

Cloudera Runtime 7.2.0

Release Notes

Date published: 2020-06-06

Date modified:

CLOUdera

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2024. 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

Overview.....	5
Cloudera Runtime Component Versions.....	5
What's New In Cloudera Runtime 7.2.0.....	6
What's new in DAS.....	6
What's New in Apache HBase.....	6
What's New in Apache Hadoop HDFS.....	6
What's New in Apache Hive.....	7
What's New in Hue.....	7
What's New in Apache Impala.....	9
What's New in Apache Kafka.....	10
What's New in Apache Knox.....	11
What's New in Apache Kudu.....	11
What's New in Apache Oozie.....	12
What's New in Apache Phoenix.....	13
What's New in Schema Registry.....	13
What's New in Cloudera Search.....	13
What's New in Apache Spark.....	13
What's New in Sqoop.....	13
What's new in Streams Messaging Manager.....	13
What's New in Apache Hadoop YARN.....	13
What's New in Apache ZooKeeper.....	14
Deprecation Notices In Cloudera Runtime 7.2.0.....	14
Deprecation notices in Apache Kudu.....	14
Deprecation Notices for Apache Kafka.....	15
Deprecation Notices in Apache HBase.....	15
Deprecation notices in Apache Impala.....	25
Fixed Issues In Cloudera Runtime 7.2.0.....	25
Fixed Issues in Atlas.....	25
Fixed issues in DAS.....	26
Fixed Issues in Hadoop.....	26
Fixed Issues in HBase.....	26
Fixed Issues in HDFS.....	26
Fixed Issues in Hive.....	26
Fixed Issues in Hue.....	26
Fixed Issues in Impala.....	27
Fixed Issues in Kafka.....	27
Fixed Issues in Kudu.....	27
Fixed Issues in Oozie.....	27
Fixed Issues in Phoenix.....	28
Fixed Issues in Ranger.....	28
Fixed Issues in Schema Registry.....	28

Fixed Issues in Spark.....	29
Fixed Issues in Streams Messaging Manager.....	29
Fixed Issues in Zeppelin.....	29

Known Issues In Cloudera Runtime 7.2.0..... 29

Known Issues in Apache Atlas.....	29
Known Issues in DAS.....	33
Known Issues in Apache HBase.....	34
Known Issues in HDFS.....	36
Known Issues in Apache Hive.....	37
Known Issues in Hue.....	39
Known Issues in Apache Impala.....	44
Known Issues in Apache Kafka.....	49
Known Issues in Kerberos.....	51
Known Issues in Apache Knox.....	52
Known Issues in Apache Kudu.....	52
Known Issues in Apache Oozie.....	52
Known Issues in Apache Phoenix.....	53
Known Issues in Apache Ranger.....	53
Known Issues in Cloudera Search.....	53
Known Issues in Apache Solr.....	58
Known Issues in Apache Spark.....	59
Known Issues for Apache Sqoop.....	60
Known issues in Streams Messaging Manager.....	61
Known Issues in MapReduce and YARN.....	61
Known Issues in Apache Zeppelin.....	64
Known Issues in Apache ZooKeeper.....	65

Overview

Welcome to the Cloudera Runtime Release Notes. This document provides you with the latest information about Cloudera Runtime 7.2.0. This document includes improvements and describes new features, bug fixes, tech previews and more. For detailed information about the runtime components themselves, see [Cloudera documentation](#).

Cloudera Runtime Component Versions

List of the official component versions for Cloudera Runtime. To know the component versions for compatibility with other applications, you must be familiar with the latest component versions in Cloudera Runtime. You should also be aware of the available Technical Preview components and use them only in a testing environment.

Apache Components

Component	Version
Apache Arrow	0.8.0
Apache Atlas	2.0.0
Apache Calcite Avatica	1.16.0
Apache Avro	1.8.2
Apache Curator	4.3.0
Apache Hadoop (Includes YARN and HDFS)	3.1.1
Apache HBase	2.2.3
Apache Hive	3.1.3000
Apache Impala	3.2.0
Apache Kafka	2.4.1
Apache Knox	1.3.0
Apache Kudu	1.12.0
Apache Livy	0.6.0
Apache MapReduce	3.1.1
Apache Oozie	5.1.0
Apache ORC	1.5.1
Apache Parquet	1.10.99
Apache Phoenix	5.0.0
Apache Ranger	2.0.0
Apache Solr	8.4.1
Apache Spark	2.4.5
Apache Sqoop	1.4.7
Apache Tez	0.9.1
Apache Zeppelin	0.8.2
Apache ZooKeeper	3.5.5

Other Components

Component	Version
Data Analytics Studio	1.4.2
GCS Connector	1.9.10
HBase Indexer	1.5.0
Hue	4.5.0
Search	1.0.0
Schema Registry	0.8.1
Streams Messaging Manager	2.1.0

Connectors and Encryption Components

Component	Version
HBase connectors	1.0.0
Hive Meta Store (HMS)	1.0.0
Hive on Tez	1.0.0
Hive Warehouse Connector	1.0.0
Spark Atlas Connector	0.1.0

What's New In Cloudera Runtime 7.2.0

This version of Cloudera Runtime provides you with several new capabilities. Learn how the new features and improvements benefit you.

What's new in DAS

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

- DAS displays all the DAG IDs that are associated with a particular Hive query, and also displays the DAG graph corresponding to each DAG ID.
- The Hive and Tez configurations are separately displayed on the DAS UI. You can view the Tez-related configurations for a query that has a DAG ID associated with it on the DAG Configurations tab in the DAG Info section on the **Query details** and the **Query Compare** page.
- The Edit button on the **Query Details** page redirects you to the **Compose** page and allows you to edit the selected query in the query composer.

What's New in Apache HBase

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

What's New in Apache Hadoop HDFS

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

For more information about HDFS, see [HDFS Overview](#)

What's New in Apache Hive

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

Auto-translation for Spark-Hive reads, no HWC session needed

Reads Hive ACID tables in HMS from Spark directly or through HWC, based on your configuration of `spark.sql.extensions`. The HWC session is created transparently. Use existing Spark application code without modification. For details, see [HWC configuration information](#).

[Hive Warehouse Connector Spark direct reads](#)

Spark Direct Reader is a Spark Datasource V1 implementation for reading Hive ACID, transactional tables from Spark fast. Low-latency analytical processing (LLAP) is not needed. Spark Direct Reader is intended to be used for Extract Transform Load/Extract Load Transform (ETL/ELT) processes.

[Authorization of external file writes from Spark](#)

Ranger now authorizes read/write access to external files from Spark through the Hive metastore API (HMS API) in addition to read/write access to managed Hive tables from Spark through HiveServer (HS2).

Scheduled Queries, Rebuilding Materialized Views Automatically

You can schedule Hive queries to run on a recurring basis, monitor query progress, temporarily ignore a query schedule, and limit the number running in parallel. You can use scheduled queries to start compaction and periodically rebuild materialized views, for example. For details, see the [Apache Hive Language Manual](#). In CDP, you need to enable scheduled queries.

SHOW DATABASES supports SQL LIKE patterns In a SHOW DATABASES LIKE statement, you can use wildcards, and in this release of Hive, specify any character or a single character. [See examples](#).

Support for quantified comparison predicates ANY/SOME/ALL in subqueries

Apache Hive now supports ALL and SOME/ANY for uncorrelated subqueries. Prior to the introduction of this feature, `<>ANY` and `=ALL` were disallowed in these constructs. [See examples](#).

New configuration parameter for Hive: `hive.server2.tez.sessions.per.default.queue`

What's New in Hue

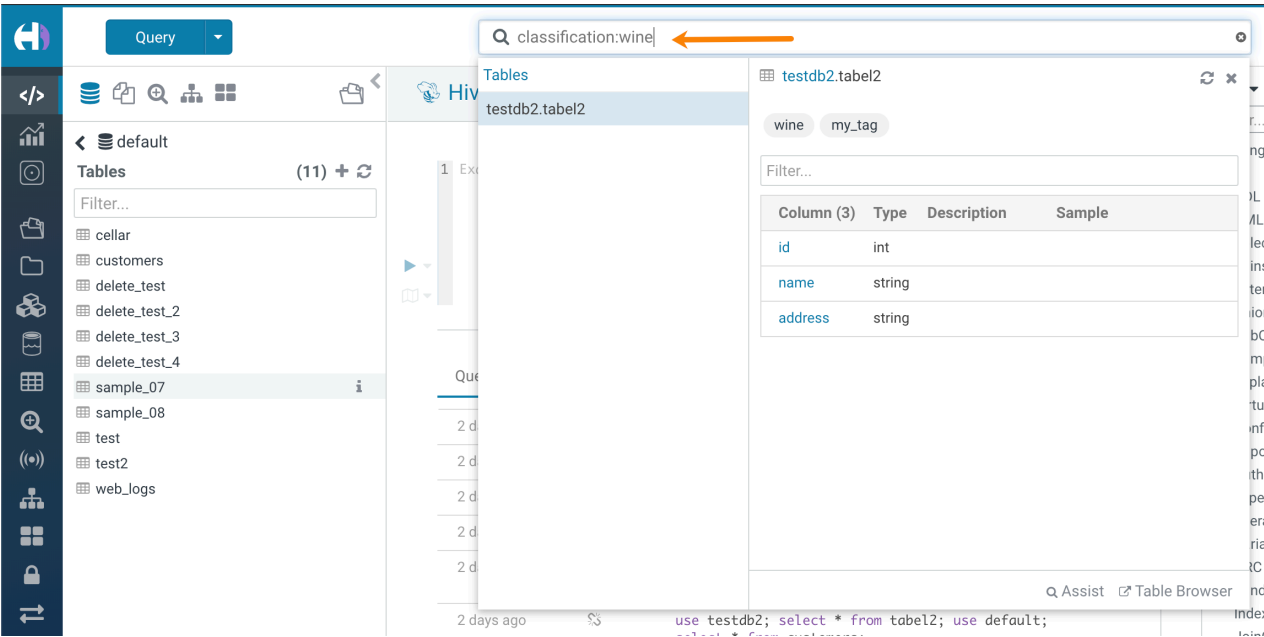
This topic lists new features for Hue in this release of Cloudera Runtime.

Support for Hive on Tez

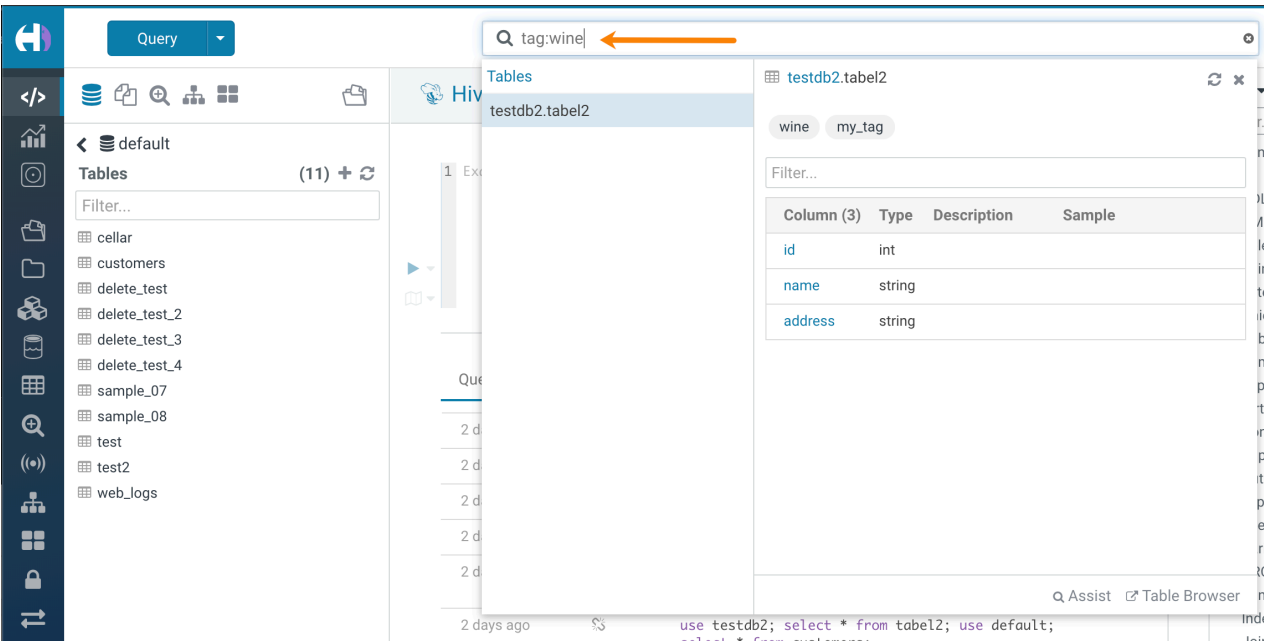
Hue in Cloudera Runtime now supports Hive using Tez as its execution engine.

Integration with Apache Atlas data catalog

You can now locate tables by searching for Atlas classifications in Hue by specifying the classification search term in the search box at the top of the page:

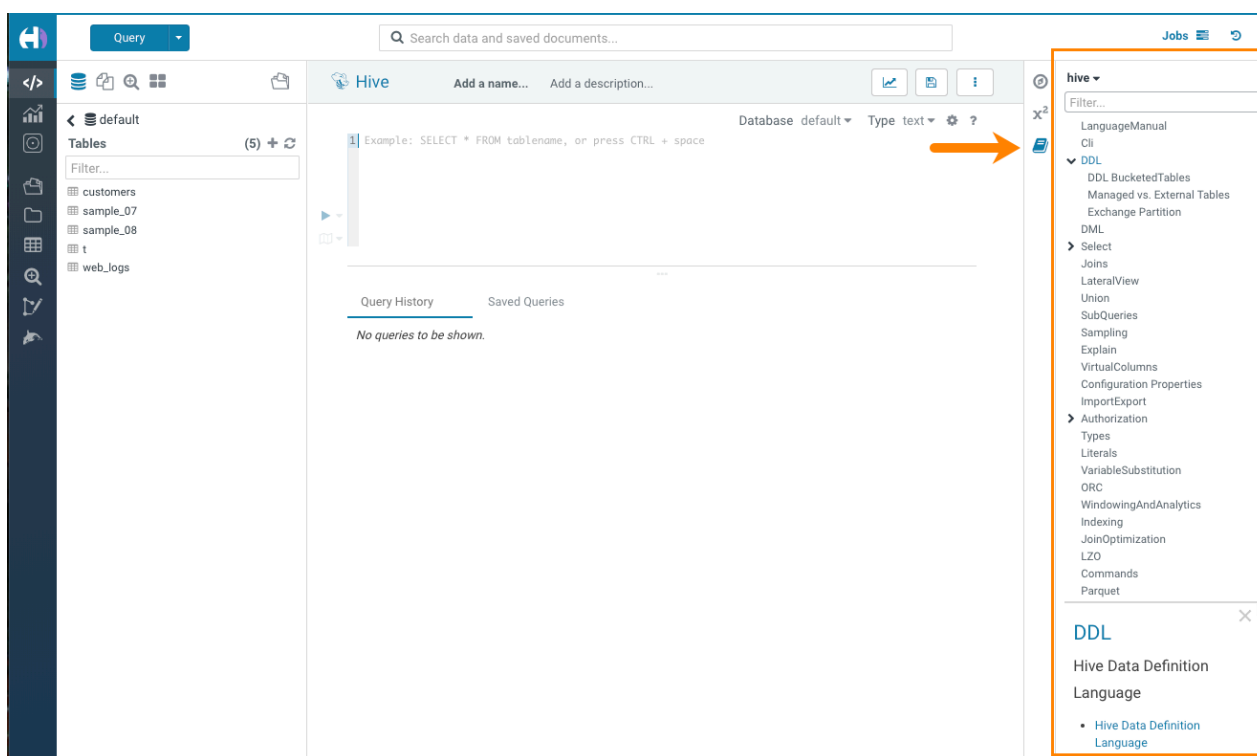


You can also locate tables by searching for Atlas classifications by specifying the tag search term:



Hive Language Reference

A language reference has been added for Hive:



When you select the Hive query editor, click the book icon  to the right of the editor window to launch the Hive language reference.

What's New in Apache Impala

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

Optimized performance for multi-threaded query execution

Multi-threaded query execution can be enabled manually on a per-query basis using the `mt_dop` query option for all `SELECT` queries. Previously queries with joins were not supported.

For details, see [MT_DOP query option](#).

Improved read performance for ORC tables with nested type columns

Improved performance for the automatic updates of metadata

If Impala inserts into a table it refreshes the underlying table/partition. When the configuration `enable_insert_events` is set to `True` Impala will generate `INSERT` event types which when received by other Impala clusters will automatically refresh the tables or partitions. Event processing must be `ON`, for this property to work.

Started generating Ranger audit logs when column masking policy is applied in a policy

Reduced the Impala runtime image size and used UBI base image

Increased scratch capacity

To help reduce spilling to disk:

- Added startup parameter to support Spill-to-disk compression to increase effective scratch capacity by 2.5x.
- Added startup parameter to reclaim space in scratch files.

Improved data cache performance

- Improved the efficiency of the data cache by providing an option to use a different cache eviction algorithm (LIRS).

Support for Kudu Date and Varchar column types

Support for reading ZSTD-compressed text files

For details, see [Using Text Data Files](#).

Improved read performance of ORC tables

Improved Impala resiliency

This release adds client retry support in the impala-shell. For details about installing the impala-shell, see [Using Impala shell](#).

broadcast_bytes_limit query option

In this release, you can set a limit for the size of a broadcast input. For details, see [Impala Query Options](#).

ORC stability and performance improvements

ORC reads enabled by default

Impala stability and performance have been improved. Consequently, ORC reads are now enabled in Impala by default. To disable, set `-enable_orc_scanner` to false when starting the cluster.

Constraints

This release adds support for primary and foreign key constraints, but in this release the constraints are advisory and intended for estimating cardinality during query planning in a future release. There is no attempt to enforce constraints. For details, see the “Constraints” section of [Create Table Statement](#).

Enhanced external Kudu table

By default HMS implicitly translates internal Kudu tables to external Kudu tables with the `'external.table.purge'` property set to true. These tables behave similar to internal tables. You can explicitly create such external Kudu tables. For details, see the “External Kudu Tables” section of [Create Table Statement](#).

Ranger column masking

This release supports Ranger column masking, which hides sensitive columnar data in Impala query output. For example, you can define a policy that reveals only the first or last four characters of column data. Column masking is enabled by default. For details, see the “Ranger Column Masking” section in [Impala Authorization](#).

What's New in Apache Kafka

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

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.

Fine-grained authorization using Ranger

Kudu now supports native fine-grained authorization via integration with Apache Ranger (in addition to integration with Apache Sentry). Kudu may now enforce access control policies defined for Kudu tables and columns stored in Ranger.

Proxy support using Knox

Kudu's web UI now supports proxying via Apache Knox. Kudu can be deployed in a firewalled state behind a Knox Gateway which will forward HTTP requests and responses between clients and the Kudu web UI.

Support for HTTP keep-alive

Kudu's web UI now supports HTTP keep-alive. Operations that access multiple URLs will now reuse a single HTTP connection, improving their performance.

Rolling-restart without stopping on-going Kudu workloads

The kudu tserver quiesce tool is added to quiesce tablet servers. While a tablet server is quiescing, it will stop hosting tablet leaders and stop serving new scan requests. This can be used to orchestrate a rolling restart without stopping on-going Kudu workloads.

Auto time source support for HybridClock timestamps

Introduced auto time source for HybridClock timestamps. With `--time_source=auto` in AWS and GCE cloud environments, Kudu masters and tablet servers use the built-in NTP client synchronized with dedicated NTP servers available via host-only networks. With `--time_source=auto` in environments other than AWS/GCE, Kudu masters and tablet servers rely on local machine's clock synchronized by NTP. The default setting for the HybridClock time source (`--time_source=system`) is backward-compatible, requiring the local machine's clock to be synchronized by the kernel's NTP discipline.

Ability to move replicas away from a tablet server

The kudu cluster rebalance tool now supports moving replicas away from specific tablet servers by supplying the `--ignored_tservers` and `--move_replicas_from_ignored_tservers` arguments.

Ability to specify table creation options using JSON

The kudu table create tool is added to allow users to specify table creation options using JSON.

Ability to automatically rebalance tablet replicas among tablet servers

An experimental feature is added to Kudu that allows it to automatically rebalance tablet replicas among tablet servers. The background task can be enabled by setting the `--auto_rebalancing_enabled` flag on the Kudu masters. Before starting auto-rebalancing on an existing cluster, the CLI rebalancer tool should be run first.

Support for DATE and VARCHAR data types

Kudu now supports DATE and VARCHAR data types.

Optimizations and improvements

- The Write Ahead Log file segments and index chunks are now managed by Kudu's file cache. With that, all the long-lived file descriptors used by Kudu are managed by the file cache, and there's no longer a need for capacity planning file descriptor usage.
- Kudu no longer requires the running of `kudu fs update_dirs` to change a directory configuration or recover from a disk failure
- Kudu tablet servers and masters now expose a tablet-level metric `num_raft_leaders` for the number of tablet replicas hosted on the server
- Kudu's maintenance operation scheduling has been updated to prioritize reducing WAL retention under memory pressure. Kudu would previously prioritize operations that yielded high-memory reduction, which could result in high WAL disk usage in workloads that contained updates
- A new maintenance operation is introduced to remove rowsets that have had all of their rows deleted and whose newest deletes operations are considered ancient
- The built-in NTP client is now fully supported as the time source for Kudu's HybridTime clock. It is no longer marked as experimental. To switch the time source from the existing system time source (which is the default) to the built-in NTP client, use `--time_source=builtin`
- Introduced additional metrics for the built-in NTP client
- Updated `/config` page of masters' and tablet servers' WebUI to display configured and effective time source.

In addition, the effective list of reference servers for the built-in NTP client is shown there as well, if applicable.

- The processing of Raft consensus vote requests has been improved to be more robust during high contention scenarios like election storms.
- Added a validator to enforce consistency between the maximum size of an RPC and the maximum size of tablet transaction memory, controlled by `--rpc_max_message_size` and `--tablet_transaction_memory` flags correspondingly.

In prior releases, if the limit on the size of RPC requests is increased and the limit on tablet transaction memory size is kept with the default setting, certain Raft transactions could be committed but not applied.

- The metrics endpoint now supports filtering metrics by a metric severity level.
- Many kudu `local_replica` tools are updated to not open the block manager, which significantly reduces the amount of IO done when running them
- The Kudu Java client now exposes a way to get the resource metrics associated with a given scanner
- Scan predicates are pushed down to RLE decoders, improving predicate-evaluation-efficiency in some workloads
- The log block manager will now attempt to use multiple threads to open blocks in each data directory, in some tests reducing startup time by up to 20%
- The `raft_term` and `time_since_last_leader_heartbeat` aggregated table metrics will now return the maximum metric reported instead of the sum
- Kudu's tablet server web UI scans page is updated to show the number of round trips per scanner
- Kudu's master and tablet server web UIs are updated to show critical partition information, including tablet count and on-disk size
- Kudu servers now expose the `last_read_elapsed_seconds` and `last_write_elapsed_seconds` tablet-level metrics that indicate how long ago the most recent read and write operations to a given tablet were
- Kudu servers now expose the `transaction_memory_limit_rejections` tablet-level metric that tracks the number of transactions rejected because a given tablet's transactional memory limit was reached

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 Phoenix

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

What's New in Schema Registry

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

What's New in Cloudera Search

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

For more information about Cloudera Search, see [Cloudera Search Overview](#).

What's New in Apache Spark

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

Data engineering cluster

You can create a data engineering cluster in Amazon AWS from within CDP by selecting the Data Engineering cluster template. A data engineering includes Spark, Livy, Hive, Zeppelin, and Oozie, along with supporting services (HDFS, YARN, and Zookeeper).

See [Creating a Cluster on AWS](#).

What's New in Sqoop

Learn what's new in the Apache Sqoop client in Cloudera Runtime 7.2.0.

To access the latest Sqoop documentation on Cloudera's documentation web site, go to [Sqoop Documentation 1.4.7.7.1.6.0](#).

Discontinued maintenance of direct mode

The Sqoop direct mode feature is no longer maintained. This feature was primarily designed to import data from an abandoned database, which is no longer updated. Using direct mode has several drawbacks:

- Imports can cause an intermittent and overlapping input split.
- Imports can generate duplicate data.
- Many problems, such as intermittent failures, can occur.
- Additional configuration is required.

Do not use the `--direct` option in Sqoop import or export commands.

What's new in Streams Messaging Manager

There are no new features for Apache Streams Messaging Manager (SMM) in this release of Cloudera Runtime.

What's New in Apache Hadoop YARN

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

Queue Manager

YARN Queue Manager is the queue management graphical user interface for Apache Hadoop YARN Capacity Scheduler. You can use YARN Queue Manager UI to manage your cluster capacity using queues to balance resource requirements of multiple applications from various users. Using YARN Queue Manager UI, you can set scheduler level properties and queue level properties. You can also view, sort, search, and filter queues using the YARN Queue Manager UI.

For more information about Queue Manager, see [Manage Queues](#).

FPGA as a resource type

You can use FPGA as a resource type. For more information, see [Use FPGA scheduling](#).

New configuration property to enable or disable the YARN recommendation engine APIs

The YARN Recommendation API now recommends scaling cluster nodes up or down based on the demand and idle state of cluster resources. This feature can be turned on/off using the YARN configuration property `yarn.cluster.scaling.recommendation.enable`.

What's New in Apache ZooKeeper

There are no new features in this release.

Deprecation Notices In Cloudera Runtime 7.2.0

Components and features that will be deprecated or removed in this release or a future release.

Terminology

Items in this section are designated as follows:

Deprecated

Technology that Cloudera is removing in a future CDP release. Marking an item as deprecated gives you time to plan for removal in a future CDP release.

Moving

Technology that Cloudera is moving from a future CDP release and is making available through an alternative Cloudera offering or subscription. Marking an item as moving gives you time to plan for removal in a future CDP release and plan for the alternative Cloudera offering or subscription for the technology.

Removed

Technology that Cloudera has removed from CDP and is no longer available or supported as of this release. Take note of technology marked as removed since it can potentially affect your upgrade plans.

Removed Components and Product Capabilities

No components are deprecated or removed in this Cloudera Runtime release.

Please contact Cloudera Support or your Cloudera Account Team if you have any questions.

Deprecation notices in Apache Kudu

This topic lists the features and functionality in Apache Kudu that will be deprecated or removed in this release or a future release.

- The Flume sink has been migrated to the Apache Flume project and removed from Kudu. Users depending on the Flume integration can use the old kudu-flume jars or migrate to the Flume jars containing the Kudu sink.
- Support for Apache Sentry authorization has been deprecated and may be removed in the next release. Users depending on the Sentry integration should migrate to the Apache Ranger integration for authorization.
- Support for Python 2 has been deprecated and may be removed in the next release.
- Support for CentOS/RHEL 6, Debian 8, Ubuntu 14 has been deprecated and may be removed in the next release.

Deprecation Notices for Apache Kafka

Features and functionality that will be deprecated or removed in this release or a future release.

Deprecated

kafka-preferred-replica-election

The kafka-preferred-replica-election.sh command line tool has been deprecated in upstream Apache Kafka 2.4.0. Its alternative in CDP, kafka-preferred.replica-election, is also deprecated.

--zookeeper

The --zookeeper option has been deprecated for all Kafka command line tools except kafka-configs and kafka-reassign-partitions. Cloudera recommends that you use the --bootstrap-server option instead.

Deprecation Notices in Apache HBase

Use this list to understand some of the deprecated items and incompatibilities if you are upgrading from HDP 2.x or CDH 5.x to CDP.

Known Incompatibilities when Upgrading from CDH and HDP

Cloudera Runtime uses Apache HBase 2.x.x whereas CDH 5.x and HDP 2.x uses Apache HBase 1.x.



Important: Some APIs that are listed as deprecated, but these APIs do not block your upgrade. You must stop using the deprecated APIs in your existing applications after upgrade, and not use these APIs in new development.

List of Major Changes

- HBASE-16189 and HBASE-18945: You cannot open the Cloudera Runtime HFiles in CDH or HDP.
- HBASE-18240: Changed the ReplicationEndpoint Interface.
- The Dynamic Jars Directory property hbase.dynamic.jars.dir is disabled by default. If you want to enable dynamic classloading, you can use the hbase.dynamic.jars.dir property in Cloudera Manager to change the default \${hbase.rootdir}/lib directory to some other location, preferably a location on HDFS. This property is flagged by Cloudera Manager as deprecated when you upgrade to CDP because the property is incompatible with HBase on cloud deployments. If you are using HBase with HDFS storage, you can ignore this warning, and keep using the hbase.use.dynamic.jars feature.

Co-processor API changes

- HBASE-16769: Deprecated Protocol Buffers references from MasterObserver and RegionServerObserver.
- HBASE-17312: [JDK8] Use default method for Observer Coprocessors. The interface classes of BaseMasterAndRegionObserver, BaseMasterObserver, BaseRegionObserver, BaseRegionServerObserver and BaseWALObserver uses JDK8's 'default' keyword to provide empty and no-op implementations.
- Interface HTableInterface introduces following changes to the methods listed below:

[#] interface CoprocessorEnvironment

Change	Result
Abstract method <code>getTable (TableName)</code> has been removed.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>getTable (TableName, ExecutorService)</code> has been removed.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.

- Public Audience

The following tables describes the coprocessor changes:

[#] class `CoprocessorRpcChannel` (1)

Change	Result
This class has become interface.	A client program may be interrupted by <code>IncompatibleClassChangeError</code> or <code>InstantiationException</code> exception depending on the usage of this class.

Class `CoprocessorHost<E>`

Classes that were Audience Private but were removed:

Change	Result
Type of field <code>coprocessors</code> has been changed from <code>java.util.SortedSet<E></code> to <code>org.apache.hadoop.hbase.util.SortedList<E></code> .	A client program may be interrupted by <code>NoSuchFieldError</code> exception.

MasterObserver changes

The following changes are introduced to the `MasterObserver` interface:

[#] interface `MasterObserver` (14)

Change	Result
Abstract method <code>voidpostCloneSnapshot (ObserverContext<MasterCoprocesorEnvironment>, HBaseProtos.SnapshotDescription, HTableDescriptor)</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>voidpostCreateTable (ObserverContext<MasterCoprocesorEnvironment>, HTableDescriptor, HRegionInfo[])</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>voidpostDeleteSnapshot (ObserverContext<MasterCoprocesorEnvironment>, HBaseProtos.SnapshotDescription)</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>voidpostGetTableDescriptors (ObserverContext<MasterCoprocesorEnvironment>, List<HTableDescriptor>)</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>voidpostModifyTable (ObserverContext<MasterCoprocesorEnvironment>, TableName, HTableDescriptor)</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>voidpostRestoreSnapshot (ObserverContext<MasterCoprocesorEnvironment>, HBaseProtos.SnapshotDescription, HTableDescriptor)</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>voidpostSnapshot (ObserverContext<MasterCoprocesorEnvironment>, HBaseProtos.SnapshotDescription, HTableDescriptor)</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.
Abstract method <code>voidpreCloneSnapshot (ObserverContext<MasterCoprocesorEnvironment>, HBaseProtos.SnapshotDescription, HTableDescriptor)</code> has been removed from this interface.	A client program may be interrupted by <code>NoSuchMethodError</code> exception.

Abstract method voidpreCreateTable (ObserverContext<MasterCoproprocessorEnvironment>, HTableDescriptor, HRegionInfo[]) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpreDeleteSnapshot (ObserverContext<MasterCoproprocessorEnvironment>, HBaseProtos.SnapshotDescription) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpreGetTableDescriptors (ObserverContext<MasterCoproprocessorEnvironment>, List<TableName>, List<HTableDescriptor>) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpreModifyTable (ObserverContext<MasterCoproprocessorEnvironment>, TableName, HTableDescriptor) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpreRestoreSnapshot (ObserverContext<MasterCoproprocessorEnvironment>, HBaseProtos.SnapshotDescription, HTableDescriptor) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpreSnapshot (ObserverContext<MasterCoproprocessorEnvironment>, HBaseProtos.SnapshotDescription, HTableDescriptor) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.

RegionObserver interface changes

The following changes are introduced to the RegionObserver interface.

[#] interface RegionObserver (13)

Change	Result
Abstract method voidpostCloseOperation (ObserverContext<RegionCoproprocessorEnvironment>, HRegion.Operation) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpostCompactSelection (ObserverContext<RegionCoproprocessorEnvironment>, Store, ImmutableList<StoreFile>) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpostCompactSelection (ObserverContext<RegionCoproprocessorEnvironment>, Store, ImmutableList<StoreFile>, CompactionRequest) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpostGetClosestRowBefore (ObserverContext<RegionCoproprocessorEnvironment>, byte[], byte[], Result) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method DeleteTrackerpostInstantiateDeleteTracker (ObserverContext<RegionCoproprocessorEnvironment>, DeleteTracker) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpostSplit (ObserverContext<RegionCoproprocessorEnvironment>, HRegion, HRegion) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpostStartRegionOperation (ObserverContext<RegionCoproprocessorEnvironment>, HRegion.Operation) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method StoreFile.ReaderpostStoreFileReaderOpen (ObserverContext<RegionCoproprocessorEnvironment>, FileSystem, Path, FSDataInputStreamWrapper, long, CacheConfig, Reference, StoreFile.Reader) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method voidpostWALRestore (ObserverContext<RegionCoproprocessorEnvironment>, HRegionInfo, HLogKey, WALEdit) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.
Abstract method InternalScannerpreFlushScannerOpen (ObserverContext<RegionCoproprocessorEnvironment>, Store, KeyValueScanner, InternalScanner) has been removed from this interface.	A client program may be interrupted by NoSuchMethodErrorException.

Abstract method voidpreGetClosestRowBefore (ObserverContext<RegionCoproprocessorEnvironment>, byte[], byte[], Result) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method StoreFile.ReaderpreStoreFileReaderOpen (ObserverContext<RegionCoproprocessorEnvironment>, FileSystem, Path, FSDataInputStreamWrapper, long, CacheConfig, Reference, StoreFile.Reader) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method voidpreWALRestore (ObserverContext<RegionCoproprocessorEnvironment>, HRegionInfo, HLogKey, WALEdit) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

WALObserver interface changes

The following changes are introduced to the WALObserver interface:

[#] interface WALObserver

Change	Result
Abstract method voidpostWALWrite (ObserverContext<WALCoproprocessorEnvironment>, HRegionInfo, HLogKey, WALEdit) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method booleanpreWALWrite (ObserverContext<WALCoproprocessorEnvironment>, HRegionInfo, HLogKey, WALEdit) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

Scheduler changes

Following methods are now changed to abstract:

[#]class RpcScheduler (1)

Change	Result
Abstract method void dispatch (CallRunner) has been removed from this class.	A client program may be interrupted by NoSuchMethodError exception.

[#] RpcScheduler.dispatch (CallRunner p1) [abstract] : void 1

org/apache/hadoop/hbase/ipc/RpcScheduler.dispatch:(Lorg/apache/hadoop/hbase/ipc/CallRunner;)V

Change	Result
Return value type has been changed from void to boolean.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

The following abstract methods have been removed:

[#]interface PriorityFunction (2)

Change	Result
Abstract method longgetDeadline (RPCProtos.RequestHeader, Message) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method int getPriority (RPCProtos.RequestHeader, Message) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

Server API changes

[#] class RpcServer (12)

Change	Result
--------	--------

Type of field CurCall has been changed from java.lang.ThreadLocal<RpcServer.Call> to java.lang.ThreadLocal<RpcCall>.	A client program may be interrupted by NoSuchFieldError exception.
Abstract method int getNumOpenConnections () has been added to this class.	This class became abstract and a client program may be interrupted by InstantiationException exception.
Field callQueueSize of type org.apache.hadoop.hbase.util.Counter has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field connectionList of type java.util.List<RpcServer.Connection> has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field maxIdleTime of type int has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field numConnections of type int has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field port of type int has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field purgeTimeout of type long has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field responder of type RpcServer.Responder has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field socketSendBufferSize of type int has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field thresholdIdleConnections of type int has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.

Following abstract methods are removed:

Change	Result
Abstract method Pair<Message,CellScanner>call (BlockingService, Descriptors.MethodDescriptor, Message, CellScanner, long, MonitoredRPCHandler) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

Replication and WAL changes

HBASE-18733: WALKey has been purged completely. Following are the changes to the WALKey:

[#] classWALKey (8)

Change	Result
Access level of field clusterIds has been changed from protected to private.	A client program may be interrupted by IllegalAccessException exception.
Access level of field compressionContext has been changed from protected to private.	A client program may be interrupted by IllegalAccessException exception.
Access level of field encodedRegionName has been changed from protected to private.	A client program may be interrupted by IllegalAccessException exception.
Access level of field tablename has been changed from protected to private.	A client program may be interrupted by IllegalAccessException exception.
Access level of field writeTime has been changed from protected to private.	A client program may be interrupted by IllegalAccessException exception.

Following fields have been removed:

Change	Result
Field LOG of type org.apache.commons.logging.Log has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field VERSION of type WALKey.Version has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field logSeqNum of type long has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.

Admin Interface API changes

You cannot administer a CDP Runtime Data Hub cluster using a client that includes RelicationAdmin, ACC, Thrift and REST usage of Admin ops. Methods returning protobufs have been changed to return POJOs instead. Returns have changed from void to Future for async methods. HBASE-18106 - Admin.listProcedures and Admin.listLocks were renamed to getProcedures and getLocks. MapReduce makes use of Admin doing following admin.getClusterStatus() to calculate Splits.

- Thrift usage of Admin API:

```
compact(ByteBuffer) createTable(ByteBuffer, List<ColumnDescriptor>) deleteTable(ByteBuffer) disableTable(ByteBuffer) enableTable(ByteBuffer) getTableNames() majorCompact(ByteBuffer)
```

- REST usage of Admin API:

```
hbase-rest org.apache.hadoop.hbase.rest RootResource getTableList() TableName[] tableNames = servlet.getAdmin().listTableNames(); SchemaResource delete(UriInfo) Admin admin = servlet.getAdmin(); update(TableSchemaModel, boolean, UriInfo) Admin admin = servlet.getAdmin(); StorageClusterStatusResource get(UriInfo) ClusterStatus status = servlet.getAdmin().getClusterStatus(); StorageClusterVersionResource get(UriInfo) model.setVersion(servlet.getAdmin().getClusterStatus().getHBaseVersion()); TableResource exists() return servlet.getAdmin().tableExists(TableName.valueOf(table));
```

[#] interface Admin (9)

Following are the changes to the Admin interface:

Change	Result
Abstract method createTableAsync (HTableDescriptor, byte[] []) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method disableTableAsync (TableName) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method enableTableAsync (TableName) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method getCompactionState (TableName) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method getCompactionStateForRegion (byte[]) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method isSnapshotFinished (HBaseProtos.SnapshotDescription) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method snapshot (String, TableName, HBaseProtos.SnapshotDescription.Type) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method snapshot (HBaseProtos.SnapshotDescription) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method takeSnapshotAsync (HBaseProtos.SnapshotDescription) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

[#] Admin.createTableAsync (HTableDescriptor p1, byte[] [] p2) [abstract] : void 1

org/apache/hadoop/hbase/client/Admin.createTableAsync:(Lorg/apache/hadoop/hbase/HTableDescriptor;[[B)V

Change	Result
Return value type has been changed from void to java.util.concurrent.Future<java.lang.Void>.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

[#] Admin.disableTableAsync (TableName p1) [abstract] : void 1

org/apache/hadoop/hbase/client/Admin.disableTableAsync:(Lorg/apache/hadoop/hbase/TableName;)V

Change	Result
Return value type has been changed from void to java.util.concurrent.Future<java.lang.Void>.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

Admin.enableTableAsync (TableName p1) [abstract] : void 1

org/apache/hadoop/hbase/client/Admin.enableTableAsync:(Lorg/apache/hadoop/hbase/TableName;)V

Change	Result
Return value type has been changed from void to java.util.concurrent.Future<java.lang.Void>.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

Admin.enableTableAsync (TableName p1) [abstract] : void 1

org/apache/hadoop/hbase/client/Admin.getCompactionState:(Lorg/apache/hadoop/hbase/TableName;)Lorg/apache/hadoop/hbase/protobuf/generated/AdminProtos\$GetRegionInfoResponse\$CompactionState;

Change	Result
Return value type has been changed from org.apache.hadoop.hbase.protobuf.generated.AdminProtos.GetRegionInfoResponse.CompactionState to CompactionState.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

[#] Admin.getCompactionStateForRegion (byte[] p1) [abstract] : AdminProtos.GetRegionInfoResponse.CompactionState 1

org/apache/hadoop/hbase/client/Admin.getCompactionStateForRegion:([B)Lorg/apache/hadoop/hbase/protobuf/generated/AdminProtos\$GetRegionInfoResponse\$CompactionState;

Change	Result
Return value type has been changed from org.apache.hadoop.hbase.protobuf.generated.AdminProtos.GetRegionInfoResponse.CompactionState to CompactionState.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

HTableDescriptor and HColumnDescriptor changes

HTableDescriptor and HColumnDescriptor has become interfaces and you can create it through Builders. HCD has become CFD. It no longer implements writable interface. package org.apache.hadoop.hbase.

[#] class HColumnDescriptor (1)

Change	Result
Removed super-interface org.apache.hadoop.io.WritableComparable<HColumnDescriptor>.	A client program may be interrupted by NoSuchMethodError exception.

class HTableDescriptor (3)

Change	Result
Removed super-interface org.apache.hadoop.io.WritableComparable<HTableDescriptor>.	A client program may be interrupted by NoSuchMethodError exception.
Field META_TABLEDESC of type HTableDescriptor has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.

[#] HTableDescriptor.getColumnFamilies () : HColumnDescriptor[] (1)

org/apache/hadoop/hbase/HTableDescriptor.getColumnFamilies():()[Lorg/apache/hadoop/hbase/HColumnDescriptor;

[#] class HColumnDescriptor (1)

Change	Result
Return value type has been changed from HColumnDescriptor[] to client.ColumnFamilyDescriptor[].	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

[#] interface Table (4)

Change	Result
Abstract method batch (List<?>) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method batchCallback (List<?>, Batch.Callback<R>) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method getWriteBufferSize () has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method setWriteBufferSize (long) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

Deprecated buffer methods

- LockTimeoutException and OperationConflictException classes have been removed.

class OperationConflictException (1)

Result	Result
This class has been removed.	A client program may be interrupted by NoClassDefFoundError exception.

class LockTimeoutException (1)

Change Result This class has been removed. A client program may be interrupted by NoClassDefFoundError exception.

Filter API changes

Following methods have been removed: package org.apache.hadoop.hbase.filter

[#] class Filter (2)

Result	Result
Abstract method getNextKeyHint (KeyValue) has been removed from this class.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method transform (KeyValue) has been removed from this class.	A client program may be interrupted by NoSuchMethodError exception.

- HBASE-12296: Filters should work with ByteBufferedCell.
- HConnection is removed in Cloudera Runtime.
- RegionLoad and ServerLoad internally moved to shaded Protocol Buffers.

[#] class RegionLoad (1)

Result	Result
Type of field regionLoadPB has been changed from protobuf.generated.ClusterStatusProtos.RegionLoad to shaded.protobuf.generated.ClusterStatusProtos.RegionLoad.	A client program may be interrupted by NoSuchFieldError exception.

[#] interface AccessControlConstants (3)

Result	Result
Field OP_ATTRIBUTE_ACL_STRATEGY of type java.lang.String has been removed from this interface.	A client program may be interrupted by NoSuchFieldError exception.
Field OP_ATTRIBUTE_ACL_STRATEGY_CELL_FIRST of type byte[] has been removed from this interface.	A client program may be interrupted by NoSuchFieldError exception.
Field OP_ATTRIBUTE_ACL_STRATEGY_DEFAULT of type byte[] has been removed from this interface.	A client program may be interrupted by NoSuchFieldError exception.

[#] ServerLoad.getNumberOfRequests () : int 1

org/apache/hadoop/hbase/ServerLoad.getNumberOfRequests():I

Result	Result
Return value type has been changed from int to long.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

[#] ServerLoad.getNumberOfRequests () : int 1

org/apache/hadoop/hbase/ServerLoad.getReadRequestsCount():I

Result	Result
Return value type has been changed from int to long.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

[#] ServerLoad.getTotalNumberOfRequests () : int 1

org/apache/hadoop/hbase/ServerLoad.getTotalNumberOfRequests():I

Result	Result
Return value type has been changed from int to long.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

[#]ServerLoad.getWriteRequestsCount () : int 1

org/apache/hadoop/hbase/ServerLoad.getWriteRequestsCount():I

Result	Result
Return value type has been changed from int to long.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

[#]class HConstants (6)

Result	Result
Field DEFAULT_HBASE_CONFIG_READ_ZOOKEEPER_CONFIG of type boolean has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field HBASE_CONFIG_READ_ZOOKEEPER_CONFIG of type java.lang.String has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field REPLICATION_ENABLE_DEFAULT of type boolean has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field REPLICATION_ENABLE_KEY of type java.lang.String has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.

Field ZOOKEEPER_CONFIG_NAME of type java.lang.String has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.
Field ZOOKEEPER_USEMULTIof type java.lang.String has been removed from this class.	A client program may be interrupted by NoSuchFieldError exception.

HBASE-18732: [compat 1-2] HBASE-14047 removed Cell methods without deprecation cycle.

[#]interface Cell 5

Result	Result
Abstract method getFamily () has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method getMvccVersion () has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method getQualifier () has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method getRow () has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.
Abstract method getValue () has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

HBASE-18795:Expose KeyValue.getBuffer() for tests alone. Allows KV#getBuffer in tests only that was deprecated previously.

Region scanner changes

[#]interface RegionScanner (1)

Result	Result
Abstract method boolean nextRow (List<Cell>, int) has been removed from this interface.	A client program may be interrupted by NoSuchMethodError exception.

StoreFile changes

[#] class StoreFile (1)

Result	Result
This class became interface.	A client program may be interrupted by IncompatibleClassChangeError or InstantiationException exception dependent on the usage of this class.

MapReduce changes

HFile*Format has been removed.

ClusterStatus changes

[#] ClusterStatus.getRegionsInTransition () : Map<String,RegionState> 1

org/apache/hadoop/hbase/ClusterStatus.getRegionsInTransition:()Ljava/util/Map;

Result	Result
Return value type has been changed from java.util.Map<java.lang.String,master.RegionState> to java.util.List<master.RegionState>.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchMethodError exception.

Other changes in ClusterStatus include removal of convert methods that were no longer necessary after purge of Protocol Buffers from API.

Purge of Protocol Buffers from API

Protocol Buffers (PB) has been deprecated in APIs.

[#] HBaseSnapshotException.getSnapshotDescription () : HBaseProtos.SnapshotDescription 1

org/apache/hadoop/hbase/snapshot/HBaseSnapshotException.getSnapshotDescription:()Lorg/apache/hadoop/hbase/protobuf/generated/HBaseProtos\$SnapshotDescription;

Result	Result
Return value type has been changed from org.apache.hadoop.hbase.protobuf.generated.HBaseProtos.SnapshotDescription to org.apache.hadoop.hbase.client.SnapshotDescription.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchElementException.

HBASE-15609: Remove PB references from Result, DoubleColumnInterpreter and any such public facing class for 2.0. hbase-client-1.0.0.jar, Result.class package org.apache.hadoop.hbase.client

[#] Result.getStats () : ClientProtos.RegionLoadStats 1

org/apache/hadoop/hbase/client/Result.getStats:()Lorg/apache/hadoop/hbase/protobuf/generated/ClientProtos\$RegionLoadStats;

Result	Result
Return value type has been changed from org.apache.hadoop.hbase.protobuf.generated.ClientProtos.RegionLoadStats to RegionLoadStats.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchElementException.

PrettyPrinter changes

hbase-server-1.0.0.jar, HFilePrettyPrinter.class package org.apache.hadoop.hbase.io.hfile

Result	Result
Return value type has been changed from void to int.	This method has been removed because the return type is part of the method signature. A client program may be interrupted by NoSuchElementException.

Deprecation notices in Apache Impala

This topic lists the features and functionality in Apache Impala that will be deprecated or removed in this release or a future release.

Deprecated support to LZ0

In CDP, we deprecated support to LZ0 in Impala. So the impala-lzo plugin is not shipped as part of GPL Extras parcel.

Fixed Issues In Cloudera Runtime 7.2.0

Fixed issues represent issues reported by Cloudera customers that are addressed in this release.

Fixed Issues in Atlas

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

CDPD-13199: Incorrect attribute values in bulk import

When importing Business Metadata attribute assignments, Atlas used only the last assigned attribute value instead of individual values for each entity in the import list.

CDPD-372: All Spark Queries from the Same Spark Session were included in a Single Atlas Process

A Spark session can include multiple queries. When Atlas reports the Spark metadata, it creates a single process entity to correspond to the Spark session. The result was that an Atlas lineage picture showed multiple input entities or multiple output entities for a process, but the inputs and outputs were only related by the fact that they were included in operations in the same Spark session. In this release, the Spark Atlas Connector produces a `spark_application` entity for each Spark job. Each data flow produced by the job creates a `spark_process` entity in Atlas, which tracks the actual input and output data sets for that process. For more information, see [Spark metadata collection](#).

Fixed issues in DAS

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

There are no fixed issues in this release of Cloudera Runtime.

Fixed Issues in Hadoop

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

CDPD-7383: The ABFS and ADL connectors compatible with Alpine Linux and other platforms which have libssl1.1-1.1.1b-r1 as their native OpenSSL implementation

See [HADOOP-16460](#) and [HADOOP-16438](#)

Fixed Issues in HBase

There are no fixed issues for <component> in Cloudera Runtime <version number>.

Fixed Issues in HDFS

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

CDPD-6100

: This improvement makes HDFS NameNode leave safemode immediately if all blocks have reported in, reducing the cluster startup time.

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.

Fixed Issues in Hive

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

There are no fixed issues in this release of Cloudera Runtime.

Fixed Issues in Hue

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

There are no fixed issues in this release of Cloudera Runtime.

Fixed Issues in Impala

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

CDPD-10444: Update the version of Atlas used by Impala

Resolves the inconsistency between the versions of jackson-databind used by Atlas and Impala so that you no longer see the exception `NoClassDefFoundError` after the call to `QueryEventHookManager#executeQueryCompleteHooks()` in Impala.

Fixed Issues in Kafka

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

Bug ID: CDPD-8546

Apache JIRA: N/A

Apache Component: Kafka

Summary: Repeated ZooKeeper client log trace in Kafka server logs

Fixed Issues in Kudu

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

There are no fixed issues in this release of Cloudera Runtime.

Fixed Issues in Oozie

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

CDPD-12742: Oozie was not able to communicate with ID Broker and hence it failed to obtain a delegation token, because of a missing Jar

That Jar is now deployed together with Oozie and hence the underlying issue is fixed.

CDPD-12283: <https://issues.apache.org/jira/browse/OOZIE-3529>: By Oozie did not allow to use s3a and abfs file systems and users had to manually specify the supportability of these via Safety Valve

Since Oozie is compatible with these filesystems we changed the default Oozie configuration to allow these so users don't have to manually specify it.

CDPD-10746: Fix to address CVE-2019-17571

CDPD-9895: <https://issues.apache.org/jira/browse/OOZIE-3529>: Various errors when trying to use an S3 filesystem

Oozie is now fully compatible with S3.

CDPD-9761: <https://issues.apache.org/jira/browse/OOZIE-3584>: There is a sub workflow run in independent mode that runs a fork action which contains two (or more) actions

These actions inside the fork action run in parallel mode, and they have some seconds delay in between them. If a parameter is passed to one of these actions, that cannot be resolved, then it changes its status to `FAILED`, and also the workflow's state to `FAILED`. The other actions' state which are not started yet will be stuck in `PREP` state forever. The correct behaviour would be to `KILL` the remaining actions as well as the workflow. Note: this bug only occurs when it is run in independent mode. If it has a parent workflow, then the parent workflow will kill this workflow after 10 minutes because of the callback process.

CDPD-9721: Upgrade built-in spark-hive in Oozie

Oozie is using the Spark-Hive library from the stack.

CDPD-9220: <https://issues.apache.org/jira/browse/OOZIE-3586>: Oozie spark actions using --keytab fail due to duplicate dist. cache

Oozie spark actions add everything in the distributed cache of the launcher job to the distributed cache of the spark job, meaning the keytab is already there, then the --keytab argument tries to add it again causing the failure.

CDPD-9189: Apache Pig support was completely removed from Oozie

CDPD-7108: <https://issues.apache.org/jira/browse/OOZIE-3561>: In case we have a workflow which has, lets say, 80 actions after each other, then the validator code "never" finishes

CDPD-7107: <https://issues.apache.org/jira/browse/OOZIE-3551>: The following were added to the spark opts section of the spark action: --conf spark

CDPD-7106: <https://issues.apache.org/jira/browse/OOZIE-2828>: query tag is not functional for Hive2 action node in oozie

Workflow is intended to create a hive table using Hive2 action node. Though workflow run successfully, table is not created.

CDPD-7105: <https://issues.apache.org/jira/browse/OOZIE-2626>: Oozie workflow processing becomes slow after the increase of rows in WF_JOBS and WF_ACTIONS tables when running against SQL Server

CDPD-6877: <https://issues.apache.org/jira/browse/OOZIE-3578>: When you create a MapReduce action which then creates more than 120 counters, an exception was thrown

CDPD-6630: <https://issues.apache.org/jira/browse/OOZIE-3575>: Oozie by default gathers delegation tokens for the nodes defined in MapReduce

CDPD-5168: <https://issues.apache.org/jira/browse/OOZIE-3381>: Logging enhancements in CoordElFunctions for better supportability

CDPD-4826: Oozies web server does not work when TLS is enabled and Open JDK 11 is in use

This issue is now fixed.

Fixed Issues in Phoenix

There are no fixed issues for <component> in Cloudera Runtime <version number>.

Fixed Issues in Ranger

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

CDPD-12848: When you try to create multiple policies using the API having same non-existing group, the group creation fails due to multiple threads trying to create the same group at once.

Separate threads are now created for retry group creation and checks if the group is previously created and associate it with policy.

CDPD-10072: Ranger Ozone plugin unable to write to solr audits in SSL enabled cluster

This issue is now resolved. A separate folder libext is added under the Ozone library path and all the ranger plugin jars are added under this new folder.

Fixed Issues in Schema Registry

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

There are no fixed issues in this release of Cloudera Runtime.

Fixed Issues in Spark

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

CDPD-10532: Update log4j to address CVE-2019-17571

Replaced log4j with an internal version to fix CVE-2019-17571.

CDPD-10515: Incorrect version of jackson-mapper-asl

Use an internal version of jackson-mapper-asl to address CVE-2017-7525.

CDPD-7882: If an insert statement specifies partitions both statically and dynamically, there is a potential for data loss

To prevent data loss, this fix throws an exception if partitions are specified both statically and dynamically. You can follow the workarounds provided in the error message.

Fixed Issues in Streams Messaging Manager

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

There are no fixed issues in this release of Cloudera Runtime.

Fixed Issues in Zeppelin

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

CDPD-10187: Incorrect version of jackson-mapper-asl.

Use internal version of jackson-mapper-asl to handle CVE-2017-7525.

CDPD-1683: Zeppelin demo users have been removed

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-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.

Workaround: None.

Known Issues In Cloudera Runtime 7.2.0

Known Issues in Apache Atlas

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

CDPD-12668: Navigator Spark lineage can fail to render in Atlas

As part of content conversion from Navigator to Atlas, the conversion of some spark applications created a cyclic lineage reference in Atlas, which the Atlas UI fails to render. The cases occur when a Spark application uses data from a table and updates the same table.

Workaround: None.

CDPD-12620: Migration progress bar not refreshed

During the import stage of Cloudera Navigator to Apache Atlas migration, the migration progress bar does not correctly refresh the migration status. The Statistics page in the Atlas UI displays the correct details of the migration.

Workaround: None.

CDPD-11941: Table creation events missed when multiple tables are created in the same Hive command

When multiple Hive tables are created in the same database in a single command, the Atlas audit log for the database may not capture all the table creation events. When there is a delay between creation commands, audits are created as expected.

Workaround: None.

CDPD-11940: Database audit record misses table delete

When a hive_table entity is created, the Atlas audit list for the parent database includes an update audit. However, at this time, the database does not show an audit when the table is deleted.

Workaround: None.

CDPD-11790: Simultaneous events on the Kafka topic queue can produce duplicate Atlas entities

In normal operation, Atlas receives metadata to create entities from multiple services on the same or separate Kafka topics. In some instances, such as for Spark jobs, metadata to create a table entity in Atlas is triggered from two separate messages: one for the Spark operation and a second for the table metadata from HMS. If the process metadata arrives before the table metadata, Atlas creates a temporary entity for any tables that are not already in Atlas and reconciles the temporary entity with the HMS metadata when the table metadata arrives.

However, in some cases such as when Spark SQL queries with the write.saveAsTable function, Atlas does not reconcile the temporary and final table metadata, resulting in two entities with the same qualified name and no lineage linking the table to the process entity.

This issue is not seen for other lineage queries from spark:

```
create table default.xx3 as select * from default.xx2
insert into yy2 select * from yy
insert overwrite table ww2 select * from ww1
```

Another case where this behavior may occur is when many REST API requests are sent at the same time.

Workaround: None.

CDPD-11692: Navigator table creation time not converted to Atlas

In converting content from Navigator to Atlas, the create time for Hive tables is not moved to Atlas.

Workaround: None.

CDPD-11338: Cluster names with upper case letters may appear in lower case in some process names

Atlas records the cluster name as lower case in qualifiedNames for some process names. The result is that the cluster name may appear in lower case for some processes (insert overwrite table) while it appears in upper case for other queries (ctas) performed on the same cluster.

Workaround: None.

CDPD-10576: Deleted Business Metadata attributes appear in Search Suggestions

Atlas search suggestions continue to show Business Metadata attributes even if the attributes have been deleted.

Workaround: None.

CDPD-10574: Suggestion order doesn't match search weights

At this time, the order of search suggestions does not honor the search weight for attributes.

Workaround: None.

CDPD-10151: Short Spark job processes may be lost

In rare occasions, it is possible for events captured for Atlas by the Spark Atlas Connector to be dropped before the metadata reaches Atlas. It is more likely that an event is lost in very short running jobs.

Workaround: None.

CDPD-9095: Duplicate audits for renaming Hive tables

Renaming a Hive table results in duplicate ENTITY_UPDATE events in the corresponding Atlas entity audits, both for the table and for its columns.

Workaround: None.

CDPD-7982: HBase bridge stops at HBase table with deleted column family

Bridge importing metadata from HBase fails when it encounters an HBase table for which a column family was previously dropped. The error indicates:

```
Metadata service API org.apache.atlas.AtlasClientV2$API_V2@58112bc4 failed with status 404 (Not Found) Response Body
({ "errorCode": "ATLAS-404-00-007", "errorMessage": "Invalid instance creation/updation parameters passed : hbase_column_family.table: mandatory attribute value missing in type hbase_column_family" })
```

Workaround: None.

CDPD-7781: TLS certificates not validated on Firefox

Atlas is not checking for valid TLS certificates when the UI is opened in FireFox browsers.

Workaround: None.

CDPD-6675: Irregular qualifiedName format for Azure storage

The qualifiedName for hdfs_path entities created from Azure blob locations (ABFS) doesn't have the clusterName appended to it as do hdfs_path entities in other location types.

Workaround: None.

CDPD-6042: Hive Default Database Location Incorrect in Atlas Metadata

The location of the default Hive database as reported through the HMS-Atlas plugin does not match the actual location of the database. This problem does not affect non-default databases.

Workaround: None.

CDPD-5933, CDPD-5931: Unexpected Search Results When Using Regular Expressions in Basic Searches on Classifications

When you include a regular expression or wildcard in the search criteria for a classification in the Basic Search, the results may differ unexpectedly from when full classification names are included. For example, the Exclude sub-classifications option is respected when using a full classification name as the search criteria; when using part of the classification name and the wildcard (*) with Exclude sub-classifications turned off, entities marked with sub-classifications are not included in the results. Other instances of unexpected results include case-sensitivity.

Workaround: None.

CDPD-4762: 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.

CDPD-4662: Lineage graph links not working

Atlas lineage graphs do not include hyperlinks from assets to the assets' detail pages and clicking an asset does not provide an error in the log. Clicking an edge in a graph still provides access to edge behavior options such as controlling how classifications propagate.

Workaround: None.

CDPD-4545: 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.

CDPD-3700: 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.

Additional Cloudera JIRAs: CDP-5027, CDPD-3700, IMPALA-9070

CDPD-3208: 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.

CDPD-3160: 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.

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 access to Atlas after logging out, close all browser windows and exit the browser.

CDPD-1892: 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.

CDPD-1884: 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.

CDPD-1823: 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.

CDPD-1664: 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.

CDPD-922: IsUnique relationship attribute not honored

The Atlas model includes the ability to ensure that an attribute can be set to a specific value in only one relationship entity across the cluster metadata. For example, if you wanted to add metadata tags to relationships that you wanted to make sure were unique in the system, you could design the relationship attribute with the property "IsUnique" equal true. However, in this release, the IsUnique attribute is not enforced.

Workaround: None.

Known Issues in DAS

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

- You may not be able to add or delete columns or change the table schema after creating a new table using the upload table feature.
- For clusters secured using Knox, you see the HTTP 401: Forbidden error message when you click the DAS quick link from Cloudera Manager and are unable to log into DAS.

Workaround: The admin user will need to provide the DAS URL from the Knox proxy topology to the users needing access to DAS.

- The download logs feature may not return the YARN application logs on a Kerberized cluster. When you download the logs, the logs contain an error-reports.json file which states that no valid Kerberos tokens are available.

Workaround: An admin user with access to the machine can use the kinit command as a hive user with hive service user keytabs and trigger the download.

- The task logs for a particular task may not be available in the task swimlane. And the zip file generated by download logs artifact may not have task logs, but instead contain an error-reports.json file with the error log of the download failures.
- 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).
- 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, DAS displays an "HTTP 404 Not Found" error.
- Join count does not get updated for tables with partitioned columns.

Known Issues in Apache HBase

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

HBASE-24885: If an operator uses HBCK2 to invoke multiple `assigns` operations against one Region or happens to invoke HBCK2 `assigns` while HBase is re-assigning a Region, it is possible that the Region will be abnormally assigned. For example, unassigned, stuck in transition, and doubly-assigned.

Obtain a fix for this issue. Operators should definitely not schedule multiple assigns for a single Region at the same time, however there is still a potential race condition.

OpDB Data Hub cluster fails to initialize if you are reusing a cloud storage location that was used by an older OpDB Data Hub cluster

Workaround: Stop HBase using Cloudera Manager before deleting an operational database Data Hub cluster.

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

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.

OPSAPS-55788: WebHDFS is always enabled. The Enable WebHDFS checkbox does not take effect.

None.

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](#))

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.

OPSAPS-54299 Installing Hive on Tez and HMS in the incorrect order causes HiveServer failure

You need to install Hive on Tez and HMS in the correct order; otherwise, HiveServer fails. You need to install additional HiveServer roles to Hive on Tez, not the Hive service; otherwise, HiveServer fails.

Workaround: Follow instructions on [Installing Hive on Tez](#).

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"."IDX$", CONSTRAINT "
IDX$_FK1" FOREIGN KEY ("ORIG_TBL_ID") REFERENCES "TBL$ ("TBL_ID"
))
```

There are two workarounds:

- Drop the foreign key "IDX\$_FK1" on the "IDX\$" table within the metastore. You can also manually drop indexes, but do not cascade any drops because the IDX\$ table includes references to "TBL\$".
- Launch an older version of Hive, such as Hive 2.3 that includes IDX\$ in the DDL, and then drop the indexes as described in [Language Manual Indexing](#).

Apache Issue: [Hive-24815](#)

CDPD-13636: Hive job fails with OutOfMemory exception in the Azure DE cluster

Workaround: Set the parameter hive.optimize.sort.dynamic.partition.threshold=0. Add this parameter in Cloudera Manager (Hive Service Advanced Configuration Snippet (Safety Valve) for hive-site.xml)

ENGESC-2214: Hiveserver2 and HMS service logs are not deleted

Update Hive log4j configurations. Hive -> Configuration -> HiveServer2 Logging Advanced Configuration Snippet (Safety Valve) Hive Metastore -> Configuration -> Hive Metastore Server Logging Advanced Configuration Snippet (Safety Valve) Add the following to the configurations: appender.DRFA.strategy.action.type=DELETE appender.DRFA.strategy.action.basepath=\${log.dir} appender.DRFA.strategy.action.maxdepth=1 appender.DRFA.strategy.action.PathConditions.glob=\${log.file}.* appender.DRFA.strategy.action.PathConditions.type=IfFileName appender.DRFA.strategy.action.PathConditions.nestedConditions.type=IfAccumulatedFileCount appender.DRFA.strategy.action.PathConditions.nestedConditions.exceeds=same value as appender.DRFA.strategy.max

HiveServer Web UI displays incorrect data

If you enabled auto-TLS for TLS encryption, the HiveServer2 Web UI does not display the correct data in the following tables: Active Sessions, Open Queries, Last Max n Closed Queries

CDPD-11890: Hive on Tez cannot run certain queries on tables stored in encryption zones

This problem occurs when the Hadoop Key Management Server (KMS) connection is SSL-encrypted and a self signed certificate is used. `SSLHandshakeException` might appear in Hive logs.

Workaround:

Use one of the workarounds:

- Install a self signed SSL certificate into cacerts file on all hosts.
- Copy `ssl-client.xml` to a directory that is available in all hosts. In Cloudera Manager, in Clusters Hive on Tez Configuration . In Hive Service Advanced Configuration Snippet for `hive-site.xml`, click +, and add the name `tez.aux.uris` and `valuepath-to-ssl-client.xml`.

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 2021-524: Intermittent data duplication if direct insert enabled

If direct insert is enabled, data is written directly to the final location with an attemptId. At the end of the insert operation, all data written before the final attempt should be deleted. However due to a bug in HIVE-21164, this does not happen.

Example: Data is written to the final location with attemptId=0, but this task fails. Hive tries the task again and writes data to the final location with attemptId=1. At the end of the insert, Hive should remove all the files with attemptId=0, but it does not.

Upstream JIRA

- [HIVE-21164](#)
- [HIVE-24322](#)

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: [TSB 2021-524: Intermittent data duplication if direct insert enabled](#)

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.

Downloading Impala query results containing special characters in CSV format fails with ASCII codec error

In CDP, Hue is compatible with Python 2.7.x, but the Tablib library for Hue has been upgraded from 0.10.x to 0.14.x, which is generally used with the Python 3 release. If you try to download Impala query results having special characters in the result set in a CSV format, then the download may fail with the ASCII unicode decode error.

To fix this issue, downgrade the Tablib library to 0.12.x.

1. SSH into the Hue server host.
2. Change directory to the following:

```
cd /opt/cloudera/parcels/CDH-7.x/lib/
```

3. Back up the hue directory:

```
cp -R hue hue_ordinal
```

4. Change to the hue directory:

```
cd hue
```

5. Install the Wheel package using pip:

```
./build/env/bin/pip install wheel
```

The Wheel package is used to avoid recompiling your software during every install.

6. Install the Python Setuptools package for Hue as follows:

```
./build/env/bin/pip setuptools==44.1.0
```

7. Install Tablib version 0.12.1 as follows:

```
./build/env/bin/pip install tablib==0.12.1
```

8. Go to Cloudera Manager and restart the Hue service.

CDPD-18959: Due to tablib upgrade for Hue in CDP from 0.10.x to 0.14.x release which is generally used for Python3 release, the CSV download has a special character in the dataset which fails with the error message "UnicodeDecodeError: 'ascii' codec can't decode byte 0xef in position 8: ordinal not in range(128)"

Workaround:

- cd /opt/cloudera/parcels/CDH/lib/
- Backup the hue folder cp -R hue hue_org
- cd hue
- Install python packages via pippip install wheel
- Install python packages for Hue
 - ./build/env/bin/pip install wheel.
 - ./build/env/bin/pip setuptools==44.1.0
 - ./build/env/bin/pip install tablib==0.12.1
- Restart Hue and check again

Impala SELECT table query fails with UTF-8 codec error

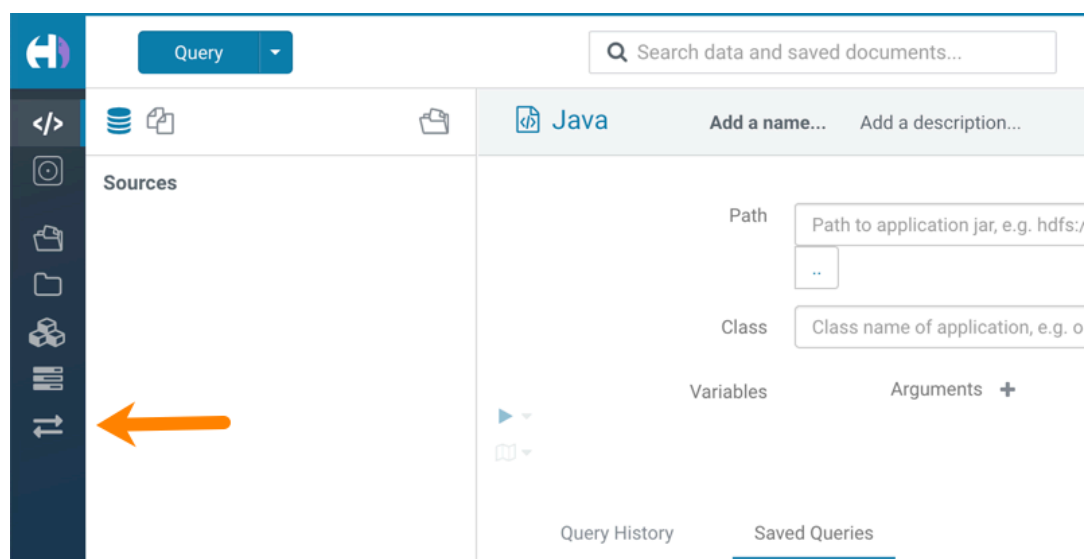
Hue cannot handle columns containing non-UTF8 data. As a result, you may see the following error while queying tables from the Impala editor in Hue: 'utf8' codec can't decode byte 0x91 in position 6: invalid start byte.

To resolve this issue, contact Cloudera Support to apply the following software patch: ENGESC-3457.

Hue Importer is not supported in the Data Engineering template

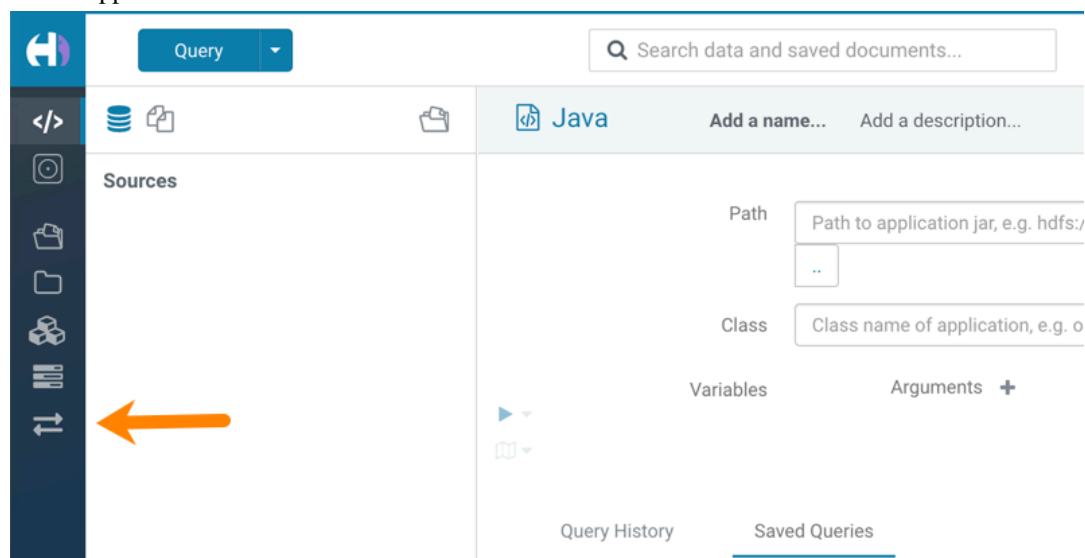
When you create a Data Hub cluster using the Data Engineering template, the Importer application is not supported in Hue.

Figure 1: Hue web UI showing Importer icon on the left assist panel



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:

CLUSTER: csl-190820-155926-zp3-dw2

STATUS: Running | NODES: 5 | CREATED AT: 08/20/19, 09:49 AM

Environment Details

NAME: csl-155927-dxy | DATA LAKE: csl-190820-155926-zp3-dd

Services: CM-UI, HUE, Name Node

Cloudera Manager Info

CM URL: <https://10.97.82.120:8443/csl-190820-155926-zp3-dw2/cdp-proxy/cm/home/>

Click this URL link.

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

Home

Switch to Classic View

Status | All Health Issues 1 | Configuration 8 | All Recent Commands

Compute Clusters | Last Updated: Aug 27, 9:19:50 PM UTC

Status	Name	Linked Data Context
Running	datamart	csl-190820-155926-zp3-dd (csl-190820-155926-zp3-dd (Proxy))

Other

Status	Name	Issues
--------	------	--------

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

The screenshot shows the Cloudera Manager interface. On the left is a dark sidebar with the 'CLUSTERA Manager' logo and a search bar. Below the search bar are navigation links: Clusters, Hosts, Diagnostics, and Charts. At the bottom of the sidebar are links for Recent Commands, Support, and CSO. The main panel on the right is titled 'datamart' and has tabs for Status, Health Issues, and Configuration. The Status tab is selected, showing a list of services: zookeeper, solr, kafka, ranger, atlas, Knox, and hive. Below this list is a section titled 'Compute Cluster, Cloudera Runtime 7.0.0 (Parcels)' which contains a list of services with their status and configuration links. An orange arrow points to the 'Hue' service entry, which has a status of 'OK' and a configuration link.

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.

BalancerMember worker hostname too long error

You may see the following error message while starting the Hue Load Balancer:

```
BalancerMember worker hostname (xxx-xxxxxxx-xxxxxxxxxxx-xxxxxxx
.xxxxxx-xxxxxx-xxxxxx.example.site) too long.
```

Cloudera Manager displays this error when you create a Data Hub cluster using the Data Engineering template and the Hue Load Balancer worker node name has exceeded 64 characters. In a CDP Public Cloud deployment, the system automatically generates the Load Balancer worker node name through AWS or Azure.

For example, if you specify `cdp-123456-scalecluster` as the cluster name, CDP creates `cdp-123456-scalecluster-master2.repro-aw.a123-4a5b.example.site` as the worker node name.

Workaround: Specify a shorter cluster name while creating a Data Hub cluster so that the final worker node name does not cross 64 characters.

For example, `cdp-123456-scale`.

Hue limitation after upgrading from CDH to CDP Data Center

The `hive.server2.parallel.ops.in.session` configuration property changes from `TRUE` to `FALSE` after upgrading from CDH to CDP Data Center. Current versions of Hue are compatible with this property change; however, if you still would like to use an earlier version of Hue that was not compatible with this property being `FALSE` and shared a single JDBC connection to issue queries concurrently, the connection will no longer work after upgrading.

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.

Queries stuck on failed HDFS calls and not timing out

In Impala 3.2 and higher, if the following error appears multiple times in a short duration while running a query, it would mean that the connection between the `impalad` and the HDFS `NameNode` is in a bad state.

```
"hdfsOpenFile() for <filename> at backend <hostname:port> failed
to finish before the <hdfs_operation_timeout_sec> second timeout
"
```

In Impala 3.1 and lower, the same issue would cause Impala to wait for a long time or not respond without showing the above error message.

Workaround: Restart the `impalad`.

Apache JIRA: [HADOOP-15720](#)

Impala should tolerate bad locale settings

If the `LC_*` environment variables specify an unsupported locale, Impala does not start.

Workaround: Add `LC_ALL="C"` to the environment settings for both the Impala daemon and the Statestore daemon.

Apache JIRA: IMPALA-532

Configuration to prevent crashes caused by thread resource limits

Impala could encounter a serious error due to resource usage under very high concurrency. The error message is similar to:

```
F0629 08:20:02.956413 29088 llvm-codegen.cc:111] LLVM hit fatal
error: Unable to allocate section memory!
terminate called after throwing an instance of 'boost::exception_
detail::clone_impl<boost::exception_detail::error_info_injector<
boost::thread_resource_error> >'
```

Workaround: To prevent such errors, configure each host running an `impalad` daemon with the following settings:

```
echo 2000000 > /proc/sys/kernel/threads-max
echo 2000000 > /proc/sys/kernel/pid_max
echo 8000000 > /proc/sys/vm/max_map_count
```

Add the following lines in `/etc/security/limits.conf`:

```
impala soft nproc 262144
impala hard nproc 262144
```

Apache JIRA: IMPALA-5605

Avro Scanner fails to parse some schemas

The default value in Avro schema must match type of first union type, e.g. if the default value is null, then the first type in the UNION must be "null".

Workaround: Swap the order of the fields in the schema specification. For example, use `["null", "string"]` instead of `["string", "null"]`. Note that the files written with the problematic schema must be rewritten with the new schema because Avro files have embedded schemas.

Apache JIRA: IMPALA-635

Process mem limit does not account for the JVM's memory usage

Some memory allocated by the JVM used internally by Impala is not counted against the memory limit for the `impalad` daemon.

Workaround: To monitor overall memory usage, use the `top` command, or add the memory figures in the Impala web UI `/memz` tab to JVM memory usage shown on the `/metrics` tab.

Apache JIRA: IMPALA-691

Ranger audit logs for applying column masking policies missing

Impala is not producing these logs.

Workaround: None.

Apache JIRA: IMPALA-9350

Impala BE cannot parse Avro schema that contains a trailing semi-colon

If an Avro table has a schema definition with a trailing semicolon, Impala encounters an error when the table is queried.

Workaround: Remove trailing semicolon from the Avro schema.

Apache JIRA: IMPALA-1024

Incorrect results with basic predicate on CHAR typed column

When comparing a CHAR column value to a string literal, the literal value is not blank-padded and so the comparison might fail when it should match.

Workaround: Use the RPAD() function to blank-pad literals compared with CHAR columns to the expected length.

Apache JIRA: IMPALA-1652

ImpalaODBC: Can not get the value in the SQLGetData(m-x th column) after the SQLBindCol(m th column)

If the ODBC `SQLGetData` is called on a series of columns, the function calls must follow the same order as the columns. For example, if data is fetched from column 2 then column 1, the `SQLGetData` call for column 1 returns `NULL`.

Workaround: Fetch columns in the same order they are defined in the table.

Apache JIRA: IMPALA-1792

Casting scenarios with invalid/inconsistent results

Using a `CAST()` function to convert large literal values to smaller types, or to convert special values such as NaN or Inf, produces values not consistent with other database systems. This could lead to unexpected results from queries.

Apache JIRA: IMPALA-1821

A failed CTAS does not drop the table if the insert fails

If a `CREATE TABLE AS SELECT` operation successfully creates the target table but an error occurs while querying the source table or copying the data, the new table is left behind rather than being dropped.

Workaround: Drop the new table manually after a failed CREATE TABLE AS SELECT

Apache JIRA: IMPALA-2005

% escaping does not work correctly when occurs at the end in a LIKE clause

If the final character in the RHS argument of a LIKE operator is an escaped \% character, it does not match a % final character of the LHS argument.

Apache JIRA: IMPALA-2422

Crash: impala::Coordinator::ValidateCollectionSlots

A query could encounter a serious error if includes multiple nested levels of INNER JOIN clauses involving subqueries.

Apache JIRA: IMPALA-2603

Incorrect result due to constant evaluation in query with outer join

Workaround: An OUTER JOIN query could omit some expected result rows due to a constant such as FALSE in another join clause. For example:

[illegible]

[illegible]

Apache JIRA: IMPALA-3094

Breakpad minidumps can be very large when the thread count is high

The size of the breakpoint minidump files grows linearly with the number of threads. By default, each thread adds 8 KB to the minidump size. Minidump files could consume significant disk space when the daemons have a high number of threads.

Workaround: Add `-\minidump_size_limit_hint_kb=size` to set a soft upper limit on the size of each minidump file. If the minidump file would exceed that limit, Impala reduces the amount of information for each thread from 8 KB to 2 KB. (Full thread information is captured for the first 20 threads, then 2 KB per thread after that.) The minidump file can still grow larger than the "hinted" size. For example, if you have 10,000 threads, the minidump file can be more than 20 MB.

Apache JIRA: IMPALA-3509

Impala requires FQDN from hostname command on Kerberized clusters

The method Impala uses to retrieve the host name while constructing the Kerberos principal is the `gethostname()` system call. This function might not always return the fully qualified domain name, depending on the network configuration. If the daemons cannot determine the FQDN, Impala does not start on a Kerberized cluster.

Workaround: Test if a host is affected by checking whether the output of the `hostname` command includes the FQDN. On hosts where `hostname`, only returns the short name, pass the command-line flag `##hostname=fully_qualified_domain_name` in the startup options of all Impala-related daemons.

Apache JIRA: IMPALA-4978

Metadata operations block read-only operations on unrelated tables

Metadata operations that change the state of a table, like `COMPUTE STATS` or `ALTER RECOVER PARTITIONS`, may delay metadata propagation of unrelated unloaded tables triggered by statements like `DESCRIBE` or `SELECT` queries.

Workaround:

Apache JIRA: IMPALA-6671

Impala does not support Heimdal Kerberos

Apache JIRA: IMPALA-7072

CDPD-28139: Set spark.hadoop.hive.stats.autogather to false by default

As an Impala user, if you submit a query against a table containing data ingested using Spark and you are concerned about the quality of the query plan, you must run `COMPUTE STATS` against such a table in any case after an ETL operation because numRows created by Spark could be incorrect. Also, use other stats computed by `COMPUTE STATS`, e.g., Number of Distinct Values (NDV) and NULL count for good selectivity estimates.

For example, when a user ingests data from a file into a partition of an existing table using Spark, if `spark.hadoop.hive.stats.autogather` is not set to `false` explicitly, `numRows` associated with this partition would be 0 even though there is at least one row in the file. To avoid this, the workaround

is to set "spark.hadoop.hive.stats.autogather=false" in the "Spark Client Advanced Configuration Snippet (Safety Valve) for spark-conf/spark-defaults.conf" in Spark's CM Configuration section.

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](#)

TSB 2021-479: Impala can return incomplete results through JDBC and ODBC clients in all CDP offerings

In CDP, we introduced a timeout on queries to Impala defaulting to 10 seconds. The timeout setting is called `FETCH_ROWS_TIMEOUT_MS`. Due to this setting, JDBC, ODBC, and Beeswax clients running Impala queries believe the data returned at 10 seconds is a complete dataset and present

it as the final output. However, in cases where there are still results to return after this timeout has passed, when the driver closes the connection, based on the timeout, it results in a scenario where the query results are incomplete.

Upstream JIRA

[IMPALA-7561](#)

TSB 2022-543: Impala query with predicate on analytic function may produce incorrect results

Apache Impala may produce incorrect results for a query which has all of the following conditions:

- There are two or more analytic functions (for example, `row_number()`) in an inline view
- Some of the functions have partition-by expression while the others do not
- There is a predicate on the inline view's output expression corresponding to the analytic function

Upstream JIRA

[IMPALA-11030](#)

Knowledge article

For the latest update on this issue, see the corresponding Knowledge article: [TSB 2022-543: Impala query with predicate on analytic function may produce incorrect results](#)

Known Issues in Apache Kafka

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

Known Issues

OPSAPS-59553: SMM's bootstrap server config should be updated based on Kafka's listeners

SMM does not show any metrics for Kafka or Kafka Connect when multiple listeners are set in Kafka.

Workaround: SMM cannot identify multiple listeners and still points to bootstrap server using the default broker port (9093 for SASL_SSL). You would have to override bootstrap server URL (hostname:port as set in the listeners for broker) in the following path:

Cloudera Manager > SMM > Configuration > Streams Messaging Manager Rest Admin Server Advanced Configuration Snippet (Safety Valve) for streams-messaging-manager.yaml > Save Changes > Restart SMM.

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.

OPSAPS-57907: The Kafka metric collector adapter generates high CPU load

If a large number of topic partitions are created on a cluster, the Cloudera Manager Agent can generate a high CPU load. This is caused by the Kafka metric collector adapter carrying out excessive regex matching.

Workaround: None.

OPSAPS-59031: Kafka cannot start if configuration is added to the Kafka Broker Advanced Configuration Snippet (Safety Valve) for ssl.properties

The Kafka Broker Advanced Configuration Snippet (Safety Valve) for ssl.properties configuration snippet does not correctly override configuration. As a result, Kafka may not start if TLS/SSL related configuration overrides are added to the this configuration snippet.

Workaround: Use the Kafka Broker Advanced Configuration Snippet (Safety Valve) for kafka.properties configuration snippet instead to override SSL related properties.

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.
- While Kafka Connect is available as part of Runtime, it is currently not supported in CDP Public Cloud. NiFi is a proven solution for batch and real time data loading that complement Kafka's message broker capability. For more information, see [Creating your first Flow Management cluster](#).
- The Kafka default authorizer is not supported. This includes setting ACLs and all related APIs, broker functionality, and command-line tools.

Limitations

Collection of Partition Level Metrics May Cause Cloudera Manager's Performance to Degrade

If the Kafka service operates with a large number of partitions, collection of partition level metrics may cause Cloudera Manager's performance to degrade.

If you are observing performance degradation and your cluster is operating with a high number of partitions, you can choose to disable the collection of partition level metrics.



Important: If you are using SMM to monitor Kafka or Cruise Control for rebalancing Kafka partitions, be aware that both SMM and Cruise Control rely on partition level metrics. If partition level metric collection is disabled, SMM will not be able to display information about partitions. In addition, Cruise Control will not operate properly.

Complete the following steps to turn off the collection of partition level metrics:

1. Obtain the Kafka service name:
 - a. In Cloudera Manager, Select the Kafka service.
 - b. Select any available chart, and select Open in Chart Builder from the configuration icon drop-down.
 - c. Find \$SERVICENAME= near the top of the display.

The Kafka service name is the value of \$SERVICENAME.

2. Turn off the collection of partition level metrics:
 - a. Go to Hosts Configuration.
 - b. Find and configure the Cloudera Manager Agent Monitoring Advanced Configuration Snippet (Safety Valve) configuration property.

Enter the following to turn off the collection of partition level metrics:

```
[KAFKA_SERVICE_NAME]_feature_send_broker_topic_partition_entity_update_enabled=false
```

Replace [KAFKA_SERVICE_NAME] with the service name of Kafka obtained in step 1. The service name should always be in lower case.

- c. Click Save Changes.

Known Issues in Kerberos

Learn about the known issues in Kerberos, the impact or changes to the functionality, and the workaround.

OPSAPS-60331: If Cloudera Manager is configured to use Active Directory as a Kerberos KDC, and is also configured to use /etc/cloudera-scm-server/cmf.keytab as the KDC admin credentials, you may encounter errors when generating Kerberos credentials.

In the Cloudera Manager Admin Console, run the "Administration > Security > Kerberos Credentials > Import KDC Account Manager Credentials" wizard. Remove `/etc/cloudera-scm-server/cmf.keytab` on the Cloudera Manager server host.

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 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.

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.

Technical Service Bulletins

TSB 2021-467: Race condition in Apache Oozie Sharelib upload

There is a race condition in Apache Oozie OozieSharelibCLI which allows a malicious attacker to replace the files in Oozie's sharelib during its creation. A race condition in OozieSharelibCLI allows an attacker to replace the contents of the sharelib.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: [TSB 2021-467: Race condition in Apache Oozie Sharelib upload](#)

Known Issues in Apache Phoenix

There are no known issues for <component> in Cloudera Runtime <version number>.

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.

CDPD-12644 Ranger Key Names cannot be reused with the Ranger KMS KTS service

Key names cannot be reused with the Ranger KMS KTS service. If the key name of a delete key is reused, the new key can be successfully created and used to create an encryption zone, but data cannot be written to that encryption zone.

Workaround:

Use only unique key names when creating keys.

Known Issues in Cloudera Search

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

Known Issues

Cloudera Bug ID:

CDPD-20577

Summary:

Splitshard of HDFS index checks local filesystem and fails

Description:

When performing a shard split on an index that is stored on HDFS, SplitShardCmd still evaluates free disk space on the local file system of the server where Solr is installed. This may cause the command to fail, perceiving that there is no adequate disk space to perform the shard split.

Workaround:

None

Cloudera Bug ID:

OPSAPS-58059

Summary:

Solr log rotation counts the number of retained log files daily instead of globally

Description:

With CDP 7.1.1, Search moved to Log4Jv2. This has affected Solr log rotation behavior in an unwanted way. With the default configuration, Solr log file names include a date and a running index, for example: solr-cmf-solr-SOLR_SERVER-solrserver-1.my.corporation.com.log.out.2020-08-31-9. The number of retained log files is configured in Cloudera Manager, however the configured number now applies for each day, instead of applying globally for all log files of the particular server.

Workaround:

Using Cloudera Manager, edit the Solr Server Logging Advanced Configuration Snippet (Safety Valve) property of your Solr service and add a new line containing: appender.DRFA.filePattern=\${log.dir}/\${log.file}.%i

Cloudera Bug ID:

DOCS-5717

Summary:

Lucene index handling limitation

Description:

The Lucene index can only be upgraded by one major version. Solr 8 will not open an index that was created with Solr 6 or earlier.

Workaround:

There is no workaround, you need to reindex collections.

Cloudera Bug ID:

CDH-82042

Summary:

Solr service with no added collections causes the upgrade process to fail

Description:

Upgrade fails while performing the bootstrap collections step of the solr-upgrade.sh script with the error message:

```
Failed to execute command Bootstrap Solr Collections on service Solr
```

if there are no collections present in Solr.

Workaround:

If there are no collections added to it, remove the Solr service from your cluster before you start the upgrade.

Cloudera Bug ID:

CDH-66345

Summary:

Solr SQL, Graph, and Stream Handlers are Disabled if Collection Uses Document-Level Security

Description:

The Solr SQL, Graph, and Stream handlers do not support document-level security, and are disabled if document-level security is enabled on the collection. If necessary, these handlers can be re-enabled by setting the following Java system properties, but document-level security is not enforced for these handlers:

- SQL: solr.sentry.enableSqlQuery=true
- Graph: solr.sentry.enableGraphQuery=true
- Stream: solr.sentry.enableStreams=true

Workaround:

None.

Cloudera Bug ID:

CDH-34050

Summary:

Collection Creation No Longer Supports Automatically Selecting A Configuration If Only One Exists

Description:

Before CDH 5.5.0, a collection could be created without specifying a configuration. If no -c value was specified, then:

- If there was only one configuration, that configuration was chosen.
- If the collection name matched a configuration name, that configuration was chosen.

Search now includes multiple built-in configurations. As a result, there is no longer a case in which only one configuration can be chosen by default.

Workaround:

Explicitly specify the collection configuration to use by passing -c <configName> to solrctl collection --create.

Cloudera Bug ID:

CDH-22190

Summary:

CrunchIndexerTool which includes Spark indexer requires specific input file format specifications

Description:

If the --input-file-format option is specified with CrunchIndexerTool, then its argument must be text, avro, or avroParquet, rather than a fully qualified class name.

Workaround:

None.

Cloudera Bug ID:

CDH-19923

Summary:

The quickstart.sh file does not validate ZooKeeper and the NameNode on some operating systems

Description:

The quickstart.sh file uses the timeout function to determine if ZooKeeper and the NameNode are available. To ensure this check can be complete as intended, the quickstart.sh determines if the operating system on which the script is running supports timeout. If the script detects that the operating system does not support timeout, the script continues without checking if the NameNode and ZooKeeper are available. If your environment is configured properly or you are using an operating system that supports timeout, this issue does not apply.

Workaround:

This issue only occurs in some operating systems. If timeout is not available, the quickstart continues and final validation is always done by the MapReduce jobs and Solr commands that are run by the quickstart.

Cloudera Bug ID:

CDH-26856

Summary:

Field value class guessing and Automatic schema field addition are not supported with the MapReduceIndexerTool nor with the HBaseMapReduceIndexerTool

Description:

The MapReduceIndexerTool and the HBaseMapReduceIndexerTool can be used with a Managed Schema created via NRT indexing of documents or via the Solr Schema API. However, neither tool supports adding fields automatically to the schema during ingest.

Workaround:

Define the schema before running the MapReduceIndexerTool or HBaseMapReduceIndexerTool. In non-schemaless mode, define in the schema using the schema.xml file. In schemaless mode, either define the schema using the Solr Schema API or index sample documents using NRT indexing before invoking the tools. In either case, Cloudera recommends that you verify that the schema is what you expect, using the List Fields API command.

Cloudera Bug ID:

CDH-19407

Summary:

The Browse and Spell Request Handlers are not enabled in schemaless mode

Description:

The Browse and Spell Request Handlers require certain fields to be present in the schema. Since those fields cannot be guaranteed to exist in a Schemaless setup, the Browse and Spell Request Handlers are not enabled by default.

Workaround:

If you require the Browse and Spell Request Handlers, add them to the solrconfig.xml configuration file. Generate a non-schemaless configuration to see the usual settings and modify the required fields to fit your schema.

Cloudera Bug ID:

CDH-17978

Summary:

Enabling blockcache writing may result in unusable indexes

Description:

It is possible to create indexes with solr.hdfs.blockcache.write.enabled set to true. Such indexes may appear corrupt to readers, and reading these indexes may irrecoverably corrupt indexes. Blockcache writing is disabled by default.

Workaround:

None.

Cloudera Bug ID:

CDH-58276

Summary:

Users with insufficient Solr permissions may receive a "Page Loading" message from the Solr Web Admin UI

Description:

Users who are not authorized to use the Solr Admin UI are not given a page explaining that access is denied to them, instead receive a web page that never finishes loading.

Workaround:

None.

Cloudera Bug ID:

CDH-15441

Summary:

Using MapReduceIndexerTool or HBaseMapReduceIndexerTool multiple times may produce duplicate entries in a collection

Description:

Repeatedly running the MapReduceIndexerTool on the same set of input files can result in duplicate entries in the Solr collection. This occurs because the tool can only insert documents and cannot update or delete existing Solr documents. This issue does not apply to the HBaseMapReduceIndexerTool unless it is run with more than zero reducers.

Workaround:

To avoid this issue, use HBaseMapReduceIndexerTool with zero reducers. This must be done without Kerberos.

Cloudera Bug ID:

CDH-58694

Summary:

Deleting collections might fail if hosts are unavailable

Description:

It is possible to delete a collection when hosts that host some of the collection are unavailable. After such a deletion, if the previously unavailable hosts are brought back online, the deleted collection may be restored.

Workaround:

Ensure all hosts are online before deleting collections.

Cloudera Bug ID:

CDPD-4139

Summary:

Collection state goes down after Solr SSL

Description:

If you enable TLS/SSL on a Solr instance with existing collections, the collections will break and become unavailable. Collections created after enabling TLS/SSL are not affected by this issue.

Workaround:

Recreate the collection after enabling TLS.

Cloudera Bug ID:

CDPD-13923

Summary:

Every Configset is Untrusted Without Kerberos

Description:

Solr 8 introduces the concept of ‘[untrusted configset](#)’, denoting configsets that were uploaded without authentication. Collections created with an untrusted configset will not initialize if `<lib>` directives are used in the configset.

Workaround:

Select one of the following options if you would like to use untrusted configsets with `<lib>` directives:

- If the configset contains external libraries, but you do not want to use them, simply upload the configsets after deleting the `<lib>` directives.
- If the configset contains external libraries, and you want to use them, choose one from the following options:
 - Secure your cluster before reuploading the configset.
 - Add the libraries to Solr’s classpath, then reupload the configset without the `<lib>` directives.

Unsupported Features

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

- [Package Management System](#)
- [HTTP/2](#)
- [Solr SQL/JDBC](#)
- [Graph Traversal](#)
- [Cross Data Center Replication \(CDCR\)](#)
- [SolrCloud Autoscaling](#)
- HDFS Federation
- Saving search results
- Solr contrib modules (Spark, MapReduce and Lily HBase indexers are not contrib modules but part of the Cloudera Search product itself, therefore they are supported).

Limitations**Default Solr core names cannot be changed**

Although it is technically possible to give user-defined Solr core names during core creation, it is to be avoided in the context of Cloudera Search. Cloudera Manager expects core names in the default "collection_shardX_replicaY" format. Altering core names results in Cloudera Manager being unable to fetch Solr metrics for the given core and this, eventually, may corrupt data collection for co-located core, or even shard and server level charts.

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 2022-535: Ranger audit retention settings in Solr are not honored**

The audits present in the `ranger_audits` collection in the Solr service of Data Lake do not get deleted based on the retention period set. The default retention period is 90 days.

This is caused by the incorrect order of processors in the configuration (solrconfig.xml) used by the ranger_audits collection.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: [TSB 2021-535: Ranger audit retention settings in Solr are not honored](#).

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-22670 and CDPD-23103: There are two configurations in Spark, "Atlas dependency" and "spark_lineage_enabled", which are conflicted. The issue is when Atlas dependency is turned off but spark_lineage_enabled is turned on.

Run Spark application, Spark will log some error message and cannot continue. That can be restored by correcting the configurations and restarting Spark component with distributing client configurations.

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.

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] ;
```

Workaround: Create the database using Hive or Impala, or specify the external data warehouse location in the create command. For example:

```
sql("create database spark_database location '/warehouse/tablesp
ace/external/hive/spark_database.db' ")
```

Technical Service Bulletins

TSB 2021-441: CDP Powered by Apache Spark may incorrectly read/write pre-Gregorian timestamps

Spark may incorrectly read or write `TIMESTAMP` data for values before the start of the Gregorian calendar ('1582-10-15 00:00:00.0'). This could happen when Spark is:

- Using dynamic partition inserts
- Reading or writing from an ORC table when the:
 - `spark.sql.hive.convertMetastoreOrc` property is set to false. Its default value is true.
 - `spark.sql.hive.convertMetastoreOrc` property is set to true but the `spark.sql.orc.impl` property is set to `hive`. Its default is `native`.
- Reading or writing from a Parquet table when the:
 - `spark.sql.hive.convertMetastoreParquet` property is set to false. Its default value is true.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: [TSB 2021-441: Spark may incorrectly read/write pre-Gregorian timestamps](#)

Known Issues for Apache Sqoop

Learn about the known issues in Sqoop, the impact or changes to the functionality, and the workaround.

Using direct mode causes problems

Using direct mode has several drawbacks:

- Imports can cause intermittent an overlapping input split.
- Imports can generate duplicate data.
- Many problems, such as intermittent failures, can occur.
- Additional configuration is required.

Stop using direct mode. Do not use the `--direct` option in Sqoop import or export commands.

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 Streams Messaging Manager

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

There are no known issues in this release of Cloudera Runtime.

OPSAPS-59553: SMM's bootstrap server config should be updated based on Kafka's listeners

SMM does not show any metrics for Kafka or Kafka Connect when multiple listeners are set in Kafka.

Workaround: SMM cannot identify multiple listeners and still points to bootstrap server using the default broker port (9093 for SASL_SSL). You would have to override bootstrap server URL (hostname:port as set in the listeners for broker). Add the bootstrap server details in SMM safety valve in the following path:

Cloudera Manager > SMM > Configuration > Streams Messaging Manager Rest Admin Server Advanced Configuration Snippet (Safety Valve) for streams-messaging-manager.yaml > Add the following value for bootstrap servers>Save Changes > Restart SMM.

```
streams.messaging.manager.kafka.bootstrap.servers=<comma-separated list of brokers>
```

OPSAPS-59828: SMM cannot connect to Schema Registry when TLS is enabled

When TLS is enabled, SMM by default cannot properly connect to Schema Registry.

As a result, when viewing topics in the SMM Data Explorer with the deserializer key or value set to Avro, the following error messages are shown:

- Error deserializing key/value for partition [***PARTITION***] at offset [***OFFSET***]. If needed, please seek past the record to continue consumption.
- Failed to fetch value schema versions for topic : '[***TOPIC***]'.

In addition, the following certificate error will also be present in the SMM log:

- javax.net.ssl.SSLHandshakeException: PKIX path building failed:...

Workaround: Additional security properties must be set for SMM.

1. In Cloudera Manager, select the SMM service.
2. Go to Configuration.
3. Find and configure the SMM_JMX_OPTS property.

Add the following JVM SSL properties:

- Djavax.net.ssl.trustStore=[***SMM TRUSTSTORE LOCATION***]
- Djavax.net.ssl.trustStorePassword=[***PASSWORD***]

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

OPSAPS-57067: Yarn Service in Cloudera Manager reports stale configuration yarn.cluster.scaling.recommendation.enable.

Workaround: This issue does not affect the functionality. Restarting Yarn service will fix this issue.

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-3181: Application logs does not work for AZURE and AWS cluster

Yarn Application Log Aggregation will fail for any YARN job (MR, Tez, Spark, etc) which do not use cloud storage, or use a cloud storage location other than the one configured for YARN logs (`yarn.nodemanager.remote-app-log-dir`).

Workaround: Configure the following:

- For MapReduce job, set `mapreduce.job.hdfs-servers` in the `mapred-site.xml` file with all filesystems required for the job including the one set in `yarn.nodemanager.remote-app-log-dir` such as `hdfs://nn1/`, `hdfs://nn2/`.
- For Spark job, set the job level with all filesystems required for the job including the one set in `yarn.nodemanager.remote-app-log-dir` such as `hdfs://nn1/`, `hdfs://nn2/` in `spark.yarn.access.hadoopFileSystems` and pass it through the `--config` option in `spark-submit`.
- For jobs submitted using the `hadoop` command, place a separate `core-site.xml` file with `fs.defaultFS` set to the filesystem set in `yarn.nodemanager.remote-app-log-dir` in a path. Add that directory path in `--config` when executing the `hadoop` command.

COMPX-1445: Queue Manager operations are failing when Queue Manager is installed separately from YARN

If Queue Manager is not selected during YARN installation, Queue Manager operation are failing. Queue Manager says 0 queues are configured and several failures are present. That is because ZooKeeper configuration store is not enabled.

Workaround:

1. In Cloudera Manager, select the YARN service.
2. Click the Configuration tab.
3. Find the Queue Manager Service property.
4. Select the Queue Manager service that the YARN service instance depends on.
5. Click Save Changes.
6. Restart all services that are marked stale in Cloudera Manager.

COMPX-1451: Queue Manager does not support multiple Resource

When YARN High Availability is enabled there are multiple Resource Managers. Queue Manager receives multiple `ResourceManager` URLs for a High Availability cluster. It picks the active `ResourceManager` URL only when Queue Manager page is loaded. Queue Manager cannot handle it gracefully when the currently active `ResourceManager` goes down while the user is still using the Queue Manager UI.

Workaround: Reload the Queue Manager page manually.

COMPX-3329: Autorestart is not enabled for Queue Manager in Data Hub

In a Data Hub cluster, Queue Manager is installed with autorestart disabled. Hence, if Queue Manager goes down, it will not restart automatically.

Workaround: If Queue Manager goes down in a Data Hub cluster, you must go to the Cloudera Manager Dashboard and restart the Queue Manager service.

Third party applications do not launch if MapReduce framework path is not included in the client configuration

MapReduce application framework is loaded from HDFS instead of being present on the NodeManagers. By default the `mapreduce.application.framework.path` property is set to the appropriate value, but third party applications with their own configurations will not launch.

Workaround: Set the `mapreduce.application.framework.path` property to the appropriate configuration for third party applications.

YARN cannot start if Kerberos principal name is changed

If the Kerberos principal name is changed in Cloudera Manager after launch, YARN will not be able to start. In such case the keytabs can be correctly generated but YARN cannot access ZooKeeper with the new Kerberos principal name and old ACLs.

There are two possible workarounds:

- Delete the znode and restart the YARN service.
- Use the reset ZK ACLs command. This also sets the znodes below /rmstore/ZKRMStateRoot to world:anyone:cdw which is less secure.

COMPX-8687: Missing access check for getAppAttempts

When the Job ACL feature is enabled using Cloudera Manager (YARN Configuration Enable 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.

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
- Rolling Log Aggregation
- Docker on YARN (DockerContainerExecutor) on Data Hub clusters
- 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-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.

Workaround: From the CDP Management Console, go to Cloudera Manager for the cluster running Zeppelin. On the Zeppelin configuration page (Zeppelin serviceConfiguration), enter shiro urls in the Search field, and then add the missing closing brace to the notebook-repositories URL, as follows:

```
/api/notebook-repositories/** = authc, roles[{{zeppelin_admin_group}}]
```

Click Save Changes, and restart the Zeppelin service.

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.

Known Issues in Apache ZooKeeper

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

Zookeeper-client does not use ZooKeeper TLS/SSL automatically

The command-line tool ‘zookeeper-client’ is installed to all Cloudera Nodes and it can be used to start the default Java command line ZooKeeper client. However even when ZooKeeper TLS/SSL is enabled, the zookeeper-client command connects to localhost:2181, without using TLS/SSL.

Workaround:

Manually configure the 2182 port, when zookeeper-client connects to a ZooKeeper cluster. The following is an example of connecting to a specific three-node ZooKeeper cluster using TLS/SSL:

```
CLIENT_JVMFLAGS="-Dzookeeper.clientCnxnSocket=org.apache.zookeeper.ClientCnxnSocketNetty -Dzookeeper.ssl.keyStore.location=<path to your configured keystore> -Dzookeeper.ssl.keyStore.password=<the password you configured for the keystore> -Dzookeeper.ssl.trustStore.location=<path to your configured truststore> -Dzookeeper.ssl.trustStore.password=<the password you configured for the truststore> -Dzookeeper.client.secure=true" zookeeper-client -server <your.zookeeper.server-1>:2182,<your.zookeeper.server-2>:2182,<your.zookeeper.server-3>:2182
```