Cloudera Runtime 7.3.3

Apache Spark Overview

Date published: 2020-07-28
Date modified: 2026-12-13



https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2026. All rights reserved.

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

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

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

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

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

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

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

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

Contents

Apache Spark Overview..... 4

Unsupported Apache Spark Features..... 4

Apache Spark Overview

Apache Spark is a distributed, in-memory data processing engine designed for large-scale data processing and analytics.

Apache Spark **3.5.8** ships with Cloudera Runtime 7.3.2.10000 SP1 and higher on the 7.3.2 line (Apache Spark 3.5.4 in 7.3.2.0).

Important:

When Cloudera rebases the Apache Spark version shipped with or CDS, minor and maintenance releases can still change SQL behavior, configuration defaults, and serialization. Recompile Java and Scala applications built against Spark, and retest PySpark and Spark SQL workloads in a non-production environment before you upgrade production clusters.

Review *Upgrading Spark* and *Migrating Spark applications* for steps that match your source and target Spark versions.

Read the *Apache Spark migration guides* for the **exact** Apache Spark version on your cluster (for example 3.5.8 or 3.5.4).

Apache Spark is a general framework for distributed computing that offers high performance for both batch and interactive processing. It exposes APIs for Java, Python, and Scala and consists of Spark core and several related projects.

You can run Spark applications locally or distributed across a cluster, either by using an interactive shell or by submitting an application. Running Spark applications interactively is commonly performed during the data-exploration phase and for ad hoc analysis.

To run applications distributed across a cluster, Spark requires a cluster manager. Cloudera Data Platform (CDP) supports only the YARN cluster manager. When run on YARN, Spark application processes are managed by the YARN ResourceManager and NodeManager roles. Spark Standalone is not supported.

For detailed API information, see the Apache Spark project site.

Note: Although this document makes some references to the external Spark site, not all the features, components, recommendations, and so on are applicable to Spark when used on CDH. Always cross-check the Cloudera documentation before building a reliance on some aspect of Spark that might not be supported or recommended by Cloudera.

CDP supports Apache Spark, Apache Livy for local and remote access to Spark through the Livy REST API, and Apache Zeppelin for browser-based notebook access to Spark. The Spark LLAP connector is not supported.

Related Information

Upgrading Spark
Upgrading Spark
Migrating Spark applications
Migrating Spark applications

Unsupported Apache Spark Features

The following Apache Spark features are not supported in Cloudera.

Spark

- Apache Spark experimental features/APIs are not supported unless stated otherwise.

- Spark Streaming (DStreams) reading from Kafka topics containing transactions such as idempotent producer being used to publish records.
- Using the JDBC Datasource API to access Hive or Impala.
- Spark with Kudu is not supported for ADLS data.
- IPython / Jupyter notebooks is not supported. The IPython notebook system (renamed to Jupyter as of IPython 4.0) is not supported.
- Certain Spark Streaming features, such as the mapWithState method, are not supported.
- Thrift JDBC/ODBC server (also known as Spark Thrift Server or STS)
- Spark SQL CLI
- GraphX
- SparkR
- GraphFrames
- Spark Connect
- Structured Streaming is supported, but the following features of it are not:
  - Continuous processing, which is still experimental, is not supported.
  - Stream static joins with HBase have not been tested and therefore are not supported.
- Structured Streaming is not supported with Iceberg tables
- Hudi
- Push-based shuffle
- ZSTD compression in ORC data source (SPARK-33978)
- spark.hadoopRDD.ignoreEmptySplits (SPARK-34809)
- LDAP authentication for livy-server (LIVY-356)
- Thrift ldap authentication, based on ldapurl, basedn, domain (LIVY-678)

Spark-Atlas Connector

- The HBase Connector's Atlas lineage generation is not supported in the Spark-Atlas Connector
- The Hive Warehouse Connector's Atlas lineage generation is not supported in the Spark-Atlas Connector