Release Notes

Date published:
Date modified:

https://docs.cloudera.com/

Legal Notice

Contents

What's New In Cloudera Runtime 7.0.2

What's New in Apache Atlas

This topic lists new features for Apache Atlas in this release of Cloudera Runtime.

Label and user-defined property support

You can now add the following metadata to entities in Atlas:

- Labels: text strings up to 50 characters
- User-defined properties: key-value pairs

Any user with privileges to update an entity can add this new metadata. The changes are stored as part of the metadata for the entity and do not require a change to the entity type definition.

What's new in DAS

There are no new features for DAS in this release of Cloudera Runtime.

What's New in Apache HBase

This topic lists new features for Apache HBase in this release of Cloudera Runtime.

For more information about Apache HBase, see Apache HBase Overview.

What's New in HDFS

There are no new features for HDFS in this release of Cloudera Runtime.

For more information about HDFS, see HDFS Overview

What's New in Apache Hive

There are no new features for Apache Hive in this release of Cloudera Runtime.

What's New in Hue

This topic lists new features for Hue in this release of Cloudera Runtime.

Support for Hive on Tez

Hue in Cloudera Runtime now supports Hive using Tez as its execution engine.

Integration with Apache Atlas data catalog

In Cloudera Runtime version 7.0.1, which was released with CDP 1.1 on September 23, 2019, integration with Apache Atlas data catalog was announced. See the September 23, 2019 release notes for Hue in Cloudera Runtime version 7.0.1 for details.

What's New in Apache Impala

This topic lists new features for Apache Impala in this release of Cloudera Runtime.

Server-side Spooling of Query Results

You can use the SPOOL_QUERY_RESULTS query option to control how query results are returned to the client.

By default, when a client fetches a set of query results, the next set of results are fetched in batches until all the result rows are produced. If a client issues a query without fetching all the results, the query fragments continue to hold on to the resources until the query is canceled and unregistered, potentially tying up resources and cause other queries to wait in admission control.

When the query result spooling feature is enabled, the result sets of queries are eagerly fetched and buffered until they are read by the client, and resources are freed and available for other queries.

See Spooling Impala Query Results for the new feature and the query options.

New Built-in Functions for Fuzzy Matching of Strings

Use the new Jaro or Jaro-Winkler functions to perform fuzzy matches on relatively short strings, e.g. to scrub user inputs of names against the records in the database.

- JARO_DISTANCE, JARO_DST
- JARO_SIMILARITY, JARO_SIM
- JARO_WINKLER_DISTANCE, JW_DST
- JARO_WINKLER_SIMILARITY, JW_SIM

See Impala String Functions for details.

Query Profile Exported to JSON

On the Query Details page of Impala Daemon Web UI, you have a new option, in addition to the existing Thrift and Text formats, to export the query profile output in the JSON format.

See Impala Daemon Web UI for generating JSON query profile outputs in Web UI.

DATE Data Type Support for Avro

You can now use the DATE data type to query date values from Avro tables.

See DATE Data Type and Using the Avro File Format with Impala Tables for details.

Cookie-based Authentication

Starting in this version, Impala supports cookies for authentication when clients connect via HiveServer2 over HTTP.

You can use the --max_cookie_lifetime_s startup flag to:

- Disable the use of cookies
- Control how long generated cookies are valid for

See Impala Clients for more information.

Capacity Quota for Scratch Disks

When configuring scratch space for intermediate files used in large sorts, joins, aggregations, or analytic function operations, use the ##scratch_dirs startup flag to optionally specify a capacity quota per scratch directory, e.g., ##scratch_dirs=/dir1:5MB,/dir2.

See Impala and HDFS for details.

TRUNCATE for Insert-only Transactional Tables

Now you can truncate insert-only transactional tables in Impala with the TRUNCATE statement.

See Impala Transactions for more information on transactional tables.

Query Option for Disabling HBase Row Estimation

During query plan generation, Impala samples underlying HBase tables to estimate row count and row size, but the sampling process can negatively impact the planning time. To alleviate the issue, when the HBase table stats do not change much in a short time, disable the sampling with the DISABLE_HBASE_NUM_ROWS_ESTIMATE query option so that the Impala planner falls back to using Hive Metastore (HMS) table stats instead.

See Impala Query Options.

Query Option for Controlling Size of Parquet Splits on Non-block Stores

To optimize query performance, Impala planner uses the value of the fs.s3a.block.size startup flag when calculating the split size on non-block based stores, e.g. S3, ADLS, etc. Starting in this release, Impala planner uses the PARQ UET_OBJECT_STORE_SPLIT_SIZE query option to get the Parquet file format specific split size.

For Parquet files, the fs.s3a.block.size startup flag is no longer used.

The default value of the PARQUET_OBJECT_STORE_SPLIT_SIZE query option is 256 MB.

See Impala with Amazon S3 for tuning Impala query performance for S3.

Support of Kerberos Authentication in Impala Web UI

Starting in this release, you can configure Kerberos authentication in Cloudera Manager to secure the debug Web UI pages for Impala Daemon, Catalog Server, and StateStore.

See Configuring Impala Web UI for the steps to enable Kerberos authentication for the Impala Web UI pages.

What's New in Apache Kafka

This topic lists new features for Apache Kafka in this release of Cloudera Runtime.

Rebase on Apache Kafka 2.3.0

Kafka shipped with this version of Cloudera Runtime is based on Apache Kafka 2.3.0. For more information, see Apache Kafka Notable Changes and Apache Kafka Release Notes in the upstream documentation.

Provision Kafka clusters with Data Hub

The Streams Messaging Heavy Duty and Streams Messaging Light Duty cluster templates and definitions are now available in Data Hub with advanced messaging and real-time processing on streaming data using Apache Kafka, centralized schema management using Schema Registry, and management and monitoring capabilities powered by Streams Messaging Manager. For more information, see the Data Hub documentation.

Note:

The Streams Messaging cluster definitions and templates are in technical preview and are only supported with Cloudera Runtime version 7.0.2 or later. Cloudera encourages you to explore these technical preview features in non-production environments and provide feedback on your experiences through the Cloudera Support Portal.

Connect Kafka clients to Data Hub provisioned clusters

Connecting clients to Data Hub clusters provisioned with the Streams Messaging cluster definitions is possible. For step-by-step instructions, see Connecting Kafka clients to Data Hub provisioned clusters.

Access to Kafka Metadata in Zookeeper is restricted by default

The Enable Zookeeper ACL (zookeeper.set.acl) property is now directly configurable in Cloudera Manager and is enabled by default. As a result of this change, access to Kafka metadata stored in Zookeeper is restricted by default. The data is still world readable, however, administrative operations, for example topic creation, deletion, any configuration changes and so on, can only be performed by authorized users. For more information, see Restrict access to Kafka metadata in Zookeeper and Unlock Kafka metadata in Zookeeper.

Ranger authorization support

Ranger support for Kafka is added. You can now use Ranger to provide authorization for Kafka. For more information, see Using Ranger to Provide Authorization in CDP as well as the documentation on Kafka Authorization with Ranger.

The resource-based Ranger service used by Kafka is user configurable

The resource-based Ranger service used by Kafka for authorization can now be manually configured with the Ranger service name for this Kafka cluster property in Cloudera Manager. In addition, if a resource-based service is set in Kafka that does not yet exist in Ranger, it will be automatically created after the Kafka service is restarted. The name of the newly created service is based on the value of the Ranger service name for this Kafka cluster property. For more information, see Configure the resource-based Ranger service used for authorization.

PAM authentication support

You can now configure Kafka to authenticate clients using PAM. For more information, see PAM Authentication.

LDAP authentication support

You can now configure Kafka to authenticate clients using LDAP. For more information, see LDAP Authentication.

New metric for monitoring garbage collector runs

A new metric called kafka_jvm_gc_runs is added to the Kafka service. This metric enables users to monitor the number of garbage collector runs performed on each broker.

What's New in Apache Knox

There are no new features for Apache Knox in this release of Cloudera Runtime.

What's New in Apache Kudu

This topic lists new features for Apache Kudu in this release of Cloudera Runtime.

Support for putting tablet servers in to maintenance mode

Kudu supports putting tablet servers into maintenance. In this mode, the tablet server's replicas are not re-replicated if they fail. Re-replication for the remaining, under-replicated tablets is triggered only when you exit from the maintenance mode. The kudu tserver state enter_maintenance and the kudu tserver state exit_maintenance tools have been added to orchestrate the tablet server maintenance. The kudu tserver list tool has been amended with a "state" column option to display the current state of each tablet server.

Built-in NTP client maintains internal time

Kudu has a built-in NTP client which maintains an internal time that is used to generate the HybridTime timestamps. If you enable the NTP client, the system clock synchronization for the nodes running Kudu is no longer necessary. This is useful for containerized deployments and in other cases when it is troublesome to maintain a properly configured system NTP service at each node of a Kudu cluster. The list of NTP servers to synchronize against is

specified with the --builtin_ntp_servers flag. By default, the Kudu masters and the tablet servers use public servers hosted by the NTP Pool project. To use the built-in NTP client, set the --time_source=builtin flag and reconfigure the --builtin_ntp_servers flag if necessary.

Support for aggregated table statistics for the Kudu clients

Aggregated table statistics are now available to the Kudu clients through the KuduClient.getTableStatistics() and the KuduTable.getTableStatistics() methods in the Kudu Java client, and by using the KuduClient.GetTableStatistics() method in the Kudu C++ client. This allows for various query optimizations.

For example, Spark now uses the aggregrated table statistics to perform join optimizations. The statistics are available via the API of both the C++ and the Java Kudu clients. In addition, per-table statistics are available via the kudu tab le statistics CLI tool. The statistics are also available via the master's Web UI at the master:8051/metrics and the mast er:8051/table?id=<uuid> URIs.

New operations using the Kudu CLI

The Kudu CLI supports the following new operations:

- Altering table columns: The following, newly introduced sub-commands allow you to alter a column of the specified table:
  - kudu table column_set_default
  - kudu table column_remove_default
  - kudu table column_set_compression
  - kudu table column_set_encoding
  - kudu table column_set_block_size
- Dropping table columns: The kudu table delete_column sub-command allows you to drop a column of the specified table.
- Getting and setting extra configuration properties: The kudu table get_extra_configs and the kudu table set_extra _config sub-commands allow you to get and set the extra extra configuration properties for a table respectively.
- Creating and dropping range partitions: The kudu table add_range_partition and the kudu table drop_range_partit ion sub-commands allow you to create and drop the range partitions for a table respectively.

Optimizations and improvements

- Tablet servers now expand a tablet's data directory group with available healthy directories when all directories of the group are full.
- For scan operations that are run with the CLOSEST_REPLICA selection mode, the Kudu Java client now picks a random available replica in case no replica is located at the same node with the client that initiated the scan operation. This helps to spread the load generated by multiple scan requests to the same tablet among all available replicas. In older versions of Kudu, all such scan requests would end up fetching data from the same tablet replica.
- The tablet servers now consider the available disk space when choosing a set of data directories for a tablet's data directory group, and when deciding in which data directory a new block should be written.
- The tablet servers reject any individual write operations that violate the schema constraints in a batch of write operations that are received from a client. The previous behavior was to reject the whole batch of write operations if a violation of the schema constraints is detected even for a single row.
- Kudu RPC now enables TCP keepalive for all outbound connections for faster detection of the no-longer-reachable nodes.
- The memory reserved by tcmalloc is now released to the OS periodically to avoid any potential OOM issues in case of read-only workloads.
- The evaluation of predicates on columns of primitive types and NULL or NOT NULL predicates has been optimized to leverage SIMD instructions.

What's New in Apache Oozie

There are no new features for Apache Oozie in this release of Cloudera Runtime.

For more information about Oozie, see Overview of Oozie.

What's New in Apache Ranger

There are no new features for Apache Ranger in this release of Cloudera Runtime.

What's New in Apache Spark

There are no new features for Apache Spark in this release of Cloudera Runtime.

For more information about Spark, see Apache Spark Overview

What's New in Apache Hadoop YARN

There are no new features for Apache Hadoop YARN in this release of Cloudera Runtime.

More reading

For more information about Apache Hadoop YARN, see Apache Hadoop YARN Overview

What's New in Apache ZooKeeper

This topic lists new features for Apache ZooKeeper in this release of Cloudera Runtime.

Enable 4LW telnet commands in ZooKeeper configuration using CM

You can configure the 4lw.commands.whitelist property using Cloudera Manager. By default the wchc and wchp commands are not enabled because of their known DOS vulnerability. For more information, see Configure four-letter-words commands in ZooKeeper using Cloudera Manager.

New diagnostics REST interface called AdminServer

AdminServer is a new diagnostics REST interface that provides simple diagnostics and metrics through HTTP. The embedded admin server is optional and disabled by default. You can enable it and change the server port using Cloudera Manager. For more information, see Enable the AdminServer.

Fixed Issues In Cloudera Runtime 7.0.2

Fixed Issues in Kudu

This section lists the issues that have been fixed since the previous version.

KUDU-2980: Fault tolerant and diff scans fail if projection contains mis-ordered primary key columns

KUDU-2980: The fault-tolerant scan operation no longer fails for a projection when the key columns are specified in an order other than that of the table schema.

KUDU-2947: A replica with slow WAL may grant votes even if established leader is alive and well

KUDU-2947: A new leader replica is no longer re-elected when the persisting Raft transactions to the WAL takes longer than the leader election timeout.

KUDU-2635: Tserver crash because some orphaned blocks are still listed when deleting metadata

KUDU-2635: The tablet servers no longer crash when the blocks are not removed due to an IO error, especially after recovering from a disk failure.

KUDU-2622: Validate read and write default value sizes when deserializing ColumnSchemaPB

KUDU-2622: Kudu validates the size of the default read and write values when de-serializing Colu mnSchemaPB, thereby preventing the master and the tablet servers from crashing.

KUDU-2871: TLS 1.3 not supported by krpc

KUDU-2871: RPC negotiation does not fail on the Linux distributions shipped or updated with OpenSSL version 1.0.2 and newer when TLS v1.3 is supported at both the client and the server side. This is a temporary workaround before the connection negotiation code is properly updated to support 1.5-RTT handshake used in TLS v1.3.

KUDU-2842: Data race in CatalogManager::GetTableLocations

KUDU-2842: The data race between GetTabletLocations() and the tablet report processing threads has been fixed, thereby preventing the Kudu master from crashing.

KUDU-2625: Unexpected behavior of WriteBatch wrt rows violating schema constraints

KUDU-2625: Tablet servers now reject any individual write operations that violate the schema constraints in a batch of write operations. In prior versions of Kudu, the behavior was designed to reject the whole batch of write operations if a violation of the schema constraints is detected even for a single row. We recommend that you revise the applications that relied on the behavior mentioned above upon upgrading to Kudu 1.11.0.

Apache bug ID unavailable

A tablet server no longer crashes when you remove a tablet replica with a pending AlterSchema transaction.

Fixed Issues in Apache Spark

This section lists the issues that have been fixed since the previous version.

CDPD-3783: Cannot create databases from Spark

Attempting to create a database using Spark results in an error similar to the following:

```
org.apache.spark.sql.AnalysisException:
            org.apache.hadoop.hive.ql.metadata.HiveException: Me
taException(message:Permission denied: user [sparkuser] does not
 have [ALL] privilege on [hdfs://ip-10-1-2-3.cloudera.site:8020/
tmp/spark/warehouse/spark_database.db]);
```

Fixed Issues in Apache Zeppelin

This section lists the issues that have been fixed since the previous version.

CDPD-3090: Due to a configuration typo, functionality involving notebook repositories does not work

Due to a missing closing brace, access to the notebook repositories API is blocked by default.

CDPD-3047: Markdown interpreter does not handle certain numbered list syntax correctly

Using the plus sign (+) or asterisk (*) to continue a numbered list using the %md interpreter results in bullet point entries instead.

Fixed Issues in Apache ZooKeeper

This section lists the issues that have been fixed since the previous version.

Technical Service Bulletins

TSB 2022-577 ZooKeeper servers assign similar negative sessionIds to multiple sessions

For the latest update on this issue see the corresponding Knowledge article: TSB 2022-577: ZooKeeper servers assign similar negative sessionIds to multiple sessions

Known Issues In Cloudera Runtime 7.0.2

Known Issues in Apache Atlas

This topic describes known issues and workarounds for using Atlas in this release of Cloudera Runtime.

Search Returns Unexpected Results for New Entity Definitions with BigInteger or BigDecimal Attributes

When new entity types are defined with BigInteger or BigDecimal type attributes, searches against values in any attributes in entities of these types return unexpected results.

Workaround: None.

Cloudera JIRA: CDPD-5309

Atlas custom properties ignored in client services

When adding a custom property for the atlas-application.properties in Atlas hook-based services such as Hive, HBase, and Impala, the custom property is not reflected in the actual configuration file that Cloudera Manager generates, causing these properties to be ignored.

Workaround: None.

Cloudera JIRA: OPSAPS-51224

Spark metadata order may affect lineage

Atlas may record unexpected lineage relationships when metadata collection from the Spark Atlas Connector occurs out of sequence from metadata collection from HMS. For example, if an ALTER TABLE operation in Spark changing a table name and is reported to Atlas before HMS has processed the change, Atlas may not show the correct lineage relationships to the altered table.

Workaround: None.

Cloudera JIRA: CDPD-4762

Searches for Qualified Names with "@" doesn't fetch the correct results

When searching Atlas qualifiedName values that include an "at" character (@), Atlas does not return the expected results or generate appropriate search suggestions.

Workaround: Consider leaving out the portion of the search string that includes the @ sign, using the wildcard character * instead.

Cloudera JIRA: CDPD-4545

Missing Impala and Spark lineage between tables and their data files

Atlas does not create lineage between Hive tables and their backing HDFS files for CTAS processes run in Impala or Spark.

Workaround: None.

Cloudera JIRA: CDP-5027, CDPD-3700, IMPALA-9070

Table alias values are not found in search

When table names are changed, Atlas keeps the old name of the table in a list of aliases. These values are not included in the search index in this release, so after a table name is changed, searching on the old table name will not return the entity for the table.

Workaround: None.

Cloudera JIRA: CDPD-3208

Hive lineage missing for INSERT OVERWRITE queries

Lineage is not generated for Hive INSERT OVERWRITE queries on partitioned tables. Lineage is generated as expected for CTAS queries from partitioned tables.

Workaround: None.

Cloudera JIRA: CDPD-3160

Logging out of Atlas does not manage the external authentication

At this time, Atlas does not communicate a log-out event with the external authentication management, Apache Knox. When you log out of Atlas, you can still open the instance of Atlas from the same web browser without re-authentication.

Workaround: To prevent access to Atlas after logging out, close all browser windows and exit the browser.

Cloudera JIRA: CDPD-3125

Ranking of top results in free-text search not intuitive

The Free-text search feature ranks results based on which attributes match the search criteria. The attribute ranking is evolving and therefore the choice of top results may not be intuitive in this release.

Workaround: If you don't find what you need in the top 5 results, use the full results or refine the search.

Cloudera JIRA: CDPD-1892

Free text search in Atlas is case sensitive

The free text search bar in the top of the screen allows you to search across entity types and through all text attributes for all entities. The search shows the top 5 results that match the search terms at any place in the text (*term* logic). It also shows suggestions that match the search terms that begin with the term (term* logic). However, in this release, the search results are case-sensitive.

Workaround: If you don't see the results you expect, repeat the search changing the case of the search terms.

Workaround: None.

Cloudera JIRA: CDPD-1884

Queries with ? wildcard return unexpected results

DSL queries in Advanced Search return incorrect results when the query text includes a question mark (?) wildcard character. This problem occurs in environments where trusted proxy for Knox is enabled, which is always the case for CDP.

Workaround: None.

Cloudera JIRA: CDPD-1823

Extra WARN messages from Solr in the Atlas log

Communication between JanuGraph and Solr for some entity attributes included in the search index produces a large number of WARN messages in the Atlas logs. Other than increasing the volume of messages in the logs, no functionality is affected.

Workaround: These messages can be ignored.

Cloudera JIRA: CDPD-1679

Guest users are redirected incorrectly

Authenticated users logging in to Atlas are redirected to the CDP Knox-based login page. However, if a guest user (without Atlas privileges) attempts to log in to Atlas, the user is redirected instead to the Atlas login page.

Workaround: To avoid this problem, open the Atlas Dashboard in a private or incognito browser window.

Cloudera JIRA: CDPD-1664, CDP-1823

Known Issues in DAS

This topic describes known issues and workarounds for using DAS in this release of Cloudera Runtime.

- You may not see any data for a report for any new queries that you run. This can happen, especially for the last one day's report.

  Workaround:

  1. Shut down the DAS Event Processor.
  2. Run the following command from the Postgres server:

     ```
     update das.report_scheduler_run_audit set status = 'FAILED' where status
      = 'READING';
     ```

  3. Start the DAS Event Processor.
- On clusters secured with Knox proxy only: You might not be able to save the changes to the JDBC URL in the DAS UI to change the server interface (HS2 or LLAP) on which you are running your queries.
- You may be unable to upload tables or get an error while browsing files to upload tables in DAS on a cluster secured using Knox proxy.
- DAS does not parse semicolons (;) and double hyphens (--) in strings and comments.

  For example, if you have a semicolon in query such as the following, the query might fail: select * from properties where prop_value = "name1;name2";

  If a semicolon is present in a comment, then execute the query after removing the semicolon from the comment, or removing the comment altogether. For example:

  ```
  select * from test; -- select * from test;
  select * from test; /* comment; comment */
  ```

  Queries with double hyphens (--) might also fail. For example:

  ```
  select * from test where option = '--name';
  ```
- You might face UI issues on Google Chrome while using faceted search. We recommend you to use the latest version of Google Chrome (version 71.x or higher).
- You are unable to see databases or the query editor on the **Compose** page even after following the troubleshooting procedure. This can be caused due to incorrect hive service name in ZooKeeper.
- Visual Explain for the same query shows different graphs on the **Compose** page and the **Query Details** page.
- While running some queries, if you restart HSI, the query execution is stopped. However, DAS does not reflect this change and the queries appear to be in the same state forever.
- After a fresh installation, when there is no data and you try to access the Reports tab, it throws HTTP 404 Not Found Error.
- Join count does not get updated for tables with partitioned columns.

Known Issues in Apache Hadoop

This topic describes known issues and workarounds for using Hive in this release of Cloudera Runtime.

Technical Service Bulletins

TSB 2021-434: KMS Load Balancing Provider Fails to invalidate Cache on Key Delete

The KMS Load balancing Provider has not been correctly invalidating the cache on key delete operations. The failure to invalidate the cache on key delete operations can result in the possibility that data can be leaked from the framework for a short period of time based on the value of the hadoop.kms.current.key.cache.timeout.ms property. Its default value is 30,000ms. When the KMS is deployed in an HA pattern the KMSLoadBalancingProvider class will only send the delete operation to one KMS role instance in a round-robin fashion. The code lacks a call to invalidate the cache across all instances and can leave key information including the metadata and key stored (the deleted key) in the cache on one or more KMS instances up to the key cache timeout.

Upstream JIRA

- HADOOP-17208
- HADOOP-17304

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2020-434: KMS Load Balancing Provider Fails to invalidate Cache on Key Delete

Known Issues in Apache HBase

This topic describes known issues and workarounds for using HBase in this release of Cloudera Runtime.

IntegrationTestReplication fails if replication does not finish before the verify phase begins

During IntegrationTestReplication, if the verify phase starts before the replication phase finishes, the test will fail because the target cluster does not contain all of the data. If the HBase services in the target cluster does not have enough memory, long garbage-collection pauses might occur.

Workaround: Use the -t flag to set the timeout value before starting verification.

HDFS encryption with HBase

Cloudera has tested the performance impact of using HDFS encryption with HBase. The overall overhead of HDFS encryption on HBase performance is in the range of 3 to 4% for both read and update workloads. Scan performance has not been thoroughly tested.

Workaround: N/A

AccessController postOperation problems in asynchronous operations

When security and Access Control are enabled, the following problems occur:

- If a Delete Table fails for a reason other than missing permissions, the access rights are removed but the table may still exist and may be used again.
- If hbaseAdmin.modifyTable() is used to delete column families, the rights are not removed from the Access Control List (ACL) table. The portOperation is implemented only for postDeleteColumn().
- If Create Table fails, full rights for that table persist for the user who attempted to create it. If another user later succeeds in creating the table, the user who made the failed attempt still has the full rights.

Workaround: N/A

Apache Issue: HBASE-6992

Bulk load is not supported when the source is the local HDFS

The bulk load feature (the completebulkload command) is not supported when the source is the local HDFS and the target is an object store, such as S3/ABFS.

Workaround: Use distcp to move the HFiles from HDFS to S3 and then run bulk load from S3 to S3.

Apache Issue: N/A

Storing Medium Objects (MOBs) in HBase is currently not supported

Storing MOBs in HBase relies on bulk loading files, and this is not currently supported when HBase is configured to use cloud storage (S3).

Workaround: N/A

Apache Issue: N/A

Technical Service Bulletins

TSB 2021-494: Accumulated WAL Files Cannot be Cleaned up When Using Phoenix Secondary Global Indexes

The Write-ahead-log (WAL) files for Phoenix tables that have secondary global indexes defined on them, cannot be automatically cleaned up by HBase, leading to excess storage usage and possible error due to filling up the storage. Accumulated WAL files can lead to lengthy restart times as they must all be played back to ensure no dataloss occurs on restart. This can have follow-on HDFS impact if the number of WAL files overwhelm HDFS Name Node.

Upstream JIRA

- HBASE-20781
- HBASE-25459
- PHOENIX-5250

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-494: Accumulated WAL Files Cannot be Cleaned up When Using Phoenix Secondary Global Indexes

TSB 2021-453: Snapshot and cloned table corruption when original table is deleted

HBASE-25206 can cause data loss either through corrupting an existing hbase snapshot or destroying data that backs a clone of a previous snapshot.

Upstream JIRA

HBASE-25206

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-453: HBASE-25206 "snapshot and cloned table corruption when original table is deleted".

TSB 2021-463: Snapshot and cloned table corruption when original table is deleted

The HDFS short-circuit setting dfs.client.read.shortcircuit is overwritten to disabled by hbase-default.xml. HDFS short-circuit reads bypass access to data in HDFS by using a domain socket (file) instead of a network socket. This alleviates the overhead of TCP to read data from HDFS which can have a meaningful improvement on HBase performance (as high as 30-40%).

Users can restore short-circuit reads by explicitly setting dfs.client.read.shortcircuit in HBase configuration via the configuration management tool for their product (e.g. Cloudera Manager or Ambari).

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-463: HBase Performance Issue .

TSB 2021-506: Active HBase MOB files can be removed

Actively used MOB files can be deleted by MobFileCleanerChore due to incorrect serialization of reference file names. This is causing data loss on MOB-enabled tables.

Upstream JIRA

- HBASE-23723
- HBASE-25970

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-506: Active HBase MOB files can be removed

TSB 2022-569: HBase normalizer can cause table inconsistencies by merging non-adjacent regions

The normalizer in HBase is a background job responsible for splitting or merging HBase regions to optimize the number of regions and the distribution of the size of the regions in HBase tables. Due to the bug described in HBASE-24376, the normalizer can cause region inconsistencies (region overlaps/holes) by merging non-adjacent regions.

Upstream JIRA

HBASE-24376

Knowledge article

For the latest update on this issue, see the corresponding Knowledge article: TSB 2022-569: HBase normalizer can cause table inconsistencies by merging non-adjacent regions

Known Issues in HDFS

This topic describes known issues and unsupported features for using HDFS in this release of Cloudera Runtime.

CDPD-2946: Slow reading and writing of erasure-coded files

The ISA-L library is not packaged with HDFS as a result of which HDFS erasure coding falls back to the Java implementation which is much slower than the native Hadoop implementation. This slows down the reading and writing of erasure-coded files.

Unsupported Features

The following HDFS features are currently not supported in Cloudera Data Platform:

- ACLs for the NFS gateway (HADOOP-11004)
- Aliyun Cloud Connector (HADOOP-12756)
- Allow HDFS block replicas to be provided by an external storage system (HDFS-9806)
- Consistent standby Serving reads (HDFS-12943)
- Cost-Based RPC FairCallQueue (HDFS-14403)
- HDFS Router Based Federation (HDFS-10467)
- More than two NameNodes (HDFS-6440)
- NameNode Federation (HDFS-1052)
- NameNode Port-based Selective Encryption (HDFS-13541)
- Non-Volatile Storage Class Memory (SCM) in HDFS Cache Directives (HDFS-13762)
- OpenStack Swift (HADOOP-8545)
- SFTP FileSystem (HADOOP-5732)
- Storage policy satisfier (HDFS-10285)
- Transparent Data Encryption (HDFS-6134)
- Upgrade Domain (HDFS-7541)

Technical Service Bulletins

TSB 2021-406: CVE-2020-9492 Hadoop filesystem bindings (ie: webhdfs) allows credential stealing

WebHDFS clients might send SPNEGO authorization header to remote URL without proper verification. A maliciously crafted request can trigger services to send server credentials to a webhdfs path (ie: webhdfs://…) for capturing the service principal.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB-2021 406: CVE-2020-9492 Hadoop filesystem bindings (ie: webhdfs) allows credential stealing

Known Issues in Apache Hive

This topic describes known issues and workarounds for using Hive in this release of Cloudera Runtime.

CDPD-23041: DROP TABLE on a table having an index does not work

If you migrate a Hive table to CDP having an index, DROP TABLE does not drop the table. Hive no longer supports indexes (HIVE-18448). A foreign key constraint on the indexed table prevents dropping the table. Attempting to drop such a table results in the following error:

```
java.sql.BatchUpdateException: Cannot delete or update a parent
row: a foreign key constraint fails ("hive"."IDXS", CONSTRAINT "
IDXS_FK1" FOREIGN KEY ("ORIG_TBL_ID") REFERENCES "TBLS ("TBL_ID"
))
```

There are two workarounds:

- Drop the foreign key "IDXS_FK1" on the "IDXS" table within the metastore. You can also manually drop indexes, but do not cascade any drops because the IDXS table includes references to "TBLS".
- Launch an older version of Hive, such as Hive 2.3 that includes IDXS in the DDL, and then drop the indexes as described in Language Manual Indexing.

Apache Issue: Hive-24815

CDPD-676: Generate Oozie workflow for microstrategy

Workaround: None

Apache JIRA: none

Technical Service Bulletins

TSB 2021-480/1: Hive produces incorrect query results when skipping a header in a binary file

In CDP, setting the table property skip.header.line.count to greater than 0 in a table stored in a binary format, such as Parquet, can cause incorrect query results. The skip header property is intended for use with Text files and typically used with CSV files. The issue is not present when you run the query on a Text file that sets the skip header property to 1 or greater.

Upstream JIRA

Apache Jira: HIVE-24827

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-480.1: Hive produces incorrect query results when skipping a header in a binary file

TSB 2021-480/2: Hive ignores the property to skip a header or footer in a compressed file

In CDP, setting the table properties skip.header.line.count and skip.footer.line.count to greater than 0 in a table stored in a compressed format, such as bzip2, can cause incorrect results from SELECT * or SELECT COUNT (*) queries.

Upstream JIRA

Apache Jira: HIVE-24224

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-480.2: Hive ignores the property to skip a header or footer in a compressed file

TSB 2021-482: Race condition in subdirectory delete/rename causes hive jobs to fail

Multiple threads try to perform a rename operation on s3. One of the threads fails to perform a rename operation, causing an error. Hive logs will report "HiveException: Error moving ..." and the log will contain an error line starting with " Exception when loading partition " -all paths listed with s3a:// prefixes.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-482: Race condition in subdirectory delete/rename causes Hive jobs to fail

TSB 2021-501: JOIN queries return wrong result for join keys with large size in Hive

JOIN queries return wrong results when performing joins on large size keys (larger than 255 bytes). This happens when the fast hash table join algorithm is enabled, which is enabled by default.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-501: JOIN queries return wrong result for join keys with large size in Hive

TSB 2021-518: Incorrect results returned when joining two tables with different bucketing versions

Incorrect results are returned when joining two tables with different bucketing versions, and with the following Hive configurations: set hive.auto.convert.join = false and set mapreduce.job.reduces = any custom value.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-518: Incorrect results returned when joining two tables with different bucketing versions

TSB 2023-627: IN/OR predicate on binary column returns wrong result

An IN or an OR predicate involving a binary datatype column may produce wrong results. The OR predicate is converted to an IN due to the setting hive.optimize.point.lookup which is true by default. Only binary data types are affected by this issue. See https://issues.apache.org/jira/browse/HIVE-26235 for example queries which may be affected.

Upstream JIRA

HIVE-26235

Knowledge article

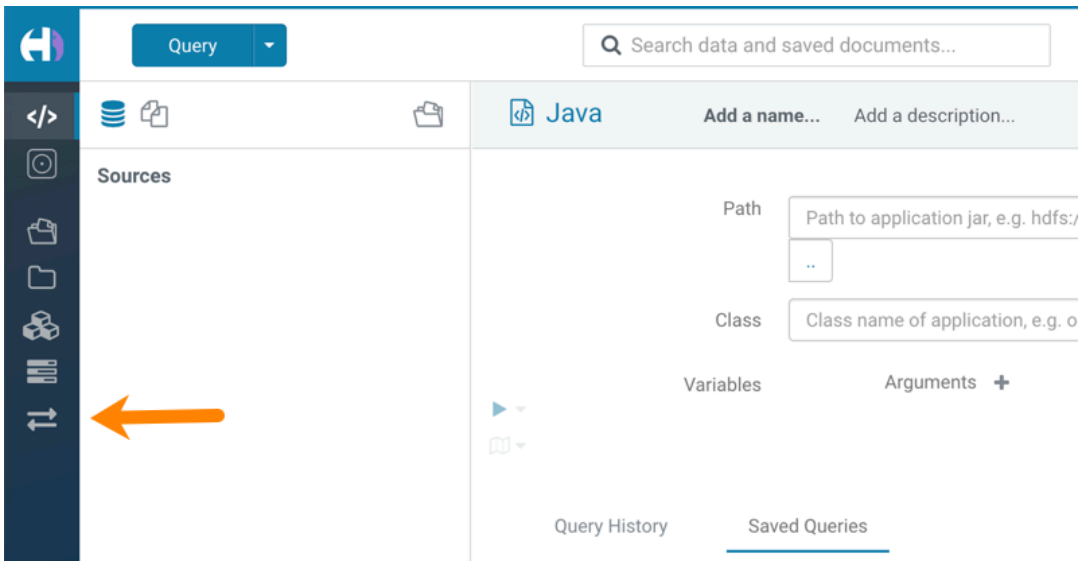
For the latest update on this issue, see the corresponding Knowledge article: TSB 2023-627: IN/OR predicate on binary column returns wrong result

Known Issues in Hue

This topic describes known issues and workarounds for using Hue in this release of Cloudera Runtime.

Hue Importer is not supported in the Data Engineering template

When you create a Data Hub cluster using the Data Engineering template, the Importer application is not supported in Hue:



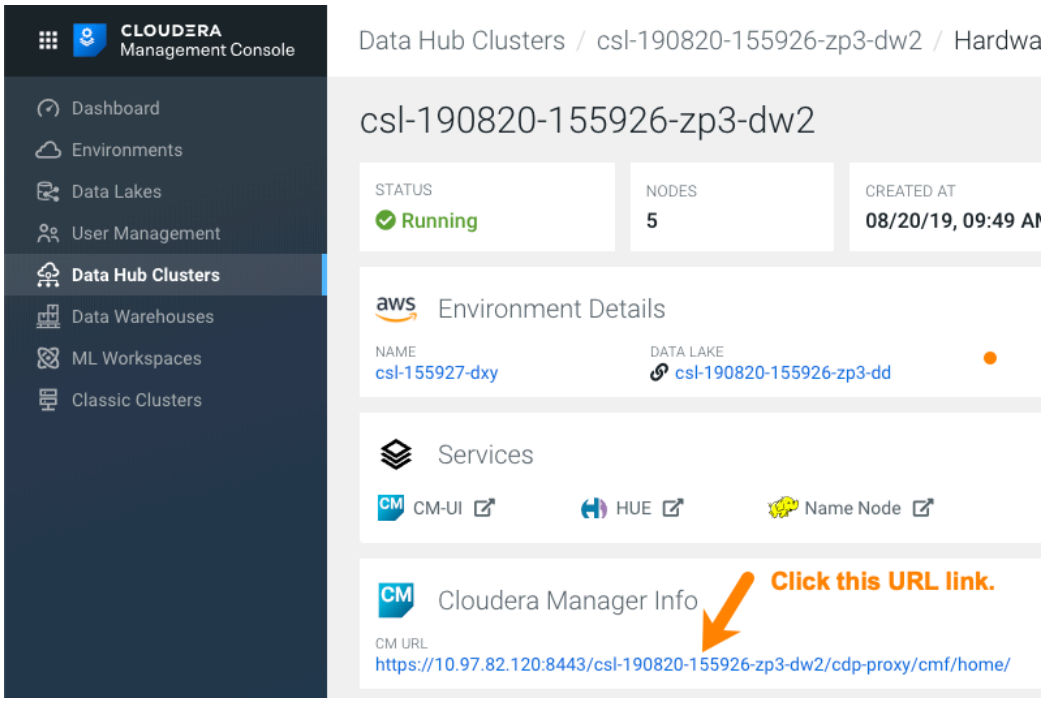
CDPD-3501: Hue-Atlas configuration information is missing on Data Mart clusters.

Problem: The configuration file hive-conf%2Fatlas-application.properties is missing on Data Mart clusters because Apache Hive is not installed. This properties file is needed for the Hue integration with Apache Atlas.

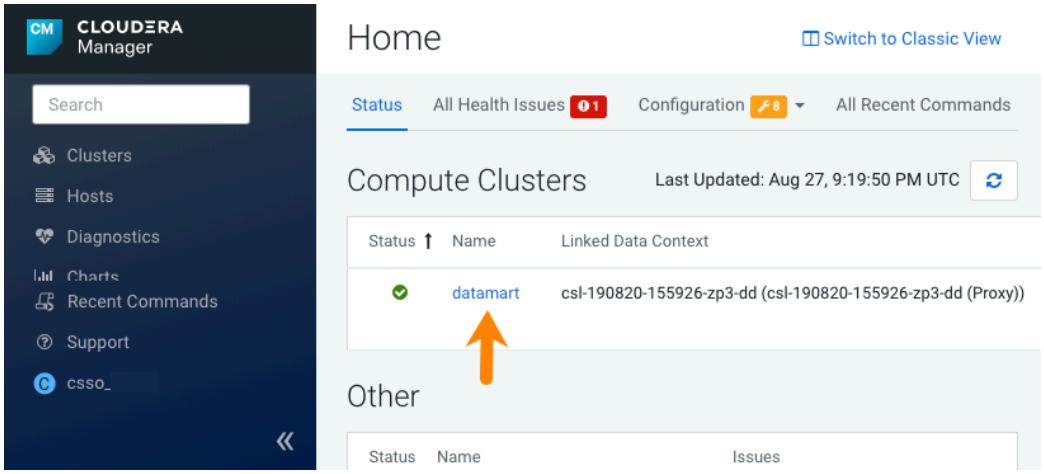
Workaround:

Note: To make the following configuration change, you must have administrative permissions on the Data Mart cluster.

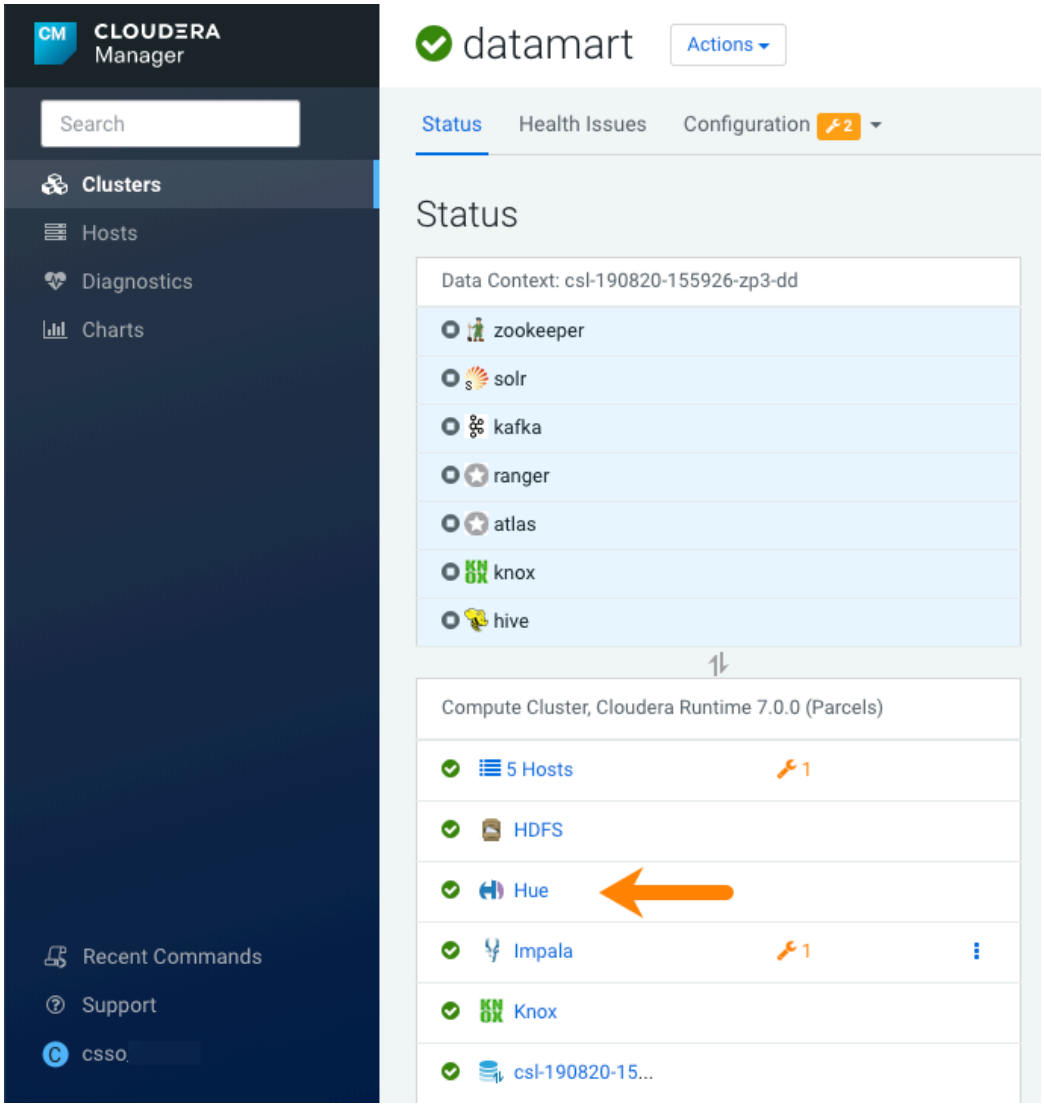
1. Log in to the CDP web interface and navigate to the Data Hub service.
2. On the Data Hub Clusters page, click the Data Mart cluster you want to work on.
3. On the Data Mart cluster page, click the URL link to Cloudera Manager Info:



4. On the Home page of Cloudera Manager, click the cluster name under Compute Clusters:



5. In the cluster page in the Status column under Compute Cluster, Cloudera Runtime, click the link to Hue:



6. On the Hue page, click the Configuration tab to view the configuration properties for Hue.

7. In the search text box, type safety and press Enter to locate the Hue Service Advanced Configuration Snippet (Safety Valve) for hue_safety_valve.ini, which appears at the top of the configuration parameters list.
8. Append the following configuration information to the existing configuration information in the Safety Valve and click Save Changes:

```
[metadata]
[[catalog]]
interface=atlas
api_url=http://master0.cloudera.site:21000/api/atlas/
kerberos_enabled=true
```

9. Restart the Hue service for the configuration change to take effect.

Technical Service Bulletins

TSB 2021-487: Cloudera Hue is vulnerable to Cross-Site Scripting attacks

Multiple Cross-Site Scripting (XSS) vulnerabilities of Cloudera Hue have been found. They allow JavaScript code injection and execution in the application context.

- CVE-2021-29994 - The Add Description field in the Table schema browser does not sanitize user inputs as expected.
- CVE-2021-32480 - Default Home direct button in Filebrowser is also susceptible to XSS attack.
- CVE-2021-32481 - The Error snippet dialog of the Hue UI does not sanitize user inputs.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-487: Cloudera Hue is vulnerable to Cross-Site Scripting attacks (CVE-2021-29994, CVE-2021-32480, CVE-2021-32481)

Known Issues in Apache Impala

This topic describes known issues and workarounds for using Impala in this release of Cloudera Runtime.

For the known issues and workarounds in Impala, see the Impala Known Issues in CDP Knowledge Base article.

Technical Service Bulletins

TSB-2021-485: Impala returns fewer rows from parquet tables on S3

IMPALA-10310 was an issue in Impala's Parquet page filtering code where the scanner did not reset state appropriately when transitioning from the first row group to subsequent row groups in a single split. This caused data from the subsequent row groups to be skipped incorrectly, leading to incorrect query results. This issue cannot occur when the Parquet page filtering is disabled by setting PARQUET_READ_PAGE_INDEX=false.

The issue is more likely to be encountered on S3/ADLS/ABFS/etc, because Spark is sometimes configured to write 128MB row groups and the PARQUET_OBJECT_STORE_SPLIT_SIZE is 256MB. This makes it more likely for Impala to process two row groups in a single split.

Parquet page filtering only works based on the min/max statistics, therefore the comparison operators it supports are “=”, “<”, “>”, “<=”, and “>=”. These operators are impacted by this bug. Expressions such as “!=”, 'LIKE' or the expressions including UDF do not use parquet page filtering.

The PARQUET_OBJECT_STORE_SPLIT_SIZE parameter is introduced in Impala 3.3 by IMPALA-5843. This means that older versions of Impala do not have this issue.

Upstream JIRA

- IMPALA-5843

- IMPALA-10310

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-485: Impala returns fewer rows from parquet tables on S3

TSB 2021-502: Impala logs the session / operation secret on most RPCs at INFO level

Impala logs contain the session / operation secret. With this information a person who has access to the Impala logs might be able to hijack other users' sessions. This means the attacker is able to execute statements for which they do not have the necessary privileges otherwise. Impala deployments where Apache Sentry or Apache Ranger authorization is enabled may be vulnerable to privilege escalation. Impala deployments where audit logging is enabled may be vulnerable to incorrect audit logging.

Restricting access to the Impala logs that expose secrets will reduce the risk of an attack. Additionally, restricting access to trusted users for the Impala deployment will also reduce the risk of an attack. Log redaction techniques can be used to redact secrets from the logs. For more information, see the *Cloudera Manager documentation*.

For log redaction, users can create a rule with a search pattern: secret \(string\) [=:].*And the replacement could be for example: secret=LOG-REDACTED

Upstream JIRA

IMPALA-10600

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-502: Impala logs the session / operation secret on most RPCs at INFO level

Known Issues in Apache Kafka

This topic describes known issues and unsupported features for using Kafka in this release of Cloudera Runtime.

Known Issues

Topics created with the kafka-topics tool are only accessible by the user who created them when the deprecated --zookeeper option is used

By default all created topics are secured. However, when topic creation and deletion is done with the kafka-topics tool using the --zookeeper option, the tool talks directly to Zookeeper. Because security is the responsibility of ZooKeeper authorization and authentication, Kafka cannot prevent users from making ZooKeeper changes. As a result, if the --zookeeper option is used, only the user who created the topic will be able to carry out administrative actions on it. In this scenario Kafka will not have permissions to perform tasks on topics created this way.

Workaround: Use kafka-topics with the --bootstrap-server option that does not require direct access to Zookeeper.

Certain Kafka command line tools require direct access to Zookeeper

The following command line tools talk directly to ZooKeeper and therefore are not secured via Kafka:

- kafka-configs
- kafka-reassign-partitions

Workaround:None.

The offsets.topic.replication.factor property must be less than or equal to the number of live brokers

The offsets.topic.replication.factor broker configuration is now enforced upon auto topic creation. Internal auto topic creation will fail with a GROUP_COORDINATOR_NOT_AVAILABLE error until the cluster size meets this replication factor requirement.

Workaround: None.

Requests fail when sending to a nonexistent topic with auto.create.topics.enable set to true

The first few produce requests fail when sending to a nonexistent topic with auto.create.topics.enable set to true.

Workaround: Increase the number of retries in the producer configuration setting retries.

Custom Kerberos principal names cannot be used for kerberized ZooKeeper and Kafka instances

When using ZooKeeper authentication and a custom Kerberos principal, Kerberos-enabled Kafka does not start. You must disable ZooKeeper authentication for Kafka or use the default Kerberos principals for ZooKeeper and Kafka.

Workaround: None.

Performance degradation when SSL Is enabled

In some configuration scenarios, significant performance degradation can occur when SSL is enabled. The impact varies depending on your CPU, JVM version, Kafka configuration, and message size. Consumers are typically more affected than producers.

Workaround: Configure brokers and clients with ssl.secure.random.implementation = SHA1PRNG. It often reduces this degradation drastically, but its effect is CPU and JVM dependent.

Apache JIRA: KAFKA-2561

OPSAPS-43236: Kafka garbage collection logs are written to the process directory

By default Kafka garbage collection logs are written to the agent process directory. Changing the default path for these log files is currently unsupported.

Workaround: None.

OPSAPS-57113: The Kafka Broker Advanced Configuration Snippet (Safety Valve) for ssl.properties does not propagate configurations correctly.

If the Kafka Broker Advanced Configuration Snippet (Safety Valve) for ssl.properties property contains configuration that has dollar signs, the configuration is not propagated to Kafka brokers correctly.

Workaround:None.

Unsupported Features

The following Kafka features are not supported in Cloudera Data Platform:

- Only Java based clients are supported. Clients developed with C, C++, Python, .NET and other languages are currently not supported.
- Kafka Connect is not supported. NiFi is a proven solution for batch and real time data loading that complements Kafka's message broker capability. For more information, see Cloudera Flow Management.
- The Kafka default authorizer is not supported. This includes setting ACLs and all related APIs, broker functionality, and command-line tools.

Known Issues in Apache Knox

This topic describes known issues and workarounds for using Knox in this release of Cloudera Runtime.
CDPD-3125: Logging out of Atlas does not manage the external authentication

At this time, Atlas does not communicate a log-out event with the external authentication management, Apache Knox. When you log out of Atlas, you can still open the instance of Atlas from the same web browser without re-authentication.

Workaround: To prevent additional access to Atlas, close all browser windows and exit the browser.

Technical Service Bulletins

TSB 2022-553: DOM based XSS Vulnerability in Apache Knox

When using Knox Single Sign On (SSO) in the affected releases, a request could be crafted to redirect a user to a malicious page due to improper URL parsing. The request includes a specially crafted request parameter that could be used to redirect the user to a page controlled by an attacker. This request URL would need to be presented to the user outside the normal request flow through a XSS or phishing campaign.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2022-553: DOM based XSS Vulnerability in Apache Knox ("Knox")

Known Issues in Apache Kudu

This topic describes known issues and workarounds for using Kudu in this release of Cloudera Runtime.

- Kudu supports only coarse grain authorization
  - Kudu does not yet support integration with Ranger
  - Kudu does not yet support integration with Atlas
- Kudu HMS Sync is disabled and is not yet supported

Known Issues in Apache Oozie

This topic describes known issues and unsupported features for using Oozie in this release of Cloudera Runtime.

Oozie jobs fail (gracefully) on secure YARN clusters when JobHistory server is down

If the JobHistory server is down on a YARN (MRv2) cluster, Oozie attempts to submit a job, by default, three times. If the job fails, Oozie automatically puts the workflow in a SUSPEND state.

Workaround: When the JobHistory server is running again, use the resume command to inform Oozie to continue the workflow from the point at which it left off.

CDPD-5340: The resourceManager property defined in an Oozie workflow might not work properly if the workflow is submitted through Knox proxy.

An Oozie workflow defined to use the resourceManager property might not work as expected in situations when the workflow is submitted through Knox proxy.

Workaround: Define the jobTracker property with the same value as that of the resourceManager property.

Unsupported Feature

The following Oozie features are currently not supported in Cloudera Data Platform:

- Non-support for Pig action (CDPD-1070)
- Conditional coordinator input logic

Cloudera does not support using Derby database with Oozie. You can use it for testing or debugging purposes, but Cloudera does not recommend using it in production environments. This could cause failures while upgrading from CDH to CDP.

Known Issues in Apache Ranger

This topic describes known issues and workarounds for using Ranger in this release of Cloudera Runtime.

CDPD-3296: Audit files for Ranger plugin components do not appear immediately in S3 after cluster creation

For Ranger plugin components (Atlas, Hive, HBase, etc.), audit data is updated when the applicable audit file is rolled over. The default Ranger audit rollover time is 24 hours, so audit data appears 24 hours after cluster creation.

Workaround:

To see the audit logs in S3 before the default rollover time of 24 hours, use the following steps to override the default value in the Cloudera Manager safety valve for the applicable service.

1. On the Configuration tab in the applicable service, select Advanced under CATEGORY.
2. Click the + icon for the <service_name> Advanced Configuration Snippet (Safety Valve) for ranger-<service_name>-audit.xml property.
3. Enter the following property in the Name box:

   xasecure.audit.destination.hdfs.file.rollover.sec.
4. Enter the desired rollover interval (in seconds) in the Value box. For example, if you specify 180, the audit log data is updated every 3 minutes.
5. Click Save Changes and restart the service.

Known Issues in Apache Solr

This topic describes known issues and workarounds for using Solr in this release of Cloudera Runtime.

Technical Service Bulletins

TSB-2021 389: CVE 2019-17558: Remote Code Execution in Solr through Velocity templates

Apache Solr in certain CDH, HDP and CDP releases are vulnerable to a Remote Code Execution through the VelocityResponseWriter. A Velocity template can be provided through Velocity templates in a configset `velocity/` directory or as a parameter. A user defined configset could contain renderable, potentially malicious, templates. Parameter provided templates are disabled by default, but can be enabled by setting `params.resource.loader.enabled` by defining a response writer with that setting set to `true`. Defining a response writer requires configuration API access. The fix removed the params resource loader entirely, and only enables the configset-provided template rendering when the configset is `trusted` (has been uploaded by an authenticated user).

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB-2021 389: CVE 2019-17558: Remote Code Execution in Solr through Velocity templates

TSB 2021-497: CVE-2021-27905: Apache Solr SSRF vulnerability with the Replication handler

The Apache Solr ReplicationHandler (normally registered at "/replication" under a Solr core) has a "masterUrl" (also "leaderUrl" alias) parameter. The “masterUrl” parameter is used to designate another ReplicationHandler on another Solr core to replicate index data into the local core. To help prevent the CVE-2021-27905 SSRF vulnerability, Solr should check these parameters against a similar configuration used for the "shards" parameter.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2021-497: CVE-2021-27905: Apache Solr SSRF vulnerability with the Replication handler

Known Issues in Apache Spark

This topic describes known issues and workarounds for using Spark in this release of Cloudera Runtime.

CDPD-217: HBase/Spark connectors are not supported

The *Apache HBase Spark Connector* (hbase-connectors/spark) and the *Apache Spark - Apache HBase Connector* (shc) are not supported in the initial CDP release.

Workaround: None

CDPD-3038: Launching pyspark displays several HiveConf warning messages

When pyspark starts, several Hive configuration warning messages are displayed, similar to the following:

```
19/08/09 11:48:04 WARN conf.HiveConf: HiveConf of name hive.vect
orized.use.checked.expressions does not exist
19/08/09 11:48:04 WARN conf.HiveConf: HiveConf of name hive.te
z.cartesian-product.enabled does not exist
```

Workaround: These errors can be safely ignored.

CDPD-2650: Spark cannot write ZSTD and LZ4 compressed Parquet to dynamically partitioned tables

Workaround: Use a different compression algorithm.

CDPD-3293: Cannot create views (CREATE VIEW statement) from Spark

Apache Ranger in CDP disallows Spark users from running CREATE VIEW statements.

Workaround: Create the view using Hive or Impala.

Known Issues for Apache Sqoop

This topic describes known issues and workarounds for using Parquet and Avro imports in this release of Cloudera Runtime.

Avro, S3, and HCat do not work together properly

Problem: Importing an Avro file into S3 with HCat fails with Delegation Token not available.

CDPD-3089

Parquet columns inadvertently renamed

Problem: Column names that start with a number are renamed when you use the --as-parquetfile option to import data.

Workaround: Prepend column names in Parquet tables with one or more letters or underscore characters.

Apache JIRA: None

Importing Parquet files might cause out-of-memory (OOM) errors

Problem: Importing multiple megabytes per row before initial-page-run check (ColumnWriter) can cause OOM. Also, rows that vary significantly by size so that the next-page-size check is based on small rows, and is set very high, followed by many large rows can also cause OOM.

PARQUET-99

Known Issues in MapReduce and YARN

This topic describes known issues, unsupported features and limitations for using MapReduce and YARN in this release of Cloudera Runtime.

Known Issues

JobHistory URL mismatch after server relocation

After moving the JobHistory Server to a new host, the URLs listed for the JobHistory Server on the ResourceManager web UI still point to the old JobHistory Server. This affects existing jobs only. New jobs started after the move are not affected.

Workaround: For any existing jobs that have the incorrect JobHistory Server URL, there is no option other than to allow the jobs to roll off the history over time. For new jobs, make sure that all clients have the updated mapred-site.xml that references the correct JobHistory Server.

CDH-49165: History link in ResourceManager web UI broken for killed Spark applications

When a Spark application is killed, the history link in the ResourceManager web UI does not work.

Workaround: To view the history for a killed Spark application, see the Spark HistoryServer web UI instead.

CDH-6808: Routable IP address required by ResourceManager

ResourceManager requires routable host:port addresses for yarn.resourcemanager.scheduler.address, and does not support using the wildcard 0.0.0.0 address.

Workaround: Set the address, in the form host:port, either in the client-side configuration, or on the command line when you submit the job.

OPSAPS-52066: Stacks under Logs Directory for Hadoop daemons are not accessible from Knox Gateway.

Stacks under the Logs directory for Hadoop daemons, such as NameNode, DataNode, ResourceManager, NodeManager, and JobHistoryServer are not accessible from Knox Gateway.

Workaround: Administrators can SSH directly to the Hadoop Daemon machine to collect stacks under the Logs directory.

CDPD-2936: Application logs are not accessible in WebUI2 or Cloudera Manager

Running Containers Logs from NodeManager local directory cannot be accessed either in Cloudera Manager or in WebUI2 due to log aggregation.

Workaround: Use the YARN log CLI to access application logs. For example:

```
yarn logs -applicationId <Application ID>
```

Apache Issue: YARN-9725

OPSAPS-50291: Environment variables HADOOP_HOME, PATH, LANG, and TZ are not getting whitelisted

It is possible to whitelist the environment variables HADOOP_HOME, PATH, LANG, and TZ, but the container launch environments do not have these variables set up automatically.

Workaround: You can manually add the required environment variables to the whitelist using Cloudera Manager.

1. In Cloudera Manager, select the YARN service.
2. Click the Configuration tab.
3. Search for Containers Environment Variable Whitelist.
4. Add the environment variables (HADOOP_HOME, PATH, LANG, TZ) which are required to the list.
5. Click Save Changes.
6. Restart all NodeManagers.
7. Check the YARN aggregated logs to ensure that newly whitelisted environment variables are set up for container launch.

COMPX-8687: Missing access check for getAppAttemps

When the Job ACL feature is enabled using Cloudera Manager (YARN Configuration Enablg JOB ACL property), the mapreduce.cluster.acls.enabled property is not generated to all configuration files, including the yarn-site.xml configuration file. As a result the ResourceManager process will use the default value of this property. The default property of mapreduce.cluster.acls.enabled is false.

Workaround: Enable the Job ACL feature using an advanced configuration snippet:

1. In Cloudera Manager select the YARN service.
2. Click Configuration.
3. Find the YARN Service MapReduce Advanced Configuration Snippet (Safety Valve) property.
4. Click the plus icon and add the following:
   - Name: mapreduce.cluster.acls.enabled
   - Value: true
5. Click Save Changes.

Limitations

Capacity Scheduler

- As Capacity Scheduler is the default scheduler, the Dynamic Resource Pool User Interface is not available by default.
- Capacity Scheduler can be configured only through safety-valves in Cloudera Manager.

Unsupported Features

The following YARN features are currently not supported in Cloudera Data Platform:

- GPU support for Docker
- Hadoop Pipes
- Fair Scheduler
- Application Timeline Server (ATS 2 and ATS 1.5)
- Container Resizing
- Distributed or Centralized Allocation of Opportunistic Containers
- Distributed Scheduling
- Native Services
- Pluggable Scheduler Configuration
- Queue Priority Support
- Reservation REST APIs
- Resource Estimator Service
- Resource Profiles
- (non-Zookeeper) ResourceManager State Store
- Shared Cache
- YARN Federation
- New Aggregated Log File Format
- Node Labels
- Rolling Log Aggregation
- YARN WebUI v2
- Docker on YARN (DockerContainerExecutor)
- Moving jobs between queues
- Dynamic Resource Pools

Known Issues in Apache Zeppelin

This topic describes known issues and workarounds for using Zeppelin in this release of Cloudera Runtime.

CDPD-1683: Zeppelin demo users have been removed

Workaround: Use cluster users to access Zeppelin. For information on provisioning users in CDP, see Onboarding users.

CDPD-880, CDPD-1685: Shell, JDBC, and Spark interpreters have been removed

Workaround: Use an available interpreter. For Spark functionality, use the Livy interpreter.

CDPD-2406: Logout button does not work

Clicking the Logout button in the Zeppelin UI logs you out, but then immediately logs you back in using SSO.

Workaround: Close the browser.

Cloudera Runtime Component Versions

Versions of Cloudera Runtime 7.0.2 components.

Component	Version
Apache Atlas	2.0.0
Apache Avro	1.8.2
Apache Hadoop	3.1.1
Apache HBase	2.2.2
HBase Indexer	1.5.0
Apache Hive	3.1.2
Hue	4.5.0
Apache Impala	3.3.0
Apache Kafka	2.3.0
Apache Kudu	1.11.0
Apache Oozie	5.1.0
Apache ORC	1.5.1
Apache Parquet	1.10.99
Apache Ranger	2.0.0
Apache Solr	7.4.0
Apache Spark	2.4.0
Apache Sqoop	1.4.7
Apache Tez	0.9.1
Apache Zookeeper	3.5.5