Cloudera Runtime 7.3.2

Apache Spark Overview

Date published: 2020-07-28
Date modified: 2026-03-31

CLOUDƎRA

Legal Notice

Contents

Apache Spark Overview

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

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.

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
- Atlas lineage is not captured for MERGE INTO operations on Iceberg tables