Apache Spark 3

Installing Apache Spark

Date of Publish: 2018-04-01

http://docs.hortonworks.com

Contents

Spark prerequisites

Before installing Spark, ensure that your cluster meets the following prerequisites.

- HDP cluster stack version 3.0 or later
- (Optional) Ambari version 2.7.0 or later
- HDFS and YARN deployed on the cluster

Only Spark version 2 is supported.

Additionally, note the following requirements and recommendations for optional Spark services and features:

- Spark Thrift server requires Hive deployed on the cluster.
- SparkR requires R binaries installed on all nodes.
- Spark access through Livy requires the Livy server installed on the cluster.
- PySpark and associated libraries require Python version 2.7 or later, or Python version 3.4 or later, installed on all nodes.
- For optimal performance with MLlib, consider installing the netlib-java library.

Related Information

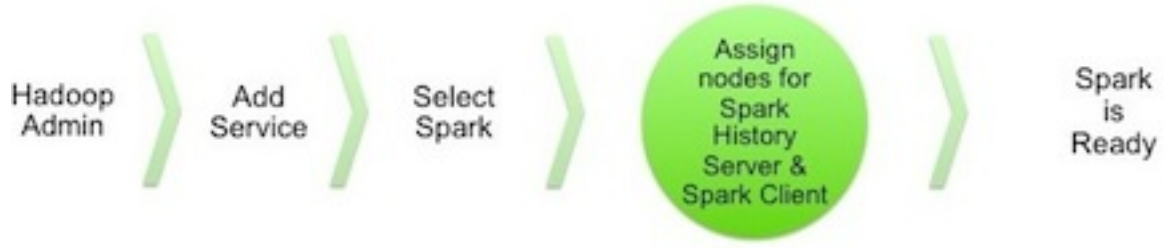
netlib-java library

Install Spark using Ambari

Use the following steps to install Apache Spark on an Ambari-managed cluster.

About this task

The following diagram shows the Spark installation process using Ambari. Before you install Spark using Ambari, refer to "Adding a Service" in the Ambari Managing and Monitoring a Cluster guide for background information about how to install Hortonworks Data Platform (HDP) components using Ambari.



Caution:

During the installation process, Ambari creates and edits several configuration files. If you configure and manage your cluster using Ambari, do not edit these files during or after installation. Instead, use the Ambari web UI to revise configuration settings.

Procedure

1. Click the ellipsis (…) symbol next to Services on the Ambari dashboard, then click Add Service.
2. On the Add Service Wizard, select Spark2, then click Next.
3. On the Assign Masters page, review the node assignment for the Spark2 History Server, then click Next.
4. On the Assign Slaves and Clients page:

a. Scroll to the right and select the client nodes where you want to run Spark clients. These are the nodes from which Spark jobs can be submitted to YARN.
b. To install the optional Livy server for security and user impersonation features, select the Livy for Spark2 Server box for the desired node assignment.
c. To install the optional Spark Thrift server for ODBC or JDBC access, review the Spark2 Thrift Server node assignments and assign one or two nodes to the Thrift Server.

   Deploying the Thrift server on multiple nodes increases scalability of the Thrift server. When specifying the number of nodes, take into consideration the cluster capacity allocated to Spark.

5. Click Next to continue.
6. On the Customize Services page, set the following configuration property for the Thrift Server:
   a. Click Advanced spark-thrift-sparkconf.
   b. Set the spark.yarn.queue property value to the name of the YARN queue that you want to use.
7. Click Next to continue.
8. If Kerberos is enabled on the cluster, review the principal and keytab settings on the Configure Identities page, modify the settings if desired, then click Next.
9. Review the configuration on the Review page, then click Deploy to begin the installation.
10. The Install, Start, and Test page displays the installation status.
11. When the progress bar reaches 100% and a "Success" message appears, click Next.
12. On the Summary page, click Complete to finish installing Spark.

Related Information

Adding a Service

Verify the Spark configuration for Hive access

Use the following steps to verify the Spark configuration for Hive access.

When you install Spark using Ambari, the hive-site.xml file is automatically populated with the Hive metastore location.

If you move Hive to a different server, edit the SPARK_HOME/conf/hive-site.xml file so that it contains only the hive.metastore.uris property. Make sure that the host name points to the URI where the Hive metastore is running, and that the Spark copy of hive-site.xml contains only the hive.metastore.uris property.

```
<configuration>
  <property>
  <name>hive.metastore.uris</name>
    <!-- hostname must point to the Hive metastore URI in your cluster -->
    <value>thrift://hostname:9083</value>
    <description>URI for client to contact metastore server</description>
  </property>
</configuration>
```

Validate the Spark installation

To validate the Spark2 installation process, run the Spark Pi and WordCount jobs supplied with the Spark package.

For more information, see "Running Spark Applications" in this guide.