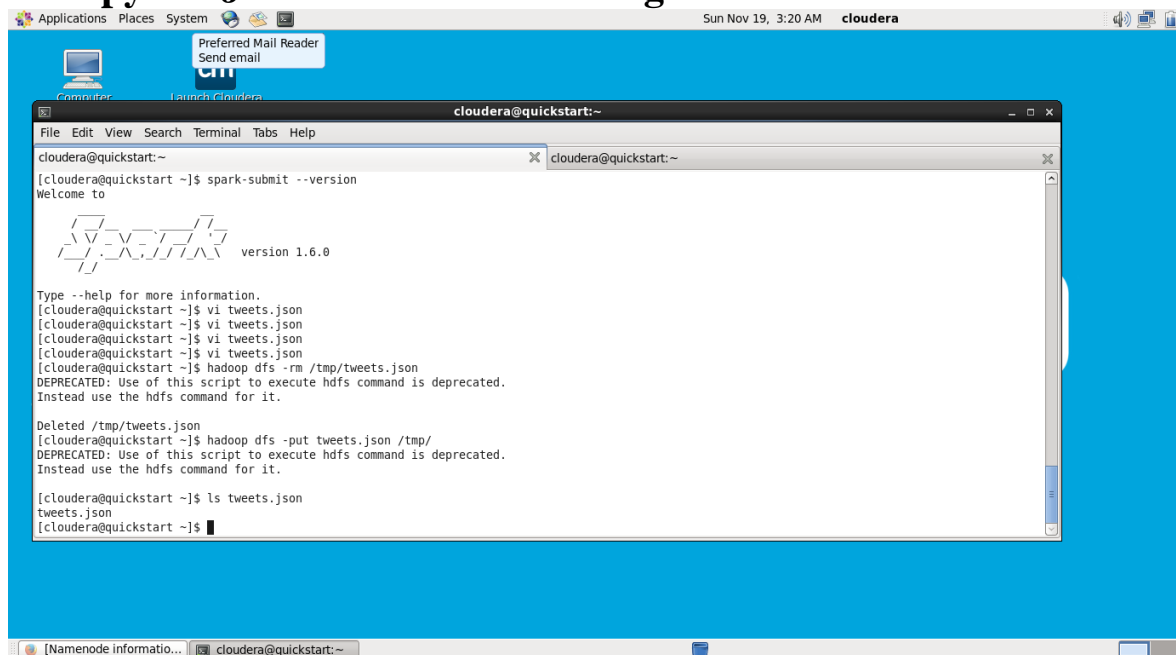


# Counting popular hashtags using spark sql

## 1. Copy the JSON file to HDFS using

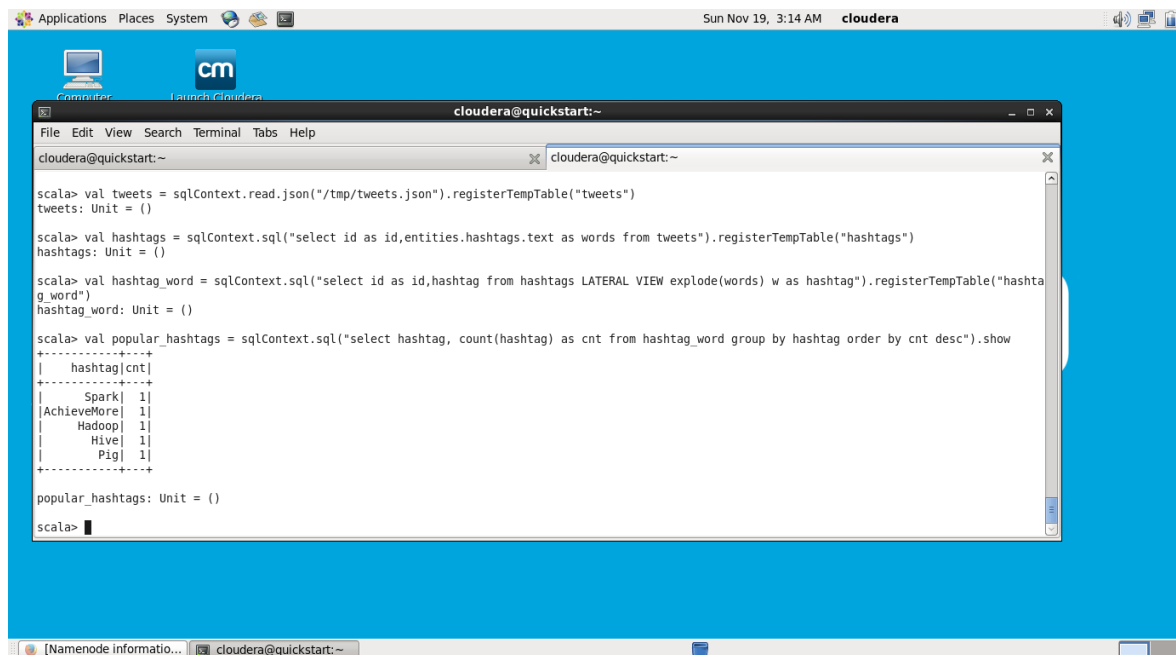


The screenshot shows a terminal window on a Cloudera system. The user is at the prompt `cloudera@quickstart:~`. They run `spark-submit --version`, which displays the Spark version 1.6.0. Then, they run `vi tweets.json` to edit the file. After saving, they run `hadoop dfs -rm /tmp/tweets.json` to delete the local copy. Finally, they run `hadoop dfs -put tweets.json /tmp/` to upload the file to HDFS. The terminal output shows the file being successfully copied to `/tmp/tweets.json`.

```
cloudera@quickstart:~$ spark-submit --version
Welcome to
Spark version 1.6.0
Type --help for more information.
[cloudera@quickstart ~]$ vi tweets.json
[cloudera@quickstart ~]$ vi tweets.json
[cloudera@quickstart ~]$ vi tweets.json
[cloudera@quickstart ~]$ vi tweets.json
[cloudera@quickstart ~]$ vi tweets.json
[cloudera@quickstart ~]$ hadoop dfs -rm /tmp/tweets.json
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Deleted /tmp/tweets.json
[cloudera@quickstart ~]$ hadoop dfs -put tweets.json /tmp/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
[cloudera@quickstart ~]$ ls tweets.json
tweets.json
[cloudera@quickstart ~]$
```

## 2. Source code to

1. To read json file
2. Extract hashtags from file.
3. Split the words and count the number of occurrence hashtags.



The screenshot shows a terminal window on a Cloudera system. The user is at the prompt `scala>`. They run the following code to read the JSON file, extract hashtags, and count their occurrences:

```
scala> val tweets = sqlContext.read.json("/tmp/tweets.json").registerTempTable("tweets")
tweets: Unit = ()

scala> val hashtags = sqlContext.sql("select id as id,entities.hashtags.text as words from tweets").registerTempTable("hashtags")
hashtags: Unit = ()

scala> val hashtag_word = sqlContext.sql("select id as id,hashtag from hashtags LATERAL VIEW explode(words) w as hashtag").registerTempTable("hashtag_word")
hashtag_word: Unit = ()

scala> val popular_hashtags = sqlContext.sql("select hashtag, count(hashtag) as cnt from hashtag_word group by hashtag order by cnt desc").show
+-----+
| hashtag|cnt|
+-----+
| Spark  | 1|
| AchieveMore| 1|
| Hadoop  | 1|
| Hive   | 1|
| Pig    | 1|
+-----+

popular_hashtags: Unit = ()

scala>
```