Release Notes

Date published:
Date modified:

CLOUDƎRA

Legal Notice

Contents

What's New In Cloudera Runtime 7.0.1

What's New in Apache Atlas

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

Spark support

Atlas collects metadata from Spark 1 and Spark 2. Spark operations are represented by Atlas processes; when the same query is run more than once, Atlas creates process execution entities to capture the volume of activity. Atlas collects metadata through Hive Metastore (HMS) for the Hive data assets that Spark operates against.

Atlas Reference: Spark Metadata Collection

Free Text Search across all entity types

Atlas provides a free-text search box that matches search criteria across all entity types and all string attributes for all entities. The top 5 matching results appear as menu selections; all results are returned ranked by which attributes matched the search terms. The text search also offers suggestions, which are the entities with the most matches across the most important attributes, such as names and descriptions.

Using Free-Text Search

Process execution entities

Atlas collects metadata for operations that occur in Hive, Impala, and other query engines. In this release, each instance of a query is represented by a process execution entity. The process execution is related to the parent process entity. The result is that you see one lineage diagram for an operation, no matter how many times the same operation runs.

Process executions are listed in the relationship tab of the parent process; they are indicated by the query text with a system generated identifier appended to the end.

Search facet counts

The entries in the basic search drop-down lists now have counts to indicate how many of each classification or entity type exist in Atlas. These counts can help you narrow your searches quickly to reach the best results.

Atlas server and message statistics

Atlas collects statistics on the metadata it processes, such as the rate of messages received, which message was most recently processed, the count and distribution of entities created. Use this information to gauge metadata collection performance and volume and to help troubleshoot problems.

Atlas Reference: Statistics

Impala support
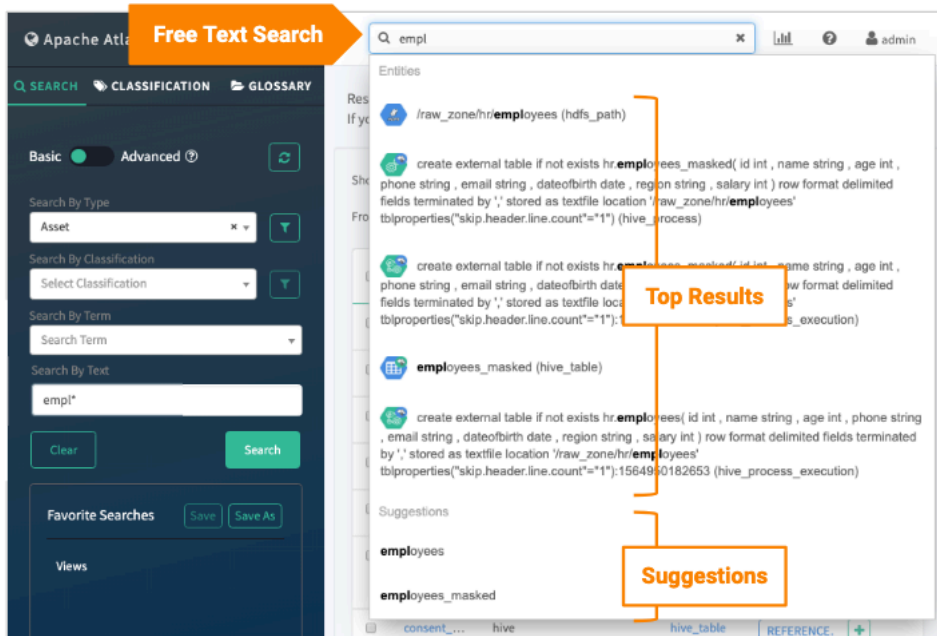
Atlas collects metadata from Impala. Impala operations are represented by Atlas processes; when the same query is run more than once, Atlas creates process execution entities to capture the volume of activity. Atlas collects metadata through Hive Metastore (HMS) for the Hive data assets that Impala operates against.

Atlas Reference: Impala Metadata Collection

What's New in Apache HBase

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

Operational database cluster

You can create an HBase cluster in Amazon AWS from within CDP by selecting the Operational Database cluster template and have a cluster running in under ten minutes.

Operational Datastore.

Amazon S3 as a storage layer

You can use Amazon S3 as a storage layer for HBase where HFiles are written to S3 and WALs are written to HDFS.

Using Amazon S3 with Apache HBase.

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

What's new in HDFS

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

ACID transaction processing

Hive 3 tables are ACID (Atomicity, Consistency, Isolation, and Durability)-compliant, which is critical to observing the right to be forgotten requirement of the GDPR (General Data Protection Regulation).

Shared metastore

Hive metastore (HMS) interoperates with multiple engines, Impala and Spark for example, simplifying interoperation between engines and user data access.

Low-latency analytical processing

Hive processes transactions using low-latency analytical processing (LLAP) or the Hive-on-Tez execution engine.

Spark integration with Hive

You can use Hive to query data from Apache Spark applications without workarounds. The Hive Warehouse Connector supports reading and writing Hive tables from Spark.

Security improvements

Apache Ranger secures Hive data by default. To meet demands for concurrency improvements, ACID support for GDPR (General Data Protection Regulation), render security, and other features, Hive tightly controls the location of the warehouse on a file system, or object store, and memory resources.

Workload management at the query level

You can configure who uses query resources, how much can be used, and how fast Hive responds to resource requests. Workload management can improve parallel query execution, cluster sharing for queries running on Hive LLAP, and performance of non-LLAP queries.

Materialized views

Because multiple queries frequently need the same intermediate roll up or joined table, you can avoid costly, repetitious query portion sharing, by precomputing and caching intermediate tables into views.

Information schema database

When launched, Hive creates two databases from JDBC data sources: information_schema and sys. All Metastore tables are mapped into your tablespace and available in sys. The information_schema data reveals the state of the system, similar to sys database data. You can query information_schema using SQL standard queries, which are portable from one DBMS to another.

What's New in Hue

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

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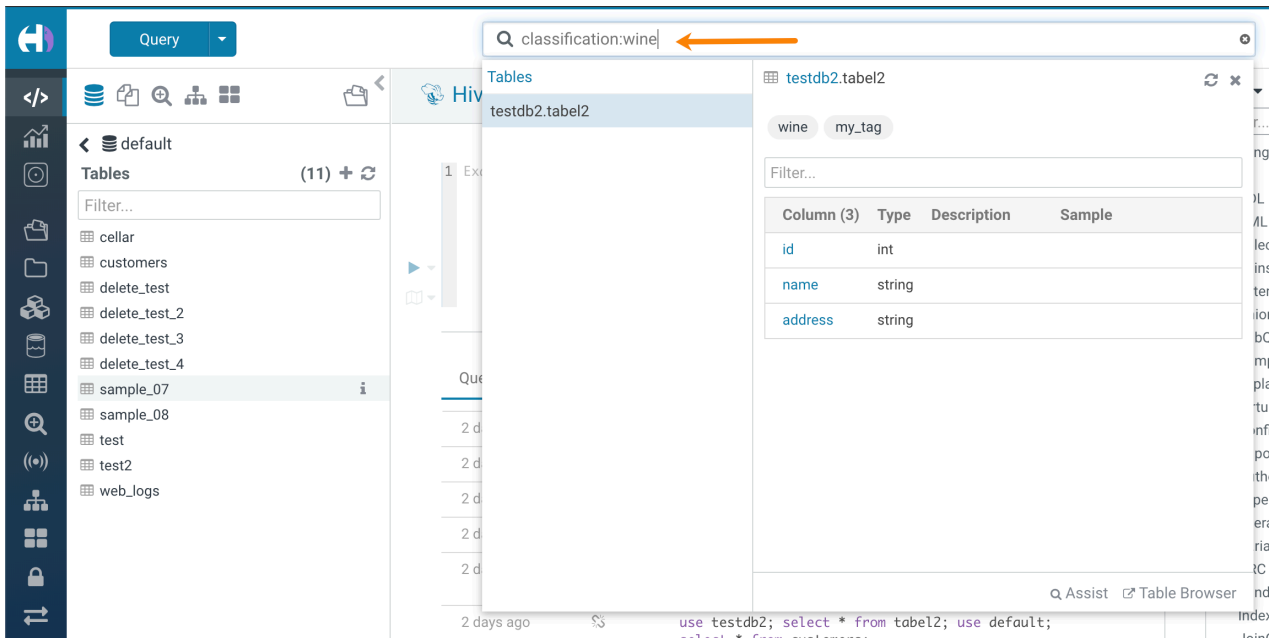
You can also locate tables by searching for Atlas classifications by specifying the tag search term:
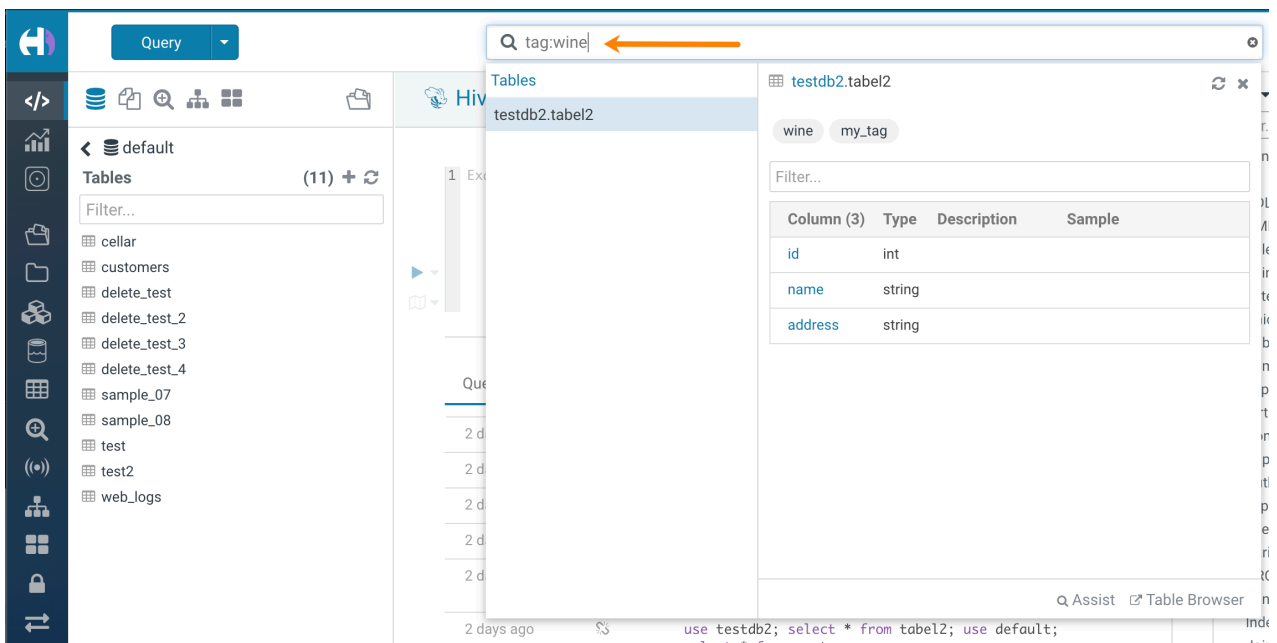
Hive Language Reference

A language reference has been added for Hive:
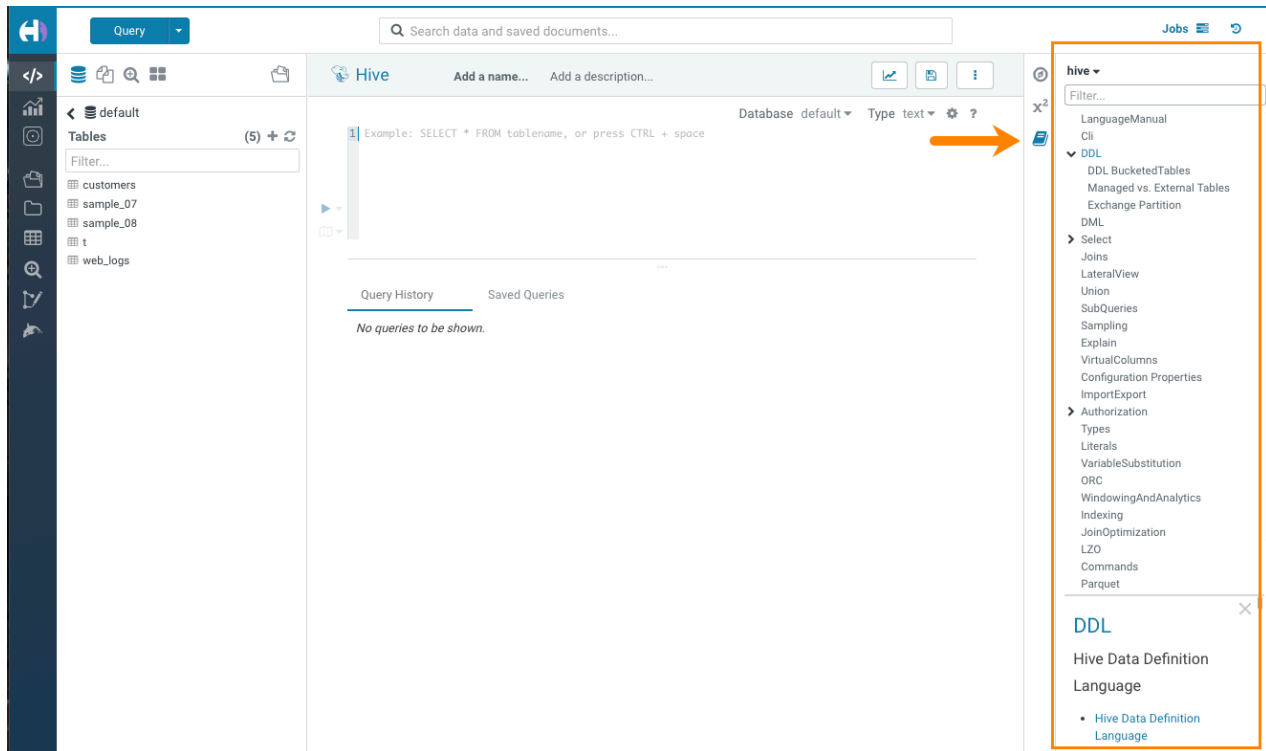
When you select the Hive query editor, click the book icon to the right of the editor window to launch the Hive language reference.

What's New in Apache Impala

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

Increased Compatibility with Components in Cloudera Data Platform

Impala is integrated with the following components:

- Apache Ranger: Use Apache Ranger to manage authorization in Impala. See Impala Authorization for details.
- Apache Atlas: Use Apache Atlas to manage data governance in Impala. See Atlas Metadata Collection Overview
- Hive 3

Parquet Page Index

To improve performance when using Parquet files, Impala can now write page indexes in Parquet files and use those indexes to skip pages for the faster scan.

See Using Parquet in Impala for more information.

The Remote File Handle Cache Supports S3

Impala can now cache remote HDFS file handles when the tables that store their data in Amazon S3 cloud storage.

See Impala Scalability Considerations for the information on remote file handle cache.

Support for Kudu Integrated with Hive Metastore

Kudu is integrated with Hive Metastore (HMS), and from Impala, you can create, update, delete, and query the tables in the Kudu services integrated with HMS.

See Using Kudu with Impala for information on using Kudu tables in Impala.

New Compressions Supported for Parquet Files

The following compressions are supported for the tables in the Parquet file format:

- Zstandard (Zstd)

  Zstd is a real-time compression algorithm offering a tradeoff between speed and ratio of compression. Compression levels from 1 up to 22 are supported. The lower the level, the faster the speed at the cost of compression ratio.
- Lz4

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

Data Cache for Remote Reads

You can execute queries faster on multi-cluster HDFS environments and on object store environments as Impala now caches data for non-local reads (e.g. S3, ABFS, ADLS) on local storage.

The data cache is enabled with the --data_cache startup flag.

See Impala Remote Data Cache for the information and steps to enable remote data cache.

Metadata Performance Improvements

The following features for improving metadata performance are enabled by default in this release:

- Incremental stats are now compressed in memory in catalogd, reducing memory footprint in catalogd.
- impalad coordinators fetch incremental stats from catalogd on-demand, reducing the memory footprint and the network requirements for broadcasting metadata.
- Time-based and memory-based automatic invalidation of metadata to keep the size of metadata bounded and to reduce the chances of catalogdcache running out of memory.
- Automatic invalidation of metadata

  With automatic metadata management enabled, you no longer have to issue INVALIDATE / REFRESH in a number of conditions.

See Impala Metadata Management for the information on the above features.

Scalable Pool Configuration in Admission Controller

To offer more dynamic and flexible resource management, Impala supports the new configuration parameters that scale with the number of executors. You can use the parameters to control the number of running queries, queued queries, and maximum amount of memory allocated for Impala resource pools.

See Impala Admission Control for the information about the new parameters and using them for admission control.

Query Profile

The following metrics were added to the Query Profile output for better monitoring and troubleshooting of query performance.

- Network I/O throughput
- System disk I/O throughput

See Impala Query Profile for generating and reading query profile.

DATE Data Type and Functions

You can use the new DATE data type to describe particular year/month/day values.

This initial DATE type supports the Text, Parquet, and HBASE file formats.

Most of the built-in functions for TIMESTAMP now allow the DATE type arguments, as well.

The support of DATE data type includes the following features:

- DATE type column as a partitioning key column
- DATE literal
- Implicit casting between DATE and other types, namely, STRING and TIMESTAMP

See DATE Data Type and Impala Date and Time Functions for using the DATE type.

Support Hive Insert-Only Transactional Tables

Impala added the support to create, drop, query, and insert into insert-only transactional tables.

Use the Hive compaction to compact small files to improve the performance and scalability of metadata in transactional tables.

See Impala Transactions for more information.

HiveServer2 HTTP Connection for Clients

Now client applications can connect to Impala over HTTP via HiveServer2 with the option to use the Kerberos SPNEGO and LDAP for authentication. See Impala Clients for details.

Default File Format Changed to Parquet

When you create a table, the default format for that table data is now Parquet unless the STORED AS clause is specified.

For backward compatibility, you can use the DEFAULT_FILE_FORMAT query option to set the default file format to the previous default, such as text or other formats.

Built-in Function to Process JSON Objects

The GET_JSON_OBJECT() function extracts JSON object from a string based on the path specified and returns the extracted JSON object.

See Impala Miscellaneous Functions.

Graceful Shutdown of Impala Daemons

You can perform a graceful shutdown of Impala Daemons in Cloudera Manager.

When you initiate a shutdown process for an Impala Daemon, the Impala daemon will notify other Impala daemons that it is shutting down, wait for a grace period, then will shut itself down once no more queries or fragments are executing on that daemon or when the configurable deadline is reached.

See Graceful Shutdown for the steps.

Object Ownership Support

Object ownership for tables, views, and databases is enabled by default in Impala. When you create a database, a table, or a view, as the owner of that object, you implicitly have the privileges on the object. The privileges that owners have are specified in Ranger on the special user, {OWNER}.

The {OWNER} user must be defined in Ranger for the object ownership privileges work in Impala.

See Authorization for details.

What's New in Apache Knox

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

IDBroker

An identity federation solution that provides temporary cloud credentials in exchange for various tokens or authentication.

Trusted Proxy

A service that is allowed to propagate an end-user request (such as hostnames, users, user-groups, and proxy server host IP) to a given backend or service.

What's New in Apache Kudu

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

Table size metrics

Disk size and live row count of a table are exposed as metrics on the master server and on the master server's web UI. And the Kudu tablet's live row count is exposed on the tablet server's web UI, as well as metrics on the tablet server.

Capability to place tablet replicas based on dimensions

You can set dimensions when creating tables and adding partitions. This ensures that the new tablets are evenly distributed within the cluster based on dimension.

Support for adding and dropping range partitions using CLI

You can add and drop range partitions through the command line by using the following commands:

```
kudu table add_range_partition <master_addresses> <table_name> <lower_bound>
 <upper_bound> [-lower_bound_type] [-upper_bound_type]
kudu table drop_range_partition <master_addresses> <table_name> <lower_bo
und> <upper_bound> [-lower_bound_type] [-upper_bound_type]
```

If you do not specify lower_bound and upper_bound, then the range partition is unbounded.

The lower_bound_type and the upper_bound_type parameters are optional. The default values are INCLUSIVE_BOUND and EXCLUSIVE_BOUND respectively. The parameters are case-sensitive.

LZ4 is updated to version 1.9.1

LZ4 is updated to version 1.9.1 for improved performance.

What's New in Apache Oozie

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

For more information about Oozie, see Overview of Oozie.

What's New in Apache Ranger

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

Ranger preloaded resource-based services and policies

Apache Ranger includes preloaded resource-based services and policies. The preloaded resource-based services appear on the Service Manager page for resource-based policies, and are prefixed with "cm_".

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 Apache Hadoop YARN

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

YARN WEB UI2

The user-friendly YARN WEB UI2 is the default user interface.

See:YARN web user interface

IFile logging format

IFile is a newer file controller than the legacy TFile file controller. In an IFile the files are indexed so it is faster to search in the aggregated log file than in a regular TFile.

See:Log Aggregation File Controllers

YARN with Knox TrustedProxy

All YARN web endpoints are behind the Knox TrustedProxy.

See:Authentication with Apache Knox

More reading

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

What's New in Apache ZooKeeper

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

For more information about ZooKeeper, see Managing Apache ZooKeeper.

Known Issues In Cloudera Runtime 7.0.1

Known Issues in Apache Atlas

This topic describes known issues and workarounds for using Atlas in this release of Cloudera Runtime.
REST /search/fulltext endpoint not supported

> The REST API endpoint for /search/fulltext is no longer used and returns an error.

> Workaround: Use the /search/basic or /search/dsl endpoints instead.

Cloudera JIRA: CDPD-3259

Configuration change required to import metadata for Kafka topics into Atlas

When using the import-kafka.sh bridge script for importing Kafka topics into Atlas, the topics present in Kafka are not imported because a configuration property is missing. This script is in the Atlas service packaging found in the CDP runtime distribution parcel.

Workaround: The following steps allow you to import Kafka topics to Atlas using import-kafka.sh:

- Login to the node where Atlas Server / Atlas Gateway role and Kafka Server role is available.
- Edit /etc/atlas/conf/atlas-application.properties and update the property atlas.kafka.zookeeper.connect to have the Kafka's zookeeper.chroot appended to the existing value.

Existing value: - atlas.kafka.zookeeper.connect=*ZOOKEEPER-NODE*: *ZOOKEEPER-PORT*

Updated value: - atlas.kafka.zookeeper.connect=*ZOOKEEPER-NODE*: *ZOOKEEPER-PORTKAFKA-ZOOKEEPER-CHROOT*

Cloudera JIRA: OPSAPS-52224

Table alias values are not found in search

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

Workaround: None.

Cloudera JIRA: CDPD-3208

Hive lineage missing for INSERT OVERWRITE queries

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

Workaround: None.

Cloudera JIRA: CDPD-3160

Logging out of Atlas does not manage the external authentication

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

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

Cloudera JIRA: CDPD-3125

Ranking of top results in free-text search not intuitive

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

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

Cloudera JIRA: CDPD-1892

Free text search in Atlas is case sensitive

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

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

Workaround: None.

Cloudera JIRA: CDPD-1884

Queries with ? wildcard return unexpected results

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

Workaround: None.

Cloudera JIRA: CDPD-1823

Extra WARN messages from Solr in the Atlas log

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

Workaround: These messages can be ignored.

Cloudera JIRA: CDPD-1679

Guest users are redirected incorrectly

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

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

Cloudera JIRA: CDPD-1664

Known Issues in Apache Hadoop

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

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

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

Upstream JIRA

- HADOOP-17208
- HADOOP-17304

Knowledge article

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

Known Issues in Apache HBase

This topic describes known issues and workarounds for using HBase in this release of Cloudera Runtime.
IntegrationTestReplication fails if replication does not finish before the verify phase begins

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

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

HDFS encryption with HBase

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

Workaround: N/A

AccessController postOperation problems in asynchronous operations

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

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

Workaround: N/A

Apache Issue: HBASE-6992

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

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

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

Apache Issue: N/A

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

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

Workaround: N/A

Apache Issue: N/A

Technical Service Bulletins

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

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

Upstream JIRA

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

Knowledge article

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

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

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

Upstream JIRA

HBASE-25206

Knowledge article

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

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

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

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

Knowledge article

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

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

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

Upstream JIRA

- HBASE-23723
- HBASE-25970

Knowledge article

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

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

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

Upstream JIRA

HBASE-24376

Knowledge article

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

Known Issues in HDFS

This topic describes known issues and unsupported features for using HDFS in this release of Cloudera Runtime.
CDPD-2946: Slow reading and writing of erasure-coded files

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

Unsupported Features

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

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

Technical Service Bulletins

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

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

Knowledge article

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

Known Issues in Apache Hive

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

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

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

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

There are two workarounds:

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

Apache Issue: Hive-24815

CDPD-676: Generate Oozie workflow for microstrategy

> Workaround: None

> Apache JIRA: none

Technical Service Bulletins

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

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

Upstream JIRA

> Apache Jira: HIVE-24827

Knowledge article

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

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

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

Upstream JIRA

> Apache Jira: HIVE-24224

Knowledge article

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

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

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

Knowledge article

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

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

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

Knowledge article

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

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

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

Knowledge article

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

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

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

Upstream JIRA

> HIVE-26235

Knowledge article

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

Known Issues in Hue

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

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

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

> Workaround:

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

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

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

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

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

7. In the search text box, type safety and press Enter to locate the Hue Service Advanced Configuration Snippet (Safety Valve) for hue_safety_valve.ini, which appears at the top of the configuration parameters list.

8. Append the following configuration information to the existing configuration information in the Safety Valve and click Save Changes:

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

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

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

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

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

Knowledge article

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

Known Issues in Apache Impala

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

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

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

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

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

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

Upstream JIRA

- IMPALA-5843
- IMPALA-10310

Knowledge article

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

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

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

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

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

Upstream JIRA

> IMPALA-10600

Knowledge article

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

Known Issues in Apache 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 in CDP 1.1.0

  - Kudu does not yet support integration with Ranger
  - Kudu does not yet support integration with Atlas

- Kudu HMS Sync is disabled and not yet supported on CDP 1.1.0

Known Issues in Apache Oozie

This topic describes known issues and unsupported features for using Oozie in this release of Cloudera Runtime.
CDH-14623: 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 feature is currently not supported in Cloudera Data Platform:

> - Non-support for Pig action (CDPD-1070)

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

Known Issues in Apache Ranger

This topic describes known issues and workarounds for using Ranger in this release of Cloudera Runtime.
CDPD-3296: Audit files for Ranger plugin components do not appear immediately in S3 after cluster creation

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

> Workaround:

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

> 1. On the Configuration tab in the applicable service, select Advanced under CATEGORY.
> 2. Click the + icon for the <service_name> Advanced Configuration Snippet (Safety Valve) for ranger-<service_name>-audit.xml property.
> 3. Enter the following property in the Name box:
>
>   xasecure.audit.destination.hdfs.file.rollover.sec.
> 4. Enter the desired rollover interval (in seconds) in the Value box. For example, if you specify 180, the audit log data is updated every 3 minutes.
> 5. Click Save Changes and restart the service.

Known Issues in Apache Solr

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

Technical Service Bulletins

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

> Apache Solr in certain CDH, HDP and CDP releases are vulnerable to a Remote Code Execution through the VelocityResponseWriter. A Velocity template can be provided through Velocity templates in a configset `velocity/` directory or as a parameter. A user defined configset could contain renderable, potentially malicious, templates. Parameter provided templates are disabled

by default, but can be enabled by setting `params.resource.loader.enabled` by defining a response writer with that setting set to `true`. Defining a response writer requires configuration API access. The fix removed the params resource loader entirely, and only enables the configset-provided template rendering when the configset is `trusted` (has been uploaded by an authenticated user).

Knowledge article

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

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

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

Knowledge article

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

Known Issues in Apache Spark

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

CDPD-217: HBase/Spark connectors are not supported

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

Workaround: None

CDPD-3038: Launching pyspark displays several HiveConf warning messages

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

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

Workaround: These errors can be safely ignored.

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

Workaround: Use a different compression algorithm.

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

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

Workaround: Create the view using Hive or Impala.

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'")
```

Known Issues for Apache Sqoop

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

Invalid method name: 'get_index_names'

Problem: During a Sqoop export operation, an invalid method is called. HCatalog returns an exception.

CDPD-3085

Corrupt data is displayed in Beeline

Problem: In Beeline, the output of selecting data from a Hive table imported by Sqoop is wrong. The actual data in the table is correct.

CDPD-3467

Avro, S3, and HCat do not work together properly

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

CDPD-3089

Parquet columns inadvertently renamed

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

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

Apache JIRA: None

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

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

PARQUET-99

Known Issues in MapReduce and YARN

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

Known Issues

JobHistory URL mismatch after server relocation

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

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

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

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

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

CDH-6808: Routable IP address required by ResourceManager

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

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

CDH-17955: Amazon S3 copy may time out

The Amazon S3 filesystem does not support renaming files, and performs a copy operation instead. If the file to be moved is very large, the operation can time out because S3 does not report progress during the operation.

Workaround: Use -Dmapred.task.timeout=15000000 to increase the MR task timeout.

Apache Issue: MAPREDUCE-972

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

Application logs cannot be accessed neither 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

COMPX-8687: Missing access check for getAppAttemps

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

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

1. In Cloudera Manager select the YARN service.
2. Click Configuration.
3. Find the YARN Service MapReduce Advanced Configuration Snippet (Safety   Valve) property.
4. Click the plus icon and add the following:

   • Name: mapreduce.cluster.acls.enabled
   • Value: true
5. Click Save Changes.

Limitations
Capacity Scheduler

   • As Capacity Scheduler is the default scheduler, the Dynamic Resource Pool User Interface is not available by default.

- Capacity Scheduler can be configured only through safety-valves in Cloudera Manager.

Unsupported Features

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

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

Known Issues in Apache Zeppelin

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

CDPD-1683: Zeppelin demo users have been removed

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

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

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

CDPD-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_gr
oup}}]
```

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.

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 Apache ZooKeeper

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

Technical Service Bulletins

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

Each Apache ZooKeeper ("ZooKeeper") client session has a 64bit long session Id, which must be unique among the online sessions. The clients can connect to any ZooKeeper server and each server assigns the session Id by continuously increasing an "initial session Id" which was set at the start of each ZooKeeper server process.

The bug ZOOKEEPER-1622 would cause:

- Generation of negative ZooKeeper client session Ids causing Apache Accumulo ("Accumulo") to fail to start
- ZooKeeper client session ID collision on clusters which are heavily using ZooKeeper, resulting in unexpected ZooKeeper session termination or possibly failing ZooKeeper client operations. This may cause unexpected session termination, data inconsistency, or operation failure, among other issues, in downstream services (e.g. HBase service failures, failovers in HA services, etc.)

Upstream JIRA

ZOOKEEPER-1622

Knowledge article

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

Cloudera Runtime Component Versions

Versions of Cloudera Runtime 7.0.1 components.

Component	Version
Apache Atlas	2.0.0
Apache Avro	1.8.2
Apache Hadoop	3.1.1
Apache HBase	2.2.1
HBase Indexer	1.5
Apache Hive	3.1.2
Hue	4.5.0
Apache Impala	3.3.0
Apache Kafka	2.3.0

Component	Version
Apache Kudu	1.11.0
Apache Oozie	5.1.0
Apache ORC	1.5.1
Apache Parquet	1.10.99
Apache Ranger	2.0.0
Apache Solr	7.4.0
Apache Spark	2.4.0
Apache Sqoop	1.4.7
Apache Tez	0.9.1
Apache Zookeeper	3.4.6