

Java-Success.com

Prepare to fast-track, choose & go places with 800+ Java & Big Data Q&As with lots of code & diagrams.

[Home](#) [Why? ▾](#) [300+ Java FAQs ▾](#) [300+ Big Data FAQs ▾](#) [Courses ▾](#)

[Membership ▾](#) [Your Career ▾](#)

[Home](#) > [bigdata-success.com](#) > [Tutorials - Big Data](#) > [TUT - Cloudera on Docker](#) > 16:

Docker Tutorial: Apache Spark (spark-shell) & parquet-tools – csv to parquet on
Cloudera quickstart

16: Docker Tutorial: Apache Spark (spark- shell) & parquet-tools – csv to parquet on Cloudera quickstart

 Posted on [June 3, 2019](#)

This extends [Docker Tutorial: BigData on Cloudera quickstart via Docker](#).

Step 1: Run the container on a command line.

```
1 ~/projects/docker-hadoop]$ docker run --hostname=q
2 --privileged=true -t -i -v /Users/arulkumarankumar
3 --publish-all=true -p 8888:8888 -p 80:80 -p 7180:71
```

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

Get parquet-tools

Step 2: Install **wget**. The “**uname -a**” gets you the info of the kernel.

```
1 [root@quickstart /]# sudo yum install wget
```

Step 3: Download “parquet-tools” from maven repository using wget.

← → ↺ <https://repo.maven.apache.org/maven2/org/apache/parquet/parquet-tools/1.9.0/>

org/apache/parquet/parquet-tools/1.9.0

../		
parquet-tools-1.9.0-javadoc.jar	2016-10-19 01:17	197543
parquet-tools-1.9.0-javadoc.jar.asc	2016-10-19 01:17	819
parquet-tools-1.9.0-javadoc.jar.md5	2016-10-19 01:17	32
parquet-tools-1.9.0-javadoc.jar.sha1	2016-10-19 01:17	40
parquet-tools-1.9.0-sources.jar	2016-10-19 01:17	41216
parquet-tools-1.9.0-sources.jar.asc	2016-10-19 01:17	819
parquet-tools-1.9.0-sources.jar.md5	2016-10-19 01:17	32
parquet-tools-1.9.0-sources.jar.sha1	2016-10-19 01:17	40
parquet-tools-1.9.0-tests.jar	2016-10-19 01:17	9162
parquet-tools-1.9.0-tests.jar.asc	2016-10-19 01:17	819
parquet-tools-1.9.0-tests.jar.md5	2016-10-19 01:17	32
parquet-tools-1.9.0-tests.jar.sha1	2016-10-19 01:17	40
parquet-tools-1.9.0.jar	2016-10-19 01:17	23897870
parquet-tools-1.9.0.jar.asc	2016-10-19 01:17	819
parquet-tools-1.9.0.jar.md5	2016-10-19 01:17	32
parquet-tools-1.9.0.jar.sha1	2016-10-19 01:17	40
parquet-tools-1.9.0.pom	2016-10-19 01:17	4754
parquet-tools-1.9.0.pom.asc	2016-10-19 01:17	819
parquet-tools-1.9.0.pom.md5	2016-10-19 01:17	32
parquet-tools-1.9.0.pom.sha1	2016-10-19 01:17	40

parquet-tools at Maven repository

```
1 [root@quickstart /]# wget https://repo.maven.apache.org/maven2/org/apache/parquet/parquet-tools/1.9.0/parquet-tools-1.9.0.jar
```

Hive table over .csv

Step 4: Create a csv file “**employee.csv**” in the local file system.

```
1 [root@quickstart /]# touch employee.csv
2 [root@quickstart /]# vi employee.csv
3
1 John, Samuel, IT
```

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

▼

```
2 Peter,Smith,Finance
3 Sean,Mendis,Marketing
```

Step 5: Copy this file onto HDFS file system.

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

spark-shell

Step 6: Download Spark-csv from maven repository for Scala version 2.10.

```
1 [root@quickstart /]# wget https://repo.maven.apache
2
3 [root@quickstart /]# wget https://repo.maven.apache
4
```

Step 7: spark-shell is useful in interactive coding in Scala.

```
1 [root@quickstart /]# spark-shell --jars spark-csv_
2 ....
3 scala>
4
```

```
1 [root@quickstart /]# ls -ltr
2 ....
3 -rw-r--r--  1 root root  165361 Sep  5  2016 spa
4 -rw-r--r--  1 root root 23897870 Oct 19  2016 par
5 -rw-r--r--  1 root root   42400 Sep 19  2018 com
6 ...
7
```

Alternatively you can use **–packages**:

```
1 [root@quickstart /]# spark-shell --packages com.da
```

Step 8: Scala code to read from a csv file into a dataframe.

```

1
2 scala> val dfCsv = sqlContext.read.format("csv").load("hdfs:///user/root/spark-warehouse/employees.csv")
3 scala> dfCsv.show()
4 +-----+-----+-----+
5 | C0 | C1 | C2 |
6 +-----+-----+-----+
7 | John | Samuel | IT |
8 | Peter | Smith | Finance |
9 | Sean | Mendis | Marketing |
10 +-----+-----+-----+
11
12
13 scala> dfCsv.write.parquet("hdfs:///user/root/spark-warehouse/employees.parquet")
14

```

Step 9: Exit out of the spark-shell.

```

1 [root@quickstart /]# hdfs dfs -ls hdfs:///user/root/spark-warehouse/employees.parquet
2 Found 5 items
3 -rw-r--r-- 1 root supergroup 0 2019-06-03 12:06:08 hdfs:///user/root/spark-warehouse/employees.parquet
4 -rw-r--r-- 1 root supergroup 386 2019-06-03 12:06:08 hdfs:///user/root/spark-warehouse/employees.parquet
5 -rw-r--r-- 1 root supergroup 1065 2019-06-03 12:06:08 hdfs:///user/root/spark-warehouse/employees.parquet
6 -rw-r--r-- 1 root supergroup 760 2019-06-03 12:06:08 hdfs:///user/root/spark-warehouse/employees.parquet
7 3-bc74-140ead01143c.gz.parquet
8 -rw-r--r-- 1 root supergroup 760 2019-06-03 12:06:08 hdfs:///user/root/spark-warehouse/employees.parquet
9 3-bc74-140ead01143c.gz.parquet
10 [root@quickstart /]#
11

```

parquet-tools

```

1 [root@quickstart /]# hadoop jar parquet-tools-1.9.0.jar read hdfs:///user/root/spark-warehouse/employees.parquet
2 af3-bc74-140ead01143c.gz.parquet
3
4 19/06/03 12:06:08 INFO hadoop.InternalParquetReader: Reading parquet file hdfs:///user/root/spark-warehouse/employees.parquet
5 C0 = John
6 C1 = Samuel
7 C2 = IT
8
9 C0 = Peter
10 C1 = Smith
11 C2 = Finance
12

```

```
13 [root@quickstart /]#  
14  
  
1 [root@quickstart /]# hadoop jar parquet-tools-1.9.0  
2  
3 19/06/03 12:12:23 INFO hadoop.InternalParquetRecon  
4 C0 = Sean  
5 C1 = Mendis  
6 C2 = Marketing  
7  
8 [root@quickstart /]#  
9
```

◀ 15: Docker Tutorial: Hive & parquet-tools – csv to parquet on Cloudera quickstart

17: Docker Tutorial: sqoop import – 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](#)