the impalad coordinators the following list of flags configure various parameters as described below. Cloudera recommends maintaining the default values for the flags; however, you can modify the MAX_STMT_METADATA_LOADER _THREADS flag to suit your environment.

- The flag local_catalog_cache_mb (defaults to -1) configures the size of the catalog cache within each coordinator. With the default set to -1, the cache is auto-configured to 60% of the configured Java heap size. Note that the Java heap size is distinct from and typically smaller than the overall Impala memory limit.
- The flag local_catalog_cache_expiration_s (defaults to 3600) configures the expiration time of the catalog cache within each impalad. Even if the configured cache capacity has not been reached, items are removed from the cache if they have not been accessed in the defined amount of time.
- The flag local_catalog_max_fetch_retries (defaults to 40) configures the maximum number of retries needed for queries to fetch a metadata object from the impalad coordinator's local catalog cache.
- The MAX_STMT_METADATA_LOADER_THREADS flag (defaults to 8) controls the maximum number of threads used to load table metadata during query compilation. In local catalog mode, you can parallelize this process to reduce waiting time due to sequential loading, This flag allows you to parallelize that process, improving performance for complex queries that involve many tables.. If a query requires metadata for only one table, or if you set this flag to 1, Impala loads the metadata sequentially. Impala also falls back to sequential loading if the system cannot successfully start the requested number of threads.

  The following example shows how to increase the parallelism for a query that joins many tables:

  ```
  SET MAX_STMT_METADATA_LOADER_THREADS=16;
  SELECT * FROM table1 JOIN table2 JOIN table3...
  ```

  The following example shows how to disable parallel metadata loading by setting the thread count to 1:

  ```
  SET MAX_STMT_METADATA_LOADER_THREADS=1;
  SELECT count(*) FROM functional.alltypes;
  ```

Limitation:

HDFS caching is not supported in On-demand metadata mode coordinators.



Note:

In Impala 3.4.0 and above, global INVALIDATE METADATA statement is supported when On-demand feature is enabled.

INVALIDATE METADATA Usage Notes:

To return accurate query results, Impala needs to keep the metadata current for the databases and tables queried. Through "automatic invalidation" or "HMS event polling" support, Impala automatically picks up most changes in metadata from the underlying systems. However there are some scenarios where you might need to run INVALIDA TE METADATA or REFRESH.

- if some other entity modifies information used by Impala in the metastore, the information cached by Impala must be updated via INVALIDATE METADATA or REFRESH,
- if you have "local catalog" enabled without "HMS event polling" and need to pick up metadata changes that were done outside of Impala in Hive and other Hive client, such as SparkSQL,
- and so on.
-  **Note:** As this is a very expensive operation compared to the incremental metadata update done by the REFRESH statement, when possible, prefer REFRESH rather than INVALIDATE METADATA.

Automatic Invalidation of Metadata Cache

To keep the size of metadata bounded, the Impala Catalog Server periodically scans all the tables and invalidates those not recently used.

There are two types of configurations in Catalog Server that control the automatic invalidation of metadata in the Catalog Server Command Line Argument Advanced Configuration Snippet (Safety Valve) field in Cloudera Manager.

Time-based cache invalidation

Catalogd invalidates tables that are not recently used in the specified time period (in seconds).

The ##invalidate_tables_timeout_s flag needs to be applied to both impalad and catalogd.

Memory-based cache invalidation

When the memory pressure reaches 60% of JVM heap size after a Java garbage collection in cata logd, Impala invalidates 10% of the least recently used tables.

The ##invalidate_tables_on_memory_pressure flag needs to be applied to both impalad and cata logd.

Automatic invalidation of metadata provides more stability with lower chances of running out of memory, but the feature could potentially cause performance issues and may require tuning.

Automatic Invalidation/Refresh of Metadata

In this release, you can invalidate or refresh metadata automatically after changes to databases, tables or partitions render metadata stale. You control the synching of tables or database metadata by basing the process on events. You learn how to access metrics and state information about the event processor.

When tools such as Hive and Spark are used to process the raw data ingested into Hive tables, new HMS metadata (database, tables, partitions) and filesystem metadata (new files in existing partitions/tables) are generated. In previous versions of Impala, in order to pick up this new information, Impala users needed to manually issue an INVALIDATE or REFRESH commands.

When automatic invalidate/refresh of metadata is enabled,, the Catalog Server polls Hive Metastore (HMS) notification events at a configurable interval and automatically applies the changes to Impala catalog.

Impala Catalog Server polls and processes the following changes.

- Refreshes the tables when it receives the ALTER TABLE event.
- Refreshes the partition when it receives the ALTER, ADD, or DROP partitions.
- Adds the tables or databases when it receives the CREATE TABLE or CREATE DATABASE events.
- Removes the tables from catalogd when it receives the DROP TABLE or DROP DATABASE events.
- Refreshes the table and partitions when it receives the INSERT events.

  If the table is not loaded at the time of processing the INSERT event, the event processor does not need to refresh the table and skips it.
- Changes the database and updates catalogd when it receives the ALTER DATABASE events. The following changes are supported. This event does not invalidate the tables in the database.

- Change the database properties
- Change the comment on the database
- Change the owner of the database
- Change the default location of the database

  Changing the default location of the database does not move the tables of that database to the new location. Only the new tables which are created subsequently use the default location of the database in case it is not provided in the create table statement.

This feature is controlled by the ##hms_event_polling_interval_s flag. Start the catalogd with the ##hms_event_polling_interval_s flag set to a positive integer to enable the feature and set the polling frequency in seconds. We recommend the value to be less than 5 seconds.

Limitations

The following use cases are not supported:

- When you bypass HMS and add or remove data into table by adding files directly on the filesystem, HMS does not generate the INSERT event, and the event processor will not invalidate the corresponding table or refresh the corresponding partition.

  It is recommended that you use the LOAD DATA command to do the data load in such cases, so that event processor can act on the events generated by the LOAD command.
- The Spark API that saves data to a specified location does not generate events in HMS, thus is not supported. For example:

```
Seq((1, 2)).toDF("i", "j").write.save("/user/hive/warehouse/spark_etl.db/
customers/date=01012019")
```

- Event processing could have delays due to the polling interval and auto-refresh on large tables also takes time. If you want the metadata to be synced up immediately, manual REFRESH/INVALIDATE is a better choice and has a better guarantee.

Disable Event Based Automatic Metadata Sync

When the ##hms_event_polling_interval_s flag is set to a non-zero value for your catalogd, the event-based automatic invalidation is enabled for all databases and tables. If you wish to have the fine-grained control on which tables or databases need to be synced using events, you can use the impala.disableHmsSync property to disable the event processing at the table or database level.

This feature can be turned off by setting the ##hms_event_polling_interval_s flag set to 0.

The disable_hms_sync_by_default option controls the default event synchronization behavior for the catalog service:

- If set to true, the catalog service skips HMS event processing for all tables and databases by default. You can opt-in specific tables or databases by setting their impala.disableHmsSync property to false.
- If set to false, the catalog service processes events as usual unless specifically disabled at the table or database level.



Note: HMS notification events polling remains independent and is not affected by this flag.

When you add the DBPROPERTIES or TBLPROPERTIES with the impala.disableHmsSync key, the HMS event based sync is turned on or off. The value of the impala.disableHmsSync property determines if the event processing needs to be disabled for a particular table or database.

- If 'impala.disableHmsSync'='true', the events for that table or database are ignored and not synced with HMS.
- If 'impala.disableHmsSync'='false' or if impala.disableHmsSync is not set, the automatic sync with HMS is enabled if the ##hms_event_polling_interval_s global flag is set to non-zero.

- To disable the event based HMS sync for a new database, set the impala.disableHmsSync database properties in Hive as currently, Impala does not support setting database properties:

```
CREATE DATABASE <name> WITH DBPROPERTIES ('impala.disableHmsSync'='true');
```

- To enable or disable the event based HMS sync for a table:

```
CREATE TABLE <name> ... TBLPROPERTIES ('impala.disableHmsSync'='true' |
'false');
```

- To change the event based HMS sync at the table level:

```
ALTER TABLE <name> SET TBLPROPERTIES ('impala.disableHmsSync'='true' | '
false');
```

When both table and database level properties are set, the table level property takes precedence. If the table level property is not set, then the database level property is used to evaluate if the event needs to be processed or not.

If the property is changed from true (meaning events are skipped) to false (meaning events are not skipped), you need to issue a manual INVALIDATE METADATA command to reset event processor because it doesn't know how many events have been skipped in the past and cannot know if the object in the event is the latest. In such a case, the status of the event processor changes to NEEDS_INVALIDATE.

Metrics for Event Based Automatic Metadata Sync

You can use the web UI of the catalogd to check the state of the automatic invalidate event processor.

By default, the debug web UI of catalogd is at http://*IMPALA-SERVER-HOSTNAME*:25020 (non-secure cluster) or https://*IMPALA-SERVER-HOSTNAME*:25020 (secure cluster).

Under the web UI, there are two pages that presents the metrics for HMS event processor that is responsible for the event based automatic metadata sync.

- /metrics#events
- /events

  This provides a detailed view of the metrics of the event processor, including min, max, mean, median, of the durations and rate metrics for all the counters listed on the /metrics#events page.

The /metrics#events page provides the following metrics about the HMS event processor.

Name	Description
events-processor.avg-events-fetch-duration	Average duration to fetch a batch of events and process it.
events-processor.avg-events-process-duration	Average time taken to process a batch of events received from the Metastore.
events-processor.events-received	Total number of the Metastore events received.
events-processor.events-received-15min-rate	Exponentially weighted moving average (EWMA) of number of events received in last 15 min. This rate of events can be used to determine if there are spikes in event processor activity during certain hours of the day.
events-processor.events-received-1min-rate	Exponentially weighted moving average (EWMA) of number of events received in last 1 min. This rate of events can be used to determine if there are spikes in event processor activity during certain hours of the day.
events-processor.events-received-5min-rate	Exponentially weighted moving average (EWMA) of number of events received in last 5 min. This rate of events can be used to determine if there are spikes in event processor activity during certain hours of the day.

Name	Description
events-processor.events-skipped	Total number of the Metastore events skipped. Events can be skipped based on certain flags are table and database level. You can use this metric to make decisions, such as: • If most of the events are being skipped, see if you might just turn off the event processing. • If most of the events are not skipped, see if you need to add flags on certain databases.
events-processor.status	Metastore event processor status to see if there are events being received or not. Possible states are: • PAUSED The event processor is paused because catalog is being reset concurrently. • ACTIVE The event processor is scheduled at a given frequency. • ERROR • The event processor is in error state and event processing has stopped. • NEEDS_INVALIDATE The event processor could not resolve certain events and needs a manual INVALIDATE command to reset the state. • STOPPED The event processing has been shutdown. No events will be processed. • DISABLED The event processor is not configured to run.

SQL transactions in Impala

A transaction is a single logical operation on the data. Impala supports transactions that satisfy a level of consistency that improves the integrity and reliability of the data before and after a transaction.

Specifically, Impala provides atomicity and isolation of insert operations on transactional tables. A single table insert is either committed in full or not committed, and the results of the insert operation are not visible to other query operations until the operation is committed.

For single table, the inserts are ordered, so if Impala doesn't see a committed insert, it won't see any insert committed after it.

For insert-only transactional tables, you can perform the following statements: CREATE TABLE, DROP TABLE, TRUNCATE, INSERT, SELECT

All transactions in Impala automatically commit at the end of the statement. Currently, Impala does not support multi-statement transactions.

Insert-only tables must be the managed and file-format based tables, such as Parquet, Avro, and text.

Note:

- Impala does not support changing transactional properties of tables. For example, you cannot alter a transactional table to a non-transactional table.
- Currently, ALTER TABLE statement is not supported on both insert-only and full acid transactional tables.