How to Execute Twitter Trending Topic Extractor

1. Unzip the folder, copy Jar file from the target directory into the Hadoop cluster
2. Execute Jar command. This accepts three parameters

Command Format : java Twitter-Topic-Extractor-jar-with-dependencies.jar [HDFS_Input_DIR] [HDFS_OUTPUT_DIR] [TWITTER_SEARCH_QUERY]

3. Once you provide the input directory and output directory path, it automatically downloads last 6 days of data into 6 different files and executes Map reduce on the same. Please note that both input and output folders are created at runtime and hence mustnt pre-exist.

hadoop jar Twitter-Topic-Extractor-jar-with-dependencies.jar /user/kpw150030/tw_b/ /user/kpw150030/tw_b_op/ "nyse stock"

The topic is kept as configurable and is accepted as the 3rd argument in the input. Data used for searching is top 100 tweets from each of the previous 6 days. These 6 days worth of data is separated into 6 separate input files. Please find trail for execution of Mapreduce algorithm with input as "nyse stock" below.

Execution Trail:

Number of tweets found for date range 2016-02-01 - 2016-02-02 : 100
.....Download complete :/user/kpw150030/tw_b/input_0.txt
Number of tweets found for date range 2016-01-31 - 2016-02-01 : 91
.....Download complete :/user/kpw150030/tw_b/input_1.txt
Number of tweets found for date range 2016-01-30 - 2016-01-31 : 100
...Download complete :/user/kpw150030/tw_b/input_2.txt
Number of tweets found for date range 2016-01-29 - 2016-01-30 : 100
.....Download complete :/user/kpw150030/tw_b/input_3.txt
Number of tweets found for date range 2016-01-28 - 2016-01-29 : 100
.....Download complete :/user/kpw150030/tw_b/input_4.txt
Number of tweets found for date range 2016-01-27 - 2016-01-28 : 100
...Download complete :/user/kpw150030/tw_b/input_5.txt
16/02/02 01:38:24 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
16/02/02 01:38:24 INFO client.RMProxy: Connecting to ResourceManager at cshadoop1.utdallas.edu/10.176.92.71:8032
16/02/02 01:38:24 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
16/02/02 01:38:24 INFO input.FileInputFormat: Total input paths to process : 6
16/02/02 01:38:24 INFO mapreduce.JobSubmitter: number of splits:6

16/02/02 01:38:25 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
16/02/02 01:38:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1451926778124_0513
16/02/02 01:38:25 INFO impl.YarnClientImpl: Submitted application application_1451926778124_0513
16/02/02 01:38:25 INFO mapreduce.Job: The url to track the job: http://cshadoop1.utdallas.edu:8088/proxy/application_1451926778124_0513/
16/02/02 01:38:25 INFO mapreduce.Job: Running job: job_1451926778124_0513
16/02/02 01:38:30 INFO mapreduce.Job: Job job_1451926778124_0513 running in uber mode : false
16/02/02 01:38:30 INFO mapreduce.Job: map 0% reduce 0%
16/02/02 01:38:35 INFO mapreduce.Job: map 67% reduce 0%
16/02/02 01:38:36 INFO mapreduce.Job: map 100% reduce 0%
16/02/02 01:38:40 INFO mapreduce.Job: map 100% reduce 100%
16/02/02 01:38:41 INFO mapreduce.Job: Job job_1451926778124_0513 completed successfully
16/02/02 01:38:41 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=1842
        FILE: Number of bytes written=657424
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=66027
        HDFS: Number of bytes written=884
        HDFS: Number of read operations=21
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=4
        Rack-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=15425
        Total time spent by all reduces in occupied slots (ms)=2463
        Total time spent by all map tasks (ms)=15425
        Total time spent by all reduce tasks (ms)=2463
        Total vcore-seconds taken by all map tasks=15425
        Total vcore-seconds taken by all reduce tasks=2463
        Total megabyte-seconds taken by all map tasks=15795200
        Total megabyte-seconds taken by all reduce tasks=2522112
    Map-Reduce Framework
        Map input records=591
        Map output records=210

```
        Map output bytes=2590
        Map output materialized bytes=1872
        Input split bytes=678
        Combine input records=210
        Combine output records=122
        Reduce input groups=80
        Reduce shuffle bytes=1872
        Reduce input records=122
        Reduce output records=80
        Spilled Records=244
        Shuffled Maps =6
        Failed Shuffles=0
        Merged Map outputs=6
        GC time elapsed (ms)=852
        CPU time spent (ms)=4950
        Physical memory (bytes) snapshot=3045306368
        Virtual memory (bytes) snapshot=8607174656
        Total committed heap usage (bytes)=3575119872
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=65349
    File Output Format Counters
        Bytes Written=884
```

Sample Output

hdfs dfs -get /user/kpw150030/tw_b_op/part-r-00000
view part-r-00000

```
#2      8
#3M     1
#Airlines       1
#American       1
#AmericanIndustrialism  1
#Aviation       1
#BigData        2
```

#BolsadeNuevaYork 2
#Delta 2
#Equi... 2
#Facebook 2
#Finance 5
#Financial 2
#Gellman 1
#GeneralMotorsNews 3
#Marijuana 1
#Market 2
#MarketMover 1
#Marketing 3
#Money 4
#NASDAQ 1
#NYSE 39
#NYSE:ASHR 1
#NYSE:FXI 1
#OTDIH 1
#PennyStocks 2
#Roc 1
#SEO 1
#SMS 3
#ShortSell 6
#Stock 5
#StockCharts 5
#Stocks 6
#TSX 2
#Trading 6
#WEED 2
#XFiles 1
#ai 1
#amazon 1
#amp 1
#amzn 1
#analysts 1
#android 2
#app 2
#auto 1
#bhive 2
#boomtobust 1
#brics 1
#chesapeake 2
#corpgov 2

#crowd 1