Cloudera Runtime 7.1.1

CDS 3 (Experimental) Powered by Apache Spark

Date published: 2020-01-30
Date modified: 2020-12-15

CLOUDERA

https://docs.cloudera.com/

Legal Notice

Contents

CDS 3 (Experimental) Powered by Apache Spark Overview

> **Warning:** This release, based on the Apache Spark 3.0 preview2 release, is experimental, and should be used for testing and evaluation only. Do not use it in production. This release is not supported by Cloudera Support. To ask questions and provide feedback, please use the Cloudera Community forum.

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.

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

CDS Powered by Apache Spark is an add-on service for CDP, distributed as a parcel and custom service descriptor.

This document describes CDS 3.0 (Experimental) Powered by Apache Spark. It enables you to install and evaluate the features of Apache Spark 3 without upgrading your CDP Data Center cluster.

On CDP Data Center, a Spark 3 service can coexist with the existing Spark 2 service. The configurations of the two services do not conflict and both services use the same YARN service. The port of the Spark History Server is 18088 for Spark 2 and 18089 for Spark 3.

Unsupported Features:

This experimental release does not support the following features:

- Phoenix Connector
- SparkR
- Hive Warehouse Connector
- Kudu
- HBase Connector
- Oozie
- Livy
- Zeppelin

CDS 3 (Experimental) Powered by Apache Spark Requirements

The following sections describe software requirements for CDS 3 (Experimental) Powered by Apache Spark.

> **Warning:** This release, based on the Apache Spark 3.0 preview2 release, is experimental, and should be used for testing and evaluation only. Do not use it in production. This release is not supported by Cloudera Support. To ask questions and provide feedback, please use the Cloudera Community forum.

CDP Versions

> **Important:** CDS 3 (Experimental) Powered by Apache Spark is an add-on service for CDP Data Center. It is not supported in CDP public cloud versions. Spark 2 is included in CDP, and does not require a separate parcel.

Supported versions of CDP are described below.

CDS Powered by Apache Spark Version	Supported CDP Versions
3.0.0.2.99.0.0-23	CDP Data Center 7.0 with Cloudera Runtime 7.0.3
3.0.0.2.99.7110.0-18	CDP Data Center 7.1 with Cloudera Runtime 7.1.1

A Spark 2 service (included in CDP) can co-exist on the same cluster as Spark 3 (installed as a separate parcel). The two services are configured to not conflict, and both run on the same YARN service. Spark 3 installs and uses its own external shuffle service.

Although Spark 2 and Spark 3 can coexist in the same CDP Data Center cluster, you cannot use multiple Spark 3 versions simultaneously. All clusters managed by the same Cloudera Manager Server must use exactly the same version of CDS Powered by Apache Spark.

Python Requirement

CDS 3 requires Python 3.4 or higher.

JDK Requirement

CDS 3 requires JDK 8 or JDK 11. Remove other JDK versions from all cluster and gateway hosts to ensure proper operation.

Installing CDS 3.0 (Experimental) Powered by Apache Spark

Warning: This release, based on the Apache Spark 3.0 preview2 release, is experimental, and should be used for testing and evaluation only. Do not use it in production. This release is not supported by Cloudera Support. To ask questions and provide feedback, please use the Cloudera Community forum.

CDS 3.0 (Experimental) Powered by Apache Spark is distributed as two files: a custom service descriptor file and a parcel, both of which must be installed on the cluster.

Note: Due to the potential for confusion between CDS Powered by Apache Spark and the initialism CSD, references to the custom service descriptor (CSD) file in this documentation use the term service descriptor.

Install CDS Powered by Apache Spark

Note:

Although Spark 2 and Spark 3 can coexist in the same CDP Data Center cluster, you cannot use multiple Spark 3 versions simultaneously. All clusters managed by the same Cloudera Manager Server must use exactly the same version of CDS Powered by Apache Spark.

Follow these steps to install CDS 3 Powered by Apache Spark:

1. Check that all the software prerequisites are satisfied. If not, you might need to upgrade or install other software components first.
2. Install the CDS Powered by Apache Spark service descriptor into Cloudera Manager.

    a. To download the CDS Powered by Apache Spark service descriptor, click the service descriptor link for the version you want to install.
    b. Log on to the Cloudera Manager Server host, and copy the CDS Powered by Apache Spark service descriptor in the location configured for service descriptor files.
    c. Set the file ownership of the service descriptor to cloudera-scm:cloudera-scm with permission 644.
    d. Restart the Cloudera Manager Server with the following command:

    ```
    systemctl restart cloudera-scm-server
    ```

3. In the Cloudera Manager Admin Console, add the CDS parcel repository to the Remote Parcel Repository URLs in Parcel Settings as described in Parcel Configuration Settings.

   Note: If your Cloudera Manager Server does not have Internet access, you can use the CDS Powered by Apache Spark parcel files: put them into a new parcel repository, and then configure the Cloudera Manager Server to target this newly created repository.

4. Download the CDS Powered by Apache Spark parcel, distribute the parcel to the hosts in your cluster, and activate the parcel. For instructions, see Managing Parcels.

5. Add the Spark 3 service to your cluster.

   a. In step 1, select any optional dependencies, such as HBase and Hive, or select No Optional Dependencies.

   b. In step 2, when customizing the role assignments, add a gateway role to every host.

   c. On the Review Changes page, you can enable TLS for the Spark History Server.

   d. Note that the History Server port is 18089 instead of the usual 18088.

   e. Complete the remaining steps in the wizard.

6. Return to the Home page by clicking the Cloudera Manager logo in the upper left corner.

7. Click the stale configuration icon to launch the Stale Configuration wizard and restart the necessary services.

CDS 3 (Experimental) Powered by Apache Spark Version and Download Information

Warning: This release, based on the Apache Spark 3.0 preview2 release, is experimental, and should be used for testing and evaluation only. Do not use it in production. This release is not supported by Cloudera Support. To ask questions and provide feedback, please use the Cloudera Community forum.

The following sections provide links to the parcel and service descriptor files for CDS 3 (Experimental).

CDS 3 Versions Available for Download

Table 1: Available CDS Versions

Version	Custom Service Descriptor	Parcel Repository
3.0.0.2.99.7110.0-18	SPARK3_ON_YARN-3.0.0.2.99.7110.0-18.jar	http://archive.cloudera.com/p/spark3/2.99.7110.0/parcels/ **Note:** This repository requires authentication. If you do not know your username or password, contact your Cloudera account representative.

Using the CDS 3 Powered by Apache Spark Maven Repository

Important: CDS 3 does not include an assembly JAR. When you build an application JAR, do not include Cloudera Runtime or CDS JARs, because they are already provided. If you do, upgrading Cloudera Runtime or CDS can break your application. To avoid this situation, set the Maven dependency scope to provided. If you have already built applications which include the Cloudera Runtime or CDS JARs, update the dependency to set scope to provided and recompile.

If you want to build applications or tools for use with CDS Powered by Apache Spark, and you are using Maven or Ivy for dependency management, you can pull the CDS artifacts from the Cloudera Maven repository. The repository is available at https://archive.cloudera.com/p/spark3/2.99.7110.0/maven-repository/.

The following is a sample POM (pom.xml) file:

```
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.
org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache
.org/maven-v4_0_0.xsd">
  <repositories>
    <repository>
      <id>cloudera</id>
      <url>https://archive.cloudera.com/p/spark3/2.99.7110.0/maven-reposit
ory/</url>
    </repository>
  </repositories>
</project>
```

To specify credentials to access the repository, add or modify the servers section in the Maven settings.xml file as follows:

```
<servers>
    <server>
      <id>cloudera</id>
      <username>username</username>
      <password>password</password>
    </server>
```

If you do not know your username or password, contact your Cloudera account representative.

CDS 3.0 (Experimental) Powered by Apache Spark Maven Artifacts

The following tables lists the groupId, artifactId, and version required to access the artifacts for CDS 3 Powered by Apache Spark:

CDS 3.0 Maven Artifacts

The following pom fragment shows how to access a CDS 3.0 artifact from a Maven POM.

```
<dependency>
      <groupId>org.apache.spark</groupId>
      <artifactId>spark-core_2.12</artifactId>
      <version>3.0.0.2.99.7110.0-18</version>
      <scope>provided</scope>
</dependency>
```

The complete artifact list for this release follows.

Project	groupId	artifactId	version
Apache Spark	org.apache.spark	spark-assembly_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-avro_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-catalyst_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-core_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-cypher_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-graph-api_2.12	3.0.0.2.99.7110.0-18

Project	groupId	artifactId	version
	org.apache.spark	spark-graph_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-graphx_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-hadoop-cloud_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-hive-thriftserver_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-hive_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-kubernetes_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-kvstore_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-launcher_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-mllib-local_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-mllib_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-network-common_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-network-shuffle_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-network-yarn_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-parent_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-repl_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-sketch_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-sql-kafka-0-10_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-sql_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-streaming-kafka-0-10-assembly_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-streaming-kafka-0-10_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-streaming_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-tags_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-token-provider-kafka-0-10_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-unsafe_2.12	3.0.0.2.99.7110.0-18
	org.apache.spark	spark-yarn_2.12	3.0.0.2.99.7110.0-18