BDA LAB DOCUMENTATION FOR ASSIGNMENTS

PERFORMED ON MACINTOSH

Covers the procedures for following **assignments done in Big Data Labs**

PI

WORD COUNT

FLIGHT DATA ANALYSIS

Submitted To:
Prof. Aditya Bharadwaj
Big Data Analytics

Submitted By:
SID - 16103104
Name - Lovedeep Singh
Bachelor of Technology - CSE 4th Year

Note :

Although, these are done in Unix Macintosh with Mojave OS 2019 edition, the steps are generic for any Linux system. Most of the things remain same across flavours of linux, but there may be slight differences in folder locations depending on the flavour or procedure followed while setting up the Hadoop ecosystem, which needs to taken care of separately.

PI

After successful setup of Hadoop, run the following command

```
(base) Lovedeeps-MacBook-Pro:mapreduce lovedeepsingh$ hadoop jar ./hadoop-mapred
uce-examples-3.1.2.jar pi 10 100
```

You will see the following output

```
                Reduce input groups=2
                Reduce shuffle bytes=280
                Reduce input records=20
                Reduce output records=0
                Spilled Records=40
                Shuffled Maps =10
                Failed Shuffles=0
                Merged Map outputs=10
                GC time elapsed (ms)=40
                Total committed heap usage (bytes)=2759852032
        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=1180
        File Output Format Counters
                Bytes Written=97
Job Finished in 2.949 seconds
Estimated value of Pi is 3.14800000000000000000
(base) Lovedeeps-MacBook-Pro:mapreduce lovedeepsingh$
```

WORD COUNT MAP REDUCE

You must have running hadoop setup on your system.

create a directory in hadoop filesystem.

Copy copy some text file to hadoop filesystem inside the input directory. Here I am copying LICENSE.txt to it. You can copy more that one files.

```
(base) Lovedeeps-MacBook-Pro:Cellar lovedeepsingh$ hdfs dfs -mkdir -p /user/hadoop/input

(base) Lovedeeps-MacBook-Pro:Cellar lovedeepsingh$ hdfs dfs -put LICENSE.txt /user/hadoop/i
nput/
```

Now run the wordcount mapreduce example using following command. Below command will read all files from input folder and process with mapreduce jar file. After successful completion of task results will be placed on output directory.

```
2019-09-22 18:58:16,225 INFO mapred.LocalJobRunner: Starting task: attempt_local
1009336737_0001_m_000000_0
2019-09-22 18:58:16,254 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 2
2019-09-22 18:58:16,254 INFO output.FileOutputCommitter: FileOutputCommitter ski
p cleanup _temporary folders under output directory:false, ignore cleanup failur
es: false
2019-09-22 18:58:16,266 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree
 currently is supported only on Linux.
2019-09-22 18:58:16,267 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
 null
2019-09-22 18:58:16,270 INFO mapred.MapTask: Processing split: hdfs://localhost:
9000/user/hadoop/input/LICENSE.txt:0+147145
2019-09-22 18:58:16,415 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2019-09-22 18:58:16,415 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2019-09-22 18:58:16,415 INFO mapred.MapTask: soft limit at 83886080
2019-09-22 18:58:16,415 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2019-09-22 18:58:16,415 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
2019-09-22 18:58:16,422 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
2019-09-22 18:58:16,610 INFO mapred.LocalJobRunner:
2019-09-22 18:58:16,613 INFO mapred.MapTask: Starting flush of map output
2019-09-22 18:58:16,613 INFO mapred.MapTask: Spilling map output
```

Now show the content of result file where you will see the result of wordcount. You will see the count of each word.

```
three-dımensıonaı       1
through 9
time    14
time.   4
time;   2
timed-relation  2
timely  1
title   3
to      466
to,     3
together        1
topography,     1
tort    2
tracking        1
trade   1
trademark       6
trademark)      5
trademark,      1
trademarks,     2
transfer        5
transformation  1
transformed,    2
translated      1
translation     5
```

Flight Data analysis

Download the csv data file from google, say for 2008, we downloaded 2008.csv
Move the file in Hadoop

```
(base) Lovedeeps-MacBook-Pro:sbin lovedeepsingh$ hdfs dfs -put /Users/lovedeepsingh/Downloa
ds/2008.csv input.csv
```

Then run the following command

```
(base) Lovedeeps-MacBook-Pro:sbin lovedeepsingh$ hadoop jar Airline_Project.jar  input.csv
output1 output2 output3
```

We have

```
                Map output bytes=824646
                Map output materialized bytes=1099562
                Input split bytes=612
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=1099562
                Reduce input records=137440
                Reduce output records=2
                Spilled Records=274880
                Shuffled Maps =6
                Failed Shuffles=0
                Merged Map outputs=6
                GC time elapsed (ms)=67
                Total committed heap usage (bytes)=2502950912
        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=689433824
        File Output Format Counters
                Bytes Written=69
```

Finally run the following command

```
(base) Lovedeeps-MacBook-Pro:hadoop lovedeepsingh$ hdfs dfs -cat /user/lovedeepsingh/output
1/*


2019-09-12 15:21:58,906 WARN util.NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Highest probabilty of airlines on schedule      0.0
AQ      0.75386995
HA      0.6939162
9E      0.6437548
Lowest probabilty of airlines on schedule       0.0
OH      0.49577072
AA      0.49908623
F9      0.50897425
```