13: Docker Tutorial: Apache Spark (spark-shell & pyspark) on Cloudera quickstart

Posted on May 30, 2019

This extends Docker Tutorial: BigData on Cloudera quickstart via Docker.

Step 1: Run the container on a command line.

```
~/projects/docker-hadoop]$ docker run --hostname=qu
--privileged=true -t -i -v /Users/arulkumarankumar
--publish-all=true -p 8888:8888 -p 80:80 -p 7180:7
cloudera/quickstart /usr/bin/docker-quickstart
```

300+ Java Interview FAQs

300+ Java FAQs

16+ Java Key Areas Q&As

150+ Java Architect FAQs

80+ Java Code Quality Q&As

150+ Java Coding Q&As

300+ Big Data Interview FAQs

300+ Big Data FAQs

Tutorials - Big Data

TUT - Starting Big Data

TUT - Starting Spark & Scala

Step 2: Create a text file "**simple.txt**" in the local file system.

```
[root@quickstart /]# touch simple.txt
[root@quickstart /]# vi simple.txt
```

```
A big brown fox jumped over a big brown fence.
```

Step 3: Copy this file onto HDFS file system.

```
[root@quickstart /]# hdfs dfs -copyFromLocal simpl
```

List the file that was copied to HDFS:

```
[root@quickstart /]# hdfs dfs -ls /user/root/
Found 1 items
-rw-r--r--   1 root supergroup         47 2019-05-
[root@quickstart /]#
```

spark-submit

Step 4: Run the Spark job in the Spark examples.

Cloudera quickstart comes with some Spark examples packaged as a jar file.

```
[root@quickstart /]# ls -ltr /usr/lib/spark/exampl
total 4
lrwxrwxrwx 1 root root 64 Apr  6  2016 spark-examp
lrwxrwxrwx 1 root root 23 Apr  6  2016 python.tar.
```

Look for the "**JavaWordCount**".

```
[root@quickstart /]# jar -tvf /usr/lib/spark/examp
  1326 Wed Mar 23 11:54:48 UTC 2016 org/apache/spa
```

TUT - Starting with Python
TUT - Kafka
TUT - Pig
TUT - Apache Storm
TUT - Spark Scala on Zeppelin
TUT - Cloudera
TUT - Cloudera on Docker
TUT - File Formats
TUT - Spark on Docker
TUT - Flume
TUT - Hadoop (HDFS)
TUT - HBase (NoSQL)
TUT - Hive (SQL)
TUT - Hadoop & Spark
TUT - MapReduce
TUT - Spark and Scala
TUT - Spark & Java
TUT - PySpark on Databricks
TUT - Zookeeper

800+ Java Interview Q&As

300+ Core Java Q&As
300+ Enterprise Java Q&As
150+ Java Frameworks Q&As
120+ Companion Tech Q&As
Tutorials - Enterprise Java

```
  3387 Wed Mar 23 11:54:48 UTC 2016 org/apache/spa
  1181 Wed Mar 23 11:54:48 UTC 2016 org/apache/spa

```

```
[root@quickstart /]# spark-submit --class org.apa
...
...
fox: 1
fence.: 1
a: 1
big: 2
A: 1
over: 1
brown: 2
jumped: 1

```

spark-shell

```
[root@quickstart /]# spark-shell
SLF4J: Class path contains multiple SLF4J binding
SLF4J: Found binding in [jar:file:/usr/lib/zookee
SLF4J: Found binding in [jar:file:/usr/jars/slf4j
SLF4J: See http://www.slf4j.org/codes.html#multip
SLF4J: Actual binding is of type [org.slf4j.impl.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLev
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Scala version 2.10.5 (Java HotSpot(TM) 64-B
Type in expressions to have them evaluated.
Type :help for more information.
19/05/29 13:45:32 WARN util.NativeCodeLoader: Una
Spark context available as sc (master = local[*],
19/05/29 13:45:35 WARN DataNucleus.General: Plugi
19/05/29 13:45:35 WARN DataNucleus.General: Plugi
19/05/29 13:45:41 WARN metastore.ObjectStore: Ver
19/05/29 13:45:41 WARN metastore.ObjectStore: Fai
19/05/29 13:45:43 WARN shortcircuit.DomainSocketF
SQL context available as sqlContext.

scala>

```

Interactive Spark code in Scala. The SparkContext will be available as "**sc**".

```
scala> var map = sc.textFile("hdfs:///user/root/s
map: org.apache.spark.rdd.RDD[(String, Int)] = Ma

scala> var counts = map.reduceByKey(_ + _)
counts: org.apache.spark.rdd.RDD[(String, Int)] =

scala> counts.foreach(println)
(big,2)
(fox,1)
(fence.,1)
(over,1)
(a,1)
(brown,2)
(A,1)
(jumped,1)

```

pyspark shell

```
[root@quickstart /]# pyspark
Python 2.6.6 (r266:84292, Jul 23 2015, 15:22:56)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-11)] on linux2
Type "help", "copyright", "credits" or "license"
SLF4J: Class path contains multiple SLF4J binding
SLF4J: Found binding in [jar:file:/usr/lib/zookee
SLF4J: Found binding in [jar:file:/usr/jars/slf4j
SLF4J: See http://www.slf4j.org/codes.html#multip
SLF4J: Actual binding is of type [org.slf4j.impl.
19/05/29 13:57:40 INFO spark.SparkContext: Runnin
19/05/29 13:57:40 WARN util.NativeCodeLoader: Una
19/05/29 13:57:41 INFO spark.SecurityManager: Cha
19/05/29 13:57:41 INFO spark.SecurityManager: Cha
19/05/29 13:57:41 INFO spark.SecurityManager: Sec
19/05/29 13:57:41 INFO util.Utils: Successfully s
19/05/29 13:57:41 INFO slf4j.Slf4jLogger: Slf4jLo
19/05/29 13:57:41 INFO Remoting: Starting remotin
19/05/29 13:57:41 INFO Remoting: Remoting started
19/05/29 13:57:41 INFO Remoting: Remoting now lis
19/05/29 13:57:41 INFO util.Utils: Successfully s
19/05/29 13:57:41 INFO spark.SparkEnv: Registerin
19/05/29 13:57:41 INFO spark.SparkEnv: Registerin
19/05/29 13:57:41 INFO storage.DiskBlockManager:
19/05/29 13:57:41 INFO storage.MemoryStore: Memor
19/05/29 13:57:41 INFO spark.SparkEnv: Registerin
19/05/29 13:57:41 INFO server.Server: jetty-8.y.z
19/05/29 13:57:42 INFO server.AbstractConnector:
```

```
19/05/29 13:57:42 INFO util.Utils: Successfully st
19/05/29 13:57:42 INFO ui.SparkUI: Started SparkUI
19/05/29 13:57:42 INFO executor.Executor: Starting
19/05/29 13:57:42 INFO util.Utils: Successfully st
19/05/29 13:57:42 INFO netty.NettyBlockTransferSer
19/05/29 13:57:42 INFO storage.BlockManagerMaster
19/05/29 13:57:42 INFO storage.BlockManagerMasterE
19/05/29 13:57:42 INFO storage.BlockManagerMaster
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Python version 2.6.6 (r266:84292, Jul 23 201
SparkContext available as sc, HiveContext availab
>>>
```

```
Using Python version 2.6.6 (r266:84292, Jul 23 201
SparkContext available as sc, HiveContext availab
>>> input_file = sc.textFile("hdfs:///user/root/st
19/05/29 14:00:11 INFO storage.MemoryStore: Block
19/05/29 14:00:11 INFO storage.MemoryStore: Block
19/05/29 14:00:11 INFO storage.BlockManagerInfo: A
19/05/29 14:00:11 INFO spark.SparkContext: Created
>>> map = input_file.flatMap(lambda line: line.spl
>>> counts = map.reduceByKey(lambda a, b: a + b)
19/05/29 14:00:54 WARN shortcircuit.DomainSocketFa
19/05/29 14:00:54 INFO mapred.FileInputFormat: Tot
>>> for x in counts.collect():
...     print(x)
...
(u'A', 1)
(u'a', 1)
(u'over', 1)
(u'fox', 1)
(u'brown', 2)
(u'jumped', 1)
(u'big', 2)
(u'fence.', 1)
>>>
```

Writing the output to another file “**simple-out.txt**“.

```
>>> counts.saveAsTextFile("hdfs:///user/root/simple
```

Quit the interactive shell with “quit()”.

List the "**simple-out.txt**" folder.

```
root@quickstart /]# hdfs dfs -ls hdfs:///user/root/
Found 3 items
-rw-r--r--   1 root supergroup          0 2019-05-2
-rw-r--r--   1 root supergroup         45 2019-05-2
-rw-r--r--   1 root supergroup         56 2019-05-2
[root@quickstart /]#
```

View the "**simple-out.txt**"

```
[root@quickstart /]# hdfs dfs -cat hdfs:///user/ro
(u'A', 1)
(u'a', 1)
(u'over', 1)
(u'fox', 1)
[root@quickstart /]# hdfs dfs -cat hdfs:///user/ro
(u'brown', 2)
(u'jumped', 1)
(u'big', 2)
(u'fence.', 1)
[root@quickstart /]#
```

‹ 12: Docker Tutorial: Hadoop Big Data configuration files on Cloudera quickstart

14: Docker Tutorial: Hive (via beeline) on Cloudera quickstart ›

Disclaimer

The contents in this Java-Success are copyrighted and from EmpoweringTech pty ltd. The EmpoweringTech pty ltd has the right to correct or enhance the current content without any prior notice. These are general advice only, and one needs to take his/her own circumstances into consideration. The EmpoweringTech pty ltd will not be held liable for any damages caused or alleged to be caused either directly or indirectly by these materials and resources. Any trademarked names or labels used in this blog remain the property of their respective trademark owners. Links to external sites do not imply endorsement of the linked-to sites. Privacy Policy

© 2022 java-success.com