Ryan Ballenger
CSCIE63 Big Data Analytics
Assignment 5

Assignment 5 Solution

Problem 1. Create your own tables KINGJAMES with columns for words and frequencies and insert into the table the result of Hadoop MapReduce GREP program which produce word counts on file all-bible. File is provided with this assignment. Tell us all words in table KINGJAMES which start with letter "w" and are 4 or more characters long and appear more than 250. There are not that many of those words so you can count them by hand. However, you want to be more automated so please change your query so that it gives you the number of such words as its output. When comparing a word with a string your use LIKE operator, like

word like 'a%' or word like '%th%'

Symbol '%' means any number of characters. You measure the length of a string using function length() and you change the case of a word to all lower characters using function lower().

// Create a table in Hue Hive Editor for the word and frequencies data for the file all-bible.
create table KINGJAMES (freq INT, word STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' stored as textfile;

// no results or log provided by the editor

// Hadoop's grep is run on the file all-bible to determine the words and frequencies.

```
[cloudera@quickstart ~]$ hadoop jar
/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar grep all-bible bible_frequency '\w+'
16/03/04 12:55:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/03/04 12:55:33 WARN mapreduce.JobResourceUploader: No job jar file set.  User classes
may not be found. See Job or Job#setJar(String).
16/03/04 12:55:33 INFO input.FileInputFormat: Total input paths to process : 1
16/03/04 12:55:33 INFO mapreduce.JobSubmitter: number of splits:1
16/03/04 12:55:34 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1457069306693_0002
16/03/04 12:55:34 INFO mapred.YARNRunner: Job jar is not present. Not adding any jar to the
list of resources.
```

16/03/04 12:55:35 INFO impl.YarnClientImpl: Submitted application application_1457069306693_0002
16/03/04 12:55:35 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1457069306693_0002/
16/03/04 12:55:35 INFO mapreduce.Job: Running job: job_1457069306693_0002
16/03/04 12:55:52 INFO mapreduce.Job: Job job_1457069306693_0002 running in uber mode : false
16/03/04 12:55:52 INFO mapreduce.Job: map 0% reduce 0%
16/03/04 12:56:11 INFO mapreduce.Job: map 100% reduce 0%
16/03/04 12:56:25 INFO mapreduce.Job: map 100% reduce 100%
16/03/04 12:56:25 INFO mapreduce.Job: Job job_1457069306693_0002 completed successfully
16/03/04 12:56:25 INFO mapreduce.Job: Counters: 49
	File System Counters
		FILE: Number of bytes read=256987
		FILE: Number of bytes written=737855
		FILE: Number of read operations=0
		FILE: Number of large read operations=0
		FILE: Number of write operations=0
		HDFS: Number of bytes read=5258808
		HDFS: Number of bytes written=346447
		HDFS: Number of read operations=6
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=2
	Job Counters
		Launched map tasks=1
		Launched reduce tasks=1
		Data-local map tasks=1
		Total time spent by all maps in occupied slots (ms)=15260
		Total time spent by all reduces in occupied slots (ms)=12120
		Total time spent by all map tasks (ms)=15260
		Total time spent by all reduce tasks (ms)=12120
		Total vcore-seconds taken by all map tasks=15260
		Total vcore-seconds taken by all reduce tasks=12120
		Total megabyte-seconds taken by all map tasks=15626240
		Total megabyte-seconds taken by all reduce tasks=12410880
	Map-Reduce Framework
		Map input records=117154
		Map output records=894145
		Map output bytes=11562331
		Map output materialized bytes=256987
		Input split bytes=120
		Combine input records=894145

```
		Combine output records=14330
		Reduce input groups=14330
		Reduce shuffle bytes=256987
		Reduce input records=14330
		Reduce output records=14330
		Spilled Records=28660
		Shuffled Maps =1
		Failed Shuffles=0
		Merged Map outputs=1
		GC time elapsed (ms)=348
		CPU time spent (ms)=6780
		Physical memory (bytes) snapshot=347787264
		Virtual memory (bytes) snapshot=3008647168
		Total committed heap usage (bytes)=226562048
	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=5258688
	File Output Format Counters
		Bytes Written=346447
16/03/04 12:56:26 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/03/04 12:56:26 WARN mapreduce.JobResourceUploader: No job jar file set.  User classes
may not be found. See Job or Job#setJar(String).
16/03/04 12:56:26 INFO input.FileInputFormat: Total input paths to process : 1
16/03/04 12:56:26 INFO mapreduce.JobSubmitter: number of splits:1
16/03/04 12:56:26 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1457069306693_0003
16/03/04 12:56:26 INFO mapred.YARNRunner: Job jar is not present. Not adding any jar to the
list of resources.
16/03/04 12:56:26 INFO impl.YarnClientImpl: Submitted application
application_1457069306693_0003
16/03/04 12:56:26 INFO mapreduce.Job: The url to track the job:
http://quickstart.cloudera:8088/proxy/application_1457069306693_0003/
16/03/04 12:56:26 INFO mapreduce.Job: Running job: job_1457069306693_0003
16/03/04 12:56:42 INFO mapreduce.Job: Job job_1457069306693_0003 running in uber mode :
false
16/03/04 12:56:42 INFO mapreduce.Job:  map 0% reduce 0%
16/03/04 12:57:00 INFO mapreduce.Job:  map 100% reduce 0%
```

```
16/03/04 12:57:20 INFO mapreduce.Job:  map 100% reduce 100%
16/03/04 12:57:20 INFO mapreduce.Job: Job job_1457069306693_0003 completed
successfully
16/03/04 12:57:20 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=256987
                FILE: Number of bytes written=736843
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=346591
                HDFS: Number of bytes written=147408
                HDFS: Number of read operations=7
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=15018
                Total time spent by all reduces in occupied slots (ms)=16862
                Total time spent by all map tasks (ms)=15018
                Total time spent by all reduce tasks (ms)=16862
                Total vcore-seconds taken by all map tasks=15018
                Total vcore-seconds taken by all reduce tasks=16862
                Total megabyte-seconds taken by all map tasks=15378432
                Total megabyte-seconds taken by all reduce tasks=17266688
        Map-Reduce Framework
                Map input records=14330
                Map output records=14330
                Map output bytes=228321
                Map output materialized bytes=256987
                Input split bytes=144
                Combine input records=0
                Combine output records=0
                Reduce input groups=614
                Reduce shuffle bytes=256987
                Reduce input records=14330
                Reduce output records=14330
                Spilled Records=28660
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
```

```
		GC time elapsed (ms)=326
		CPU time spent (ms)=6930
		Physical memory (bytes) snapshot=351039488
		Virtual memory (bytes) snapshot=3008765952
		Total committed heap usage (bytes)=226562048
	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=346447
	File Output Format Counters
		Bytes Written=147408


// In Hue, the resulting words and frequencies are loaded into the KINGJAMES table.

LOAD DATA INPATH "/user/cloudera/bible_frequency/part-r-00000" INTO TABLE KINGJAMES;
INFO  : Loading data to table default.kingjames from
hdfs://quickstart.cloudera:8020/user/cloudera/bible_frequency/part-r-00000
INFO  : Table default.kingjames stats: [numFiles=1, totalSize=147408]

// A query is run to count the words starting with ‘w’ that have a length of 4 or more and occur
// more than 250 times.

select count(*)
from kingjames
where LOWER(word) LIKE "w%"
AND freq > 250
AND LENGTH(word) >= 4;

// Result:
28

// Log from query above:
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
```

```
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : Starting Job = job_1457132565553_0003, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1457132565553_0003/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1457132565553_0003
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2016-03-04 15:15:17,231 Stage-1 map = 0%,  reduce = 0%
INFO  : 2016-03-04 15:15:38,428 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.06
sec
INFO  : 2016-03-04 15:15:57,928 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.67
sec
INFO  : MapReduce Total cumulative CPU time: 6 seconds 670 msec
INFO  : Ended Job = job_1457132565553_0003

// Next, the same query is run but the actual word list is retrieved.
select *
from kingjames
where LOWER(word) LIKE "w%"
AND freq > 250
AND LENGTH(word) >= 4;

// No Logs provided by Hue for the above query

// Result:
        kingjames.freq          kingjames.word
0       6057    with
1       4297    which
2       3819    will
3       2767    were
4       2487    when
5       1399    went
6       732     whom
7       694     word
8       652     what
9       546     words
10      512     work
11      443     would
12      436     without
13      407     wife
14      396     water
15      355     woman
16      349     When
17      343     wicked
```

18	335	What
19	335	where
20	304	wilderness
21	301	works
22	288	world
23	286	waters
24	284	whose
25	283	written
26	261	Wherefore
27	253	well

Problem 2. Create your own table SHAKE similar to the one we used in class and populate it with results of MapReduce GREP program applied to the file all-shakespeare which is provided with this assignment. Create your own MERGED table similar to the one we used in class. The table will list all the word and the frequencies with which they appear in either table SHAKE or KINGJAMES. Your table will be "better" than the one we used in class. In class we only inserted into that table words that appear in both texts. Please use **outer joins** to populate the table with words that also appear in one but not the other text. Tell us how many words appear in table SHAKE but not in KINGJAMES and how many appear in KINGJAMES and not in SHAKE. Select 10 words from each group for us. To solve this problem you will have to consult Hive Tutorial at https://cwiki.apache.org/confluence/display/Hive/Tutorial or simply Google around the Web.

```
// In Hue Hive Editor, I create the SHAKE table to store the words and frequencies
// of the text in the all-shakespeare file
create table SHAKE (freq INT, word STRING) ROW FORMAT DELIMITED FIELDS
TERMINATED BY '\t' stored as textfile;

// No Logs or Result provided by Hue


// Hadoop's grep runs on the file all-shakespeare to determine the words and counts

[cloudera@quickstart ~]$ hadoop jar
/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar grep all-shakespeare shake_freq
'\w+'
16/03/04 14:02:59 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/03/04 14:03:00 WARN mapreduce.JobResourceUploader: No job jar file set.  User classes
may not be found. See Job or Job#setJar(String).
16/03/04 14:03:00 INFO input.FileInputFormat: Total input paths to process : 1
16/03/04 14:03:01 INFO mapreduce.JobSubmitter: number of splits:1
```

16/03/04 14:03:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1457128255540_0003
16/03/04 14:03:01 INFO mapred.YARNRunner: Job jar is not present. Not adding any jar to the list of resources.
16/03/04 14:03:02 INFO impl.YarnClientImpl: Submitted application application_1457128255540_0003
16/03/04 14:03:02 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1457128255540_0003/
16/03/04 14:03:02 INFO mapreduce.Job: Running job: job_1457128255540_0003
16/03/04 14:03:17 INFO mapreduce.Job: Job job_1457128255540_0003 running in uber mode : false
16/03/04 14:03:17 INFO mapreduce.Job: map 0% reduce 0%
16/03/04 14:03:32 INFO mapreduce.Job: map 100% reduce 0%
16/03/04 14:03:46 INFO mapreduce.Job: map 100% reduce 100%
16/03/04 14:03:46 INFO mapreduce.Job: Job job_1457128255540_0003 completed successfully
16/03/04 14:03:46 INFO mapreduce.Job: Counters: 49
	File System Counters
		FILE: Number of bytes read=525117
		FILE: Number of bytes written=1274127
		FILE: Number of read operations=0
		FILE: Number of large read operations=0
		FILE: Number of write operations=0
		HDFS: Number of bytes read=5284357
		HDFS: Number of bytes written=707255
		HDFS: Number of read operations=6
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=2
	Job Counters
		Launched map tasks=1
		Launched reduce tasks=1
		Data-local map tasks=1
		Total time spent by all maps in occupied slots (ms)=13877
		Total time spent by all reduces in occupied slots (ms)=10666
		Total time spent by all map tasks (ms)=13877
		Total time spent by all reduce tasks (ms)=10666
		Total vcore-seconds taken by all map tasks=13877
		Total vcore-seconds taken by all reduce tasks=10666
		Total megabyte-seconds taken by all map tasks=14210048
		Total megabyte-seconds taken by all reduce tasks=10921984
	Map-Reduce Framework
		Map input records=173126
		Map output records=964453

```
		Map output bytes=12641942
		Map output materialized bytes=525117
		Input split bytes=126
		Combine input records=964453
		Combine output records=29183
		Reduce input groups=29183
		Reduce shuffle bytes=525117
		Reduce input records=29183
		Reduce output records=29183
		Spilled Records=58366
		Shuffled Maps =1
		Failed Shuffles=0
		Merged Map outputs=1
		GC time elapsed (ms)=283
		CPU time spent (ms)=7910
		Physical memory (bytes) snapshot=347222016
		Virtual memory (bytes) snapshot=3008622592
		Total committed heap usage (bytes)=226562048
	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=5284231
	File Output Format Counters
		Bytes Written=707255
16/03/04 14:03:47 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/03/04 14:03:47 WARN mapreduce.JobResourceUploader: No job jar file set.  User classes
may not be found. See Job or Job#setJar(String).
16/03/04 14:03:47 INFO input.FileInputFormat: Total input paths to process : 1
16/03/04 14:03:47 INFO mapreduce.JobSubmitter: number of splits:1
16/03/04 14:03:47 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1457128255540_0004
16/03/04 14:03:47 INFO mapred.YARNRunner: Job jar is not present. Not adding any jar to the
list of resources.
16/03/04 14:03:47 INFO impl.YarnClientImpl: Submitted application
application_1457128255540_0004
16/03/04 14:03:47 INFO mapreduce.Job: The url to track the job:
http://quickstart.cloudera:8088/proxy/application_1457128255540_0004/
16/03/04 14:03:47 INFO mapreduce.Job: Running job: job_1457128255540_0004
```

16/03/04 14:04:00 INFO mapreduce.Job: Job job_1457128255540_0004 running in uber mode : false
16/03/04 14:04:00 INFO mapreduce.Job: map 0% reduce 0%
16/03/04 14:04:12 INFO mapreduce.Job: map 100% reduce 0%
16/03/04 14:04:25 INFO mapreduce.Job: map 100% reduce 100%
16/03/04 14:04:26 INFO mapreduce.Job: Job job_1457128255540_0004 completed successfully
16/03/04 14:04:27 INFO mapreduce.Job: Counters: 49
	File System Counters
		FILE: Number of bytes read=525117
		FILE: Number of bytes written=1273093
		FILE: Number of read operations=0
		FILE: Number of large read operations=0
		FILE: Number of write operations=0
		HDFS: Number of bytes read=707399
		HDFS: Number of bytes written=299379
		HDFS: Number of read operations=7
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=2
	Job Counters
		Launched map tasks=1
		Launched reduce tasks=1
		Data-local map tasks=1
		Total time spent by all maps in occupied slots (ms)=10663
		Total time spent by all reduces in occupied slots (ms)=10078
		Total time spent by all map tasks (ms)=10663
		Total time spent by all reduce tasks (ms)=10078
		Total vcore-seconds taken by all map tasks=10663
		Total vcore-seconds taken by all reduce tasks=10078
		Total megabyte-seconds taken by all map tasks=10918912
		Total megabyte-seconds taken by all reduce tasks=10319872
	Map-Reduce Framework
		Map input records=29183
		Map output records=29183
		Map output bytes=466745
		Map output materialized bytes=525117
		Input split bytes=144
		Combine input records=0
		Combine output records=0
		Reduce input groups=631
		Reduce shuffle bytes=525117
		Reduce input records=29183
		Reduce output records=29183

```
		Spilled Records=58366
		Shuffled Maps =1
		Failed Shuffles=0
		Merged Map outputs=1
		GC time elapsed (ms)=269
		CPU time spent (ms)=5520
		Physical memory (bytes) snapshot=343154688
		Virtual memory (bytes) snapshot=3008622592
		Total committed heap usage (bytes)=226562048
	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=707255
	File Output Format Counters
		Bytes Written=299379
```

// The resulting word counts are loaded into the shakespeare table named SHAKE

```
LOAD DATA INPATH "/user/cloudera/shake_freq/part-r-00000" INTO TABLE shake;
```

```
INFO  : Loading data to table default.shake from
hdfs://quickstart.cloudera:8020/user/cloudera/shake_freq/part-r-00000
INFO  : Table default.shake stats: [numFiles=1, totalSize=299379]
```

// I create a merged table for the upcoming join of the kingjames and shake tables

```
create table merged (wordKJ STRING, wordS STRING, freqKJ INT, freqS INT) ROW FORMAT
DELIMITED FIELDS TERMINATED BY '\t' stored as textfile;
```

// No Results or Logs

// Insert into the merged table an outer join on the rows' words being equivalent. The rows
// are from the SHAKE and KINGJAMES tables.

```
INSERT INTO merged
select K.word, S.word, K.freq, S.freq
from shake S FULL OUTER JOIN kingjames K
ON S.word = K.word;
```

INFO : Number of reduce tasks not specified. Estimated from input data size: 1
INFO : In order to change the average load for a reducer (in bytes):
INFO :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO : In order to limit the maximum number of reducers:
INFO :   set hive.exec.reducers.max=<number>
INFO : In order to set a constant number of reducers:
INFO :   set mapreduce.job.reduces=<number>
INFO : Starting Job = job_1457132565553_0006, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1457132565553_0006/
INFO : Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1457132565553_0006
INFO : Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
INFO : 2016-03-04 15:41:26,240 Stage-1 map = 0%, reduce = 0%
INFO : 2016-03-04 15:42:03,683 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 7.49 sec
INFO : 2016-03-04 15:42:04,805 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 8.79
sec
INFO : 2016-03-04 15:42:22,726 Stage-1 map = 100%, reduce = 78%, Cumulative CPU 13.34
sec
INFO : 2016-03-04 15:42:24,111 Stage-1 map = 100%, reduce = 100%, Cumulative CPU
14.43 sec
INFO : MapReduce Total cumulative CPU time: 14 seconds 430 msec
INFO : Ended Job = job_1457132565553_0006
INFO : Loading data to table default.merged from
hdfs://quickstart.cloudera:8020/user/hive/warehouse/merged/.hive-staging_hive_2016-03-04_15
-41-05_929_1200754675307790080-2/-ext-10000
INFO : Table default.merged stats: [numFiles=1, numRows=35758, totalSize=614805,
rawDataSize=579047]

// The words present in all-shakespeare but not in all-bible are determined by the query.
// The count is 21428 and 10 selected words are listed below.

```
select count(*) from merged
where wordKJ is NULL AND wordS is NOT NULL;
```

// Result
21428

// Logs
INFO : Number of reduce tasks determined at compile time: 1
INFO : In order to change the average load for a reducer (in bytes):
INFO :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO : In order to limit the maximum number of reducers:
INFO :   set hive.exec.reducers.max=<number>
INFO : In order to set a constant number of reducers:

INFO : set mapreduce.job.reduces=<number>
INFO : Starting Job = job_1457132565553_0007, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1457132565553_0007/
INFO : Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1457132565553_0007
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2016-03-04 15:43:32,879 Stage-1 map = 0%, reduce = 0%
INFO : 2016-03-04 15:43:50,800 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.5 sec
INFO : 2016-03-04 15:44:06,131 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.62
sec
INFO : MapReduce Total cumulative CPU time: 5 seconds 620 msec
INFO : Ended Job = job_1457132565553_0007

// The same query is run except for finding the results instead of counting them.

```
select *
from merged
where wordKJ is NULL AND wordS is NOT NULL;
```

// No log this time and the selected 10 results from the above query are listed

339	NULL	Antonio	NULL	70
340	NULL	Antonius	NULL	10
341	NULL	Antony	NULL	209
342	NULL	Apace	NULL	1
343	NULL	Apemantus	NULL	24
344	NULL	Apennines	NULL	1
345	NULL	Apollinem	NULL	1
346	NULL	Apollo	NULL	27
347	NULL	Apollodorus	NULL	1
348	NULL	Apothecary	NULL	7
349	NULL	Appals	NULL	2

// Words present in all-bible but not in all-shakespeare are determined by the query below.
// The count is 6575 and 10 words are listed below.

```
select count(*) from merged
where wordKJ is NOT NULL AND wordS is NULL;
```

// Result
6575

// Logs

INFO : Number of reduce tasks determined at compile time: 1
INFO : In order to change the average load for a reducer (in bytes):
INFO :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO : In order to limit the maximum number of reducers:
INFO :   set hive.exec.reducers.max=<number>
INFO : In order to set a constant number of reducers:
INFO :   set mapreduce.job.reduces=<number>
INFO : Starting Job = job_1457132565553_0008, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1457132565553_0008/
INFO : Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1457132565553_0008
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2016-03-04 15:46:09,927 Stage-1 map = 0%, reduce = 0%
INFO : 2016-03-04 15:46:24,877 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.1 sec
INFO : 2016-03-04 15:46:42,660 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.41
sec
INFO : MapReduce Total cumulative CPU time: 5 seconds 410 msec
INFO : Ended Job = job_1457132565553_0008

// The same query is run, without counting, to determine specific word results.

select * from merged
where wordKJ is NOT NULL AND wordS is NULL;

// No logs and selected 10 results for the above query

379	Adoraim	NULL	1	NULL
380	Adoram	NULL	2	NULL
381	Adrammelech	NULL	3	NULL
382	Adramyttium	NULL	1	NULL
383	Adria	NULL	1	NULL
384	Adriel	NULL	2	NULL
385	Adullam	NULL	8	NULL
386	Adullamite	NULL	3	NULL
387	Adummim	NULL	2	NULL
388	Aenon	NULL	1	NULL
389	Afterward	NULL	9	NULL

Problem 3. When you have your three queries for counting common words, words that are present in Bible but not in Shakespeare and the words present in Shakespeare but not in Bible refined and working, collect the execution times of those queries. This is not straightforward, since Hive does not give you a simple tool to time your queries. You can look in query logs (a tab next to the Results tab) and sum execution times of map and reduce jobs. That is close

enough. Then change your Hue Query Editor and switch to Impala Editor. Run your queries in that editor. This time you have no way of read the time. You just make a subjective estimate. Compare the execution time of queries with Impala and Hive. Impala is usually much faster. One thing to notice here is that you can use Impala on some of Hive tables. Unfortunately not all. Hive is more versatile than Impala.

// In Hue Hive Editor, words not in the Bible but in Shakespeare are found.
// The logs are analyzed to determine the amount of time taken.

```
select count(*) from merged
where wordKJ is NULL AND wordS is NOT NULL;
```

// Logs from Hue Hive editor

```
INFO  : 2016-03-03 21:05:36,361 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.61
sec
INFO  : 2016-03-03 21:05:58,311 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.29
sec
INFO  : MapReduce Total cumulative CPU time: 6 seconds 290 msec
```

// Logs from Impala editor for the same query as above.

```
Query 95417f658d3d5037:12ffb95b0a41c1ac 100% Complete (1 out of 1)
```

// Impala time: 2.78 sec

// Comparison: As noted in the handout, the Impala Editor is much faster and finishes in less
// than half the time of the Hue Hive Editor. Specifically Impala finishes in 2.78 seconds while
the
// Hue Hive Editor finishes in 6.290 seconds. The Impala editor has less capabilities but is useful
// for performing small queries quickly.

// Results for both queries

```
21428
```

// The query for words in all-bible but not in all-shakespeare is run.
// The time taken for this query is analyzed below with data from the logs.

```
select count(*) from merged
where wordKJ is NOT NULL AND wordS is NULL;
```

// Log from Hue Hive editor

```
INFO  : 2016-03-03 21:12:22,480 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.58
sec
```

INFO : 2016-03-03 21:12:42,639 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.92 sec
INFO : MapReduce Total cumulative CPU time: 5 seconds 920 msec

// Log from Impala editor
Query e84825f7b08cd585:ab0d7975dd2ec2ab 100% Complete (1 out of 1)
// Impala time: 2.90 sec

// Result for both queries
 6575

// Comparison: As in the first query, the Impala editor finishes much more quickly than the Hue
// Hive Editor. Specifically it finishes in less than half the time (5.920s vs. 2.9s). The Impala
// Editor is better at performing small queries quickly.

Problem 4. Please create Hive table APACHELOG for extraction of the content of Apache server logs:

```
CREATE TABLE apachelog (
host STRING,
identity STRING,
user STRING,
time STRING,
request STRING,
status STRING,
size STRING,
referer STRING,
agent STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe' WITH
SERDEPROPERTIES ( "input.regex" = "([^ ]*) ([^ ]*) ([^ ]*) (-|\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)(?: ([^ \"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\"))?", "output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s %8$s %9$s" )
STORED AS TEXTFILE;
```

Please expand the above regular expression to single line before copying the entire statement to Hue Hive editor.
Test success of creation of that table using two single line samples of Apache logs contained in files apache.access.2.log and apache.access.log (note files do not have .txt suffix) contained in the attached file examples_older.zip. Once you are convinced that you can safely insert those two samples into your table apachelog, insert a bigger log contained in file apache_log_1.txt. Tell us how many lines of apache logs you have in table apachelog.

// I run the create table command above and no results or logs are given by the editor.

// apache.access.log is loaded with the Hue Hive Editor into apachelog table.
LOAD DATA INPATH "/user/cloudera/apache.access.log" INTO TABLE apachelog;

// Log:
INFO : Loading data to table default.apachelog from
hdfs://quickstart.cloudera:8020/user/cloudera/apache.access.log
INFO : Table default.apachelog stats: [numFiles=1, totalSize=86]

// Similarly, apache.access.2.log is loaded with the Hue Hive Editor into apachelog table.

LOAD DATA INPATH "/user/cloudera/apache.access.2.log" INTO TABLE apachelog;

INFO : Loading data to table default.apachelog from
hdfs://quickstart.cloudera:8020/user/cloudera/apache.access.2.log
INFO : Table default.apachelog stats: [numFiles=2, totalSize=305]

// Next, the larger file access_log_1.txt is loaded into the apachelog table.
LOAD DATA INPATH "/user/cloudera/access_log_1.txt" INTO TABLE apachelog;

// Log
INFO : Loading data to table default.apachelog from
hdfs://quickstart.cloudera:8020/user/cloudera/access_log_1.txt
INFO : Table default.apachelog stats: [numFiles=1, totalSize=8754118]

// Querying to count the number of lines in apachelog
select count(*)
from apachelog;

INFO : Number of reduce tasks determined at compile time: 1
INFO : In order to change the average load for a reducer (in bytes):
INFO :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO : In order to limit the maximum number of reducers:
INFO :   set hive.exec.reducers.max=<number>
INFO : In order to set a constant number of reducers:
INFO :   set mapreduce.job.reduces=<number>
INFO : Starting Job = job_1457132565553_0009, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1457132565553_0009/
INFO : Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1457132565553_0009
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO : 2016-03-04 17:31:30,890 Stage-1 map = 0%, reduce = 0%

```
INFO  : 2016-03-04 17:31:52,513 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.06
sec
INFO  : 2016-03-04 17:32:11,385 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.06
sec
INFO  : MapReduce Total cumulative CPU time: 4 seconds 60 msec
INFO  : Ended Job = job_1457132565553_0009

// Result:
39346

// The result of 39346 includes the two inputs from apache.access.log and apache.access.2.log.
// The file access_log_1.txt therefore contributed 39344 of the 39346 lines into
// the table apachelog.
```