DEV-CDM-Spark02

November-10-17 1:15 PM

Installing Spark

Check OS Release

```
[administrator@dev-cdm-spark0* ~]$ cat /etc/os-release
NAME="CentOS Linux"
VERSION="7 (Core)"
ID="centos"
ID_LIKE="rhel fedora"
VERSION_ID="7"
PRETTY_NAME="CentOS Linux 7 (Core)"
ANSI_COLOR="0;31"
CPE_NAME="cpe:/o:centos:centos:7"
HOME_URL="https://www.centos.org/"
BUG_REPORT_URL="https://bugs.centos.org/"

CENTOS_MANTISBT_PROJECT="CentOS-7"
CENTOS_MANTISBT_PROJECT_VERSION="7"
REDHAT_SUPPORT_PRODUCT="centos"
REDHAT_SUPPORT_PRODUCT_VERSION="7"
```

Check what Operating System Was installed

Install Java

```
[administrator@dev-cdm-spark0* ~]$ yum list java-1.*openjdk.*
Loaded plugins: fastestmirror, langpacks
Loading mirror speeds from cached hostfile
 * base: mirror.gpmidi.net
 * extras: mirror.gpmidi.net
 * updates: mirror.gpmidi.net
Available Packages
java-1.6.0-openjdk.x86_64        1:1.6.0.41-1.13.13.1.el7_3        updates
java-1.7.0-openjdk.x86_64        1:1.7.0.141-2.6.10.1.el7_3        updates
java-1.8.0-openjdk.i686        1:1.8.0.131-3.b12.el7_3        updates
java-1.8.0-openjdk.x86_64        1:1.8.0.131-3.b12.el7_3        updates
```

List available Java

Pick latest

```
[administrator@dev-cdm-spark0* ~]$ sudo yum install java-1.8.0-openjdk
```

Install 1.8

```
[administrator@dev-cdm-spark0* ~]$ java -version
openjdk version "1.8.0_131"
OpenJDK Runtime Environment (build 1.8.0_131-b12)
OpenJDK 64-Bit Server VM (build 25.131-b12, mixed mode)
```

Verify

Supported Interpreters

Thanks to many Zeppelin contributors, we can provide much more interpreters in every release. Please check the below table before you download Zeppelin package.

Note : Only Spark interpreter is included in the netinst binary package by default. If you want to use the other interpreters, you need to install them using net-install script.

Zeppelin	0.7.3	0.7.1 - 0.7.2	0.7.0	0.6.2 - 0.6.1	0.6.0
Spark	1.4.x, 1.5.x, 1.6.x, 2.0.x, 2.1.x, **2.2.0**	1.4.x, 1.5.x, 1.6.x, 2.0.x **2.1.0**	1.4.x, 1.5.x, 1.6.x, 2.0.x **2.1.0**	1.1.x, 1.2.x, 1.3.x, 1.4.x, 1.5.x, 1.6.x, **2.0.0**	1.1.x, 1.2.x, 1.3.x, 1.4.x, 1.5.x, 1.6.x
				Support Scala 2.11	SparkR is available
JDBC	PostgreSQL, MySQL,MariaDB, Redshift, Hive, Phoenix, Drill ,Tajo are available	PostgreSQL, MySQL,MariaDB, Redshift, Hive, Phoenix, Drill ,Tajo are available	PostgreSQL, MySQL,MariaDB, Redshift, Hive, Phoenix, Drill ,Tajo are available	PostgreSQL, MySQL ,MariaDB, Redshift, Hive, Phoenix, Drill, Tajo are available	PostgreSQL,MySQL,MariaDB,Redshift, Hive,Phoenix,Drill, Tajoare available
Pig	O	O	O	N/A	N/A
Beam	O	O	O	N/A	N/A

Scio	O	O	O	N/A	N/A
BigQuery	O	O	O	O	N/A
Python	O	O	O	O	O
Livy	O	O	O	O	O
HDFS	O	O	O	O	O
Alluxio	O	O	O	O	O
Hbase	O	O	O	O	O
Scalding	O	O	O	O	O
Elasticsearch	O	O	O	O	O
Angular	O	O	O	O	O
Markdown	O	O	O	O	O
Shell	O	O	O	O	O
Flink	O	O	O	O	O
Cassandra	O	O	O	O	O
Geode	O	O	O	O	O
Ignite	1.9.0	1.9.0	1.7.0	1.7.0	1.6.0
Kylin	O	O	O	O	O
Lens	O	O	O	O	O
PostgreSQL	O	O	O	O	O

From <https://zeppelin.apache.org/supported_interpreters.html>

We need to install Scala

Install Scala

Used Scala 2.11.8-0 check for other versions at http://www.scala-lang.org/files/archive/

[administrator@dev-cdm-spark0* ~]$ **cd $HOME** **pwd** /home/administrator	Set HOME
[administrator@dev-cdm-spark0* ~]$ **wget http://www.scala-lang.org/files/archive/scala-2.11.8-0.rpm** Saving to: ‘scala-2.10.6.rpm’ 100%[==>] 26,067,733 3.02MB/s in 7.3s 2017-06-27 12:23:05 (3.40 MB/s) - ‘scala-2.10.6.rpm’ saved [26067733/26067733]	Get the 2-10-6
[administrator@dev-cdm-spark0* ~]$ **sudo yum install scala-2.11.8-0.rpm** Loaded plugins: fastestmirror, langpacks Complete!	Install it
[administrator@dev-cdm-spark0* ~]$ **scala** Welcome to Scala version 2.10.6 (OpenJDK 64-Bit Server VM, Java 1.8.0_131). Type in expressions to have them evaluated. Type :help for more information. scala> **1+2** res0: Int = 3 scala> **exit** warning: there were 1 deprecation warning(s); re-run with -deprecation for details [administrator@dev-cdm-spark01 ~]$	Verify

Install Apache Spark

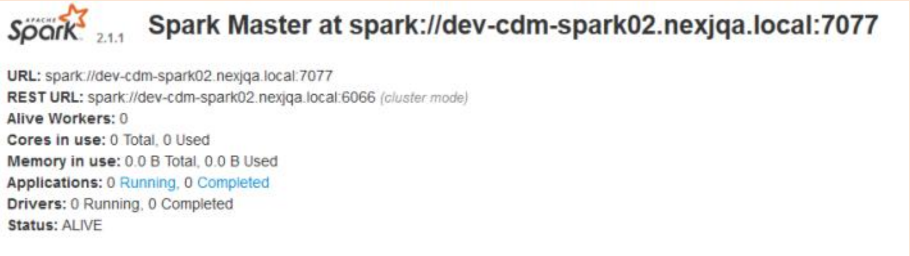
Navigate to http://spark.apache.org/downloads.html and select the 2.1.1 version

[administrator@dev-cdm-spark0* ~]$ **cd $HOME** **pwd** /home/administrator	Set HOME
[administrator@dev-cdm-spark0* ~]$	Get based on the

wget https://d3kbcqa49mib13.cloudfront.net/spark-2.1.1-bin-hadoop2.7.tgz 100%[===>] 279,470,513 5.92MB/s in 51s	URL
[administrator@dev-cdm-spark0* ~]$ **sudo mkdir /usr/local/spark** **ls -al /usr/local/ | grep spark** drwxr-xr-x 2 root root 4096 Jun 27 15:23 spark	Target folder
[administrator@dev-cdm-spark0* ~]$ **sudo tar xzf spark-1.6.3-bin-hadoop2.6.tgz -C /usr/local/spark --strip-components 1**	Unpack without leading folder
[administrator@dev-cdm-spark0* spark]$ **sudo useradd hadoop**	Create hadoop user if needed
[administrator@dev-cdm-spark0* spark]$ **cd /usr/local/spark** [administrator@dev-cdm-spark0* spark]$ **sudo chown administrator.hadoop . -R**	Chown root.root ownership
[administrator@dev-cdm-spark0* ~]$ **ll /usr/local/spark/** total 1412 drwxr-xr-x 2 hadoop hadoop 4096 Nov 2 2016 bin -rw-r--r-- 1 hadoop hadoop 1343562 Nov 2 2016 CHANGES.txt drwxr-xr-x 2 hadoop hadoop 4096 Jun 27 17:00 conf drwxr-xr-x 3 hadoop hadoop 4096 Nov 2 2016 data drwxr-xr-x 3 hadoop hadoop 4096 Nov 2 2016 ec2 drwxr-xr-x 3 hadoop hadoop 4096 Nov 2 2016 examples drwxr-xr-x 2 hadoop hadoop 4096 Nov 2 2016 lib -rw-r--r-- 1 hadoop hadoop 17352 Nov 2 2016 LICENSE drwxr-xr-x 2 hadoop hadoop 4096 Nov 2 2016 licenses drwxr-xr-x 2 hadoop hadoop 4096 Jun 28 10:33 logs -rw-r--r-- 1 hadoop hadoop 23529 Nov 2 2016 NOTICE drwxr-xr-x 6 hadoop hadoop 4096 Nov 2 2016 python drwxr-xr-x 3 hadoop hadoop 4096 Nov 2 2016 R -rw-r--r-- 1 hadoop hadoop 3359 Nov 2 2016 README.md -rw-r--r-- 1 hadoop hadoop 120 Nov 2 2016 RELEASE drwxr-xr-x 2 hadoop hadoop 4096 Nov 2 2016 sbin	Verify
vi $HOME/.bashrc <<append>> **# Spark HOME** **export SPARK_HOME=/usr/local/spark** **# Spark PATH** **PATH=$PATH:$SPARK_HOME/bin** **export PATH**	Add variables
[administrator@dev-cdm-spark0* ~]$. **.bashrc** [administrator@dev-cdm-spark0* ~]$ **echo $PATH** /usr/local/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/home/administrator/.local/bin:/home/administrator/bin:/home/administrator/.local/bin:/home/administrator/bin:/usr/local/spark/bin	Check
[administrator@dev-cdm-spark0* ~]$ **spark-shell** Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties Setting default log level to "WARN". To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel). 17/11/10 13:26:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable 17/11/10 13:26:24 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException Spark context Web UI available at http://10.10.20.92:4040 Spark context available as 'sc' (master = local[*], app id = local-1510338380917). Spark session available as 'spark'. Welcome to ____ __ / __/__ ___ _____/ /__ _\ \/ _ \/ _ `/ __/ '_/ /___/ .__/_,_/_/ /_/_\ version 2.1.1 /_/ Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_131) Type in expressions to have them evaluated. Type :help for more information. scala> **val file=sc.textFile("/usr/local/spark/README.md")** file: org.apache.spark.rdd.RDD[String] = /usr/local/spark/README.md MapPartitionsRDD[1] at textFile at <console>:24	Validate Spark

scala> **file.count();** res0: Long = 104	
Failed to create database 'metastore_db' error? **cd $HOME** **spark-shell**	Must have rwx privs to current directory
[administrator@dev-cdm-spark02 ~]$ **cd $SPARK_HOME/conf** **sudo cp spark-env.sh.template spark-env.sh** **ls -al spark-env.sh** -rwxr-xr-x 1 root root 4209 Jun 27 16:58 spark-env.sh	Copy the template shell
[administrator@dev-cdm-spark02 ~]$ **sudo vi $SPARK_HOME/conf/spark-env.sh** <<update>> **export SPARK_MASTER_HOST=dev-cdm-spark01.nexjqa.local:7077** **SPARK_MASTER_WEBUI_PORT=8880**	Edit the conf shell and add the Master variable
[administrator@dev-cdm-spark02 ~]$ **. $SPARK_HOME/conf/spark-env.sh** [administrator@dev-cdm-spark02 ~]$ **echo $SPARK_MASTER_HOST** dev-cdm-spark01.nexjqa.local:7077	Register the variable
[administrator@dev-cdm-spark02 ~]$ **cd $HOME** [administrator@dev-cdm-spark02 ~]$ **sudo /usr/local/spark/sbin/start-master.sh** starting org.apache.spark.deploy.worker.Worker, logging to /usr/local/spark/logs/spark-root-org.apache.spark.deploy.worker.Worker-1-dev-cdm-spark02.nexjqa.local.out	Start the worker slave
[administrator@dev-cdm-spark02 ~]$ **cat /usr/local/spark/logs/spark-root-org.apache.spark.deploy.worker.Worker-1-dev-cdm-spark02.nexjqa.local.out** Spark Command: /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.131-3.b12.el7_3.x86_64/jre/bin/java -cp /usr/local/spark/conf/:/usr/local/spark/jars/* -Xmx1g org.apache.spark.deploy.master.Master --host dev-cdm-spark02.nexjqa.local --port 7077 --webui-port 8880 == Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties 17/11/10 10:23:47 INFO Master: Started daemon with process name: 11597@dev-cdm-spark02.nexjqa.local 17/11/10 10:23:47 INFO SignalUtils: Registered signal handler for TERM 17/11/10 10:23:47 INFO SignalUtils: Registered signal handler for HUP 17/11/10 10:23:47 INFO SignalUtils: Registered signal handler for INT 17/11/10 10:23:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable 17/11/10 10:23:47 INFO SecurityManager: Changing view acls to: root 17/11/10 10:23:47 INFO SecurityManager: Changing modify acls to: root 17/11/10 10:23:47 INFO SecurityManager: Changing view acls groups to: 17/11/10 10:23:47 INFO SecurityManager: Changing modify acls groups to: 17/11/10 10:23:47 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); groups with view permissions: Set(); users with modify permissions: Set(root); groups with modify permissions: Set() 17/11/10 10:23:47 INFO Utils: Successfully started service 'sparkMaster' on port 7077. 17/11/10 10:23:47 INFO Master: Starting Spark master at spark://dev-cdm-spark02.nexjqa.local:7077 17/11/10 10:23:47 INFO Master: Running Spark version 2.1.1 17/11/10 10:23:47 INFO Utils: Successfully started service 'MasterUI' on port 8880. 17/11/10 10:23:47 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://10.10.20.92:8880 17/11/10 10:23:47 INFO Utils: Successfully started service on port 6066. 17/11/10 10:23:47 INFO StandaloneRestServer: Started REST server for submitting applications on port 6066 17/11/10 10:23:48 INFO Master: I have been elected leader! New state: ALIVE	Check that it started

Check Installation

http://dev-cdm-spark01.nexjqa.local:8080/  Spark Master at spark://dev-cdm-spark02.nexjqa.local:7077 **URL:** spark://dev-cdm-spark02.nexjqa.local:7077 **REST URL:** spark://dev-cdm-spark02.nexjqa.local:6066 *(cluster mode)* **Alive Workers:** 0 **Cores in use:** 0 Total, 0 Used **Memory in use:** 0.0 B Total, 0.0 B Used **Applications:** 0 Running, 0 Completed **Drivers:** 0 Running, 0 Completed **Status:** ALIVE	Check URL

RIAK Basho Spark Adapter

https://github.com/basho/spark-riak-connector

★ **spark-riak-connector from the repository comes without pyspark support. But you can build it yourself and attach to pyspark** From <https://stackoverflow.com/questions/40799372/how-to-use-spark-riak-connector-with-pyspark>	NO Python Support by default
[administrator@dev-cdm-spark0* ~]$ **sudo yum install git** **git clone https://github.com/basho/spark-riak-connector.git** **cd spark-riak-connector/** **python connector/python/setup.py bdist_egg # creates egg file inside connector/python/dist/**	Git clone repo and create egg
[administrator@dev-cdm-spark0* ~]$ **sudo chown hadoop.hadoop -R *** **cd $HOME** **sudo mv ./spark-riak-connector /usr/local/spark/python**	Chmod and chmove
[administrator@dev-cdm-spark0* ~]$ **vi $HOME/.bashrc** <<update> **# Python** **PYTHONPATH=$SPARK_HOME/python:$SPARK_HOME/python/lib/py4j-0.9-src.zip:$SPARK_HOME/python/spark-riak-connector/connector/python/dist/pyspark_riak-1.6.3-py3.4.egg:$PYTHONPATH** **export PYTHONPATH**	Add to PYTHONPATH
[administrator@dev-cdm-spark0* ~]$ **pyspark** **Import pyspark_riak**	Validate
★ **Riak Python Client** http://basho.github.io/riak-python-client/index.html	Riak Python Client
[administrator@dev-cdm-spark0* ~]$ **sudo pip3 install riak**	Install

[administrator@dev-cdm-spark0* ~]$
pyspark
from riak import RiakClient, RiakNode
RiakClient()
	Validate

★ **scala**	
[administrator@dev-cdm-spark0* ~]$ **wget "http://repo1.maven.org/maven2/com/basho/riak/spark-riak-connector_2.10/1.6.3/spark-riak-connector_2.10-1.6.3-uber.jar"** **sudo chown hadoop.hadoop spark-riak-connector_2.10-1.6.3-uber.jar** **sudo mv spark-riak-connector_2.10-1.6.3-uber.jar /usr/local/spark/lib**	Install it
https://github.com/basho/spark-riak-connector/blob/master/docs/using-connector.md#using-the-spark-riak-connector	Spark Conf BASHO Riak Properties
[administrator@dev-cdm-spark0* ~]$ **cd /usr/local/spark/conf** **sudo cp spark-defaults.conf.template spark-defaults.conf** **sudo chown hadoop.hadoop spark-defaults.conf** **sudo vi /usr/local/spark/conf/spark-defaults.conf** <<update>> **# BASHO RIAK** **spark.riak.connection.host dev-cdm-riak01.nexjqa.local:8098** **spark.jars /usr/local/spark/lib/spark-riak-connector_2.10-1.6.3-uber.jar**	Create CONF from template

[administrator@dev-cdm-spark0* ~]$ **cd $HOME && spark-shell**
/* Required Import */
import org.apache.spark.sql.SaveMode
import java.sql.Timestamp
import com.basho.riak.spark.rdd.connector.RiakConnector
 	
/* Setup the Spark Context (sc is created for you) */
	Validate

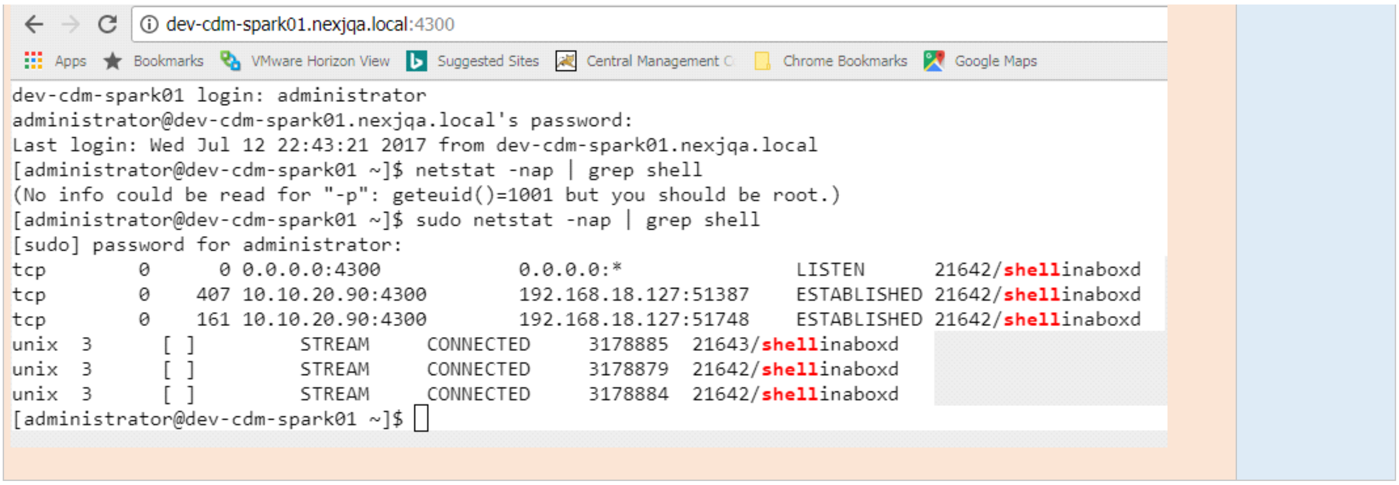
```
val sqlContext = new org.apache.spark.sql.SQLContext(sc)
import sqlContext.implicits._

/* Create an RDD */
val testRDD = sc.parallelize(Seq(
  (1, "f", Timestamp.valueOf("1980-1-1 10:00:00"), "v1"),
  (1, "f", Timestamp.valueOf("1980-1-1 10:10:00"), "v2"),
  (1, "f", Timestamp.valueOf("1980-1-1 10:20:00"), "v3")))

/* Convert to DataFrame (yuck!) */
val df = testRDD.toDF("k", "family", "ts", "value")
df.printSchema()

/* Create a TS Table */
val tableName = "ts_table_c"
val connector = RiakConnector(sc.getConf)

connector.withSessionDo(session =>{
        val request = new com.basho.riak.client.api.commands.timeseries.Query.Builder(
          s"""
            |  CREATE TABLE $tableName  (
            |     k      SINT64    not null,
            |     family  VARCHAR   not null,
            |     ts      TIMESTAMP not null,
            |     value   VARCHAR,
            |
            |     primary key ((k, family, quantum(ts,1,h)), k, family, ts)
            |  )
            |
          """.stripMargin)
          .build()

val response = session.execute(request)})


/* Write the Data Frame */
df.write.format("org.apache.spark.sql.riak").mode(SaveMode.Append).save(tableName)
/* Validate with query */
val tableName = "ts_table_c"
val test_query = "ts >= CAST('1980-1-1 10:00:00' AS TIMESTAMP) AND ts <= CAST('1980-1-1 10:30:00' AS TIMESTAMP) AND k = 1
AND family = 'f'"

val df2 = sqlContext.read.format("org.apache.spark.sql.riak").load(tableName).filter(test_query)

df.toJSON.collect.foreach(println)
```

Utilities

SSH Web Client (Terminal on port 4200) Shell in a Box

https://www.unixmen.com/shellinabox-a-web-based-ajax-terminal-emulator/

[administrator@dev-cdm-spark0* ~]$ **sudo yum install shellinabox** **sudo vi /etc/sysconfig/shellinaboxd** <<updated>> **# Basic options** **USER=shellinabox** **GROUP=shellinabox** **CERTDIR=/var/lib/shellinabox** **PORT=4300** **OPTS="-t -s /:SSH:dev-cdm-spark01/2.nexjqa.local"**	Add it
[administrator@dev-cdm-spark0* ~]$ **sudo systemctl enable shellinaboxd** **sudo systemctl start shellinaboxd**	Enable & Restart service
http://dev-cdm-spark01.nexjqa.local:4300/	Validate it



```
dev-cdm-spark01 login: administrator
administrator@dev-cdm-spark01.nexjqa.local's password:
Last login: Wed Jul 12 22:43:21 2017 from dev-cdm-spark01.nexjqa.local
[administrator@dev-cdm-spark01 ~]$ netstat -nap | grep shell
(No info could be read for "-p": geteuid()=1001 but you should be root.)
[administrator@dev-cdm-spark01 ~]$ sudo netstat -nap | grep shell
[sudo] password for administrator:
tcp        0      0 0.0.0.0:4300            0.0.0.0:*               LISTEN      21642/shellinaboxd
tcp        0    407 10.10.20.90:4300        192.168.18.127:51387    ESTABLISHED 21642/shellinaboxd
tcp        0    161 10.10.20.90:4300        192.168.18.127:51748    ESTABLISHED 21642/shellinaboxd
unix  3      [ ]         STREAM     CONNECTED     3178885  21643/shellinaboxd
unix  3      [ ]         STREAM     CONNECTED     3178879  21642/shellinaboxd
unix  3      [ ]         STREAM     CONNECTED     3178884  21642/shellinaboxd
[administrator@dev-cdm-spark01 ~]$ 
```

From <http://www.aodba.com/how-to-install-apache-spark-in-centos-standalone/>