UNIVERSITY OF MALAYA

WQD7007 LAB TEST

20220611

GROUP 2

Dr. ALI FEIZOLLAH

ZHANG ZITENG

S2149768

Part 1

Q1

cd ..

cd WQD7007

cd hadoop

 # Create a folder in hdfs to Import dataset and put the dataset to the hdfs

bin/hadoop fs -mkdir -p /data/lab_p1

bin/hadoop fs -put /home/student/Downloads/Set5.csv /data/lab_p1

```
student@student-VirtualBox:~$ cd ..
student@student-VirtualBox:/home$ cd WQD7007
student@student-VirtualBox:/home/WQD7007$ cd hadoop
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop fs -mkdir -p /data/lab_p1
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop fs -put /home/student/Downloads/Set5.csv /data/lab_p1
```

 # Check if successful

bin/hadoop fs -cat /data/lab_p1/Set5.csv

```
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop fs -cat /data/lab_p1/Set5.csv
No,gender,race/ethnicity,parental level of education,1unch,test preparation course,math score,6eading score,writing score
1,female,group D,some high school,standard,completed,97,100,100
2,female,group D,bachelor's degree,standard,completed,68,75,81
3,male,group D,associate's degree,standard,none,80,68,72
4,female,group B,some college,standard,none,79,86,92
5,male,group D,associate's degree,free/reduced,none,90,87,75
6,male,group C,bachelor's degree,standard,completed,83,82,84
7,female,group C,some college,standard,none,73,80,82
8,male,group B,associate's degree,standard,none,65,54,57
9,female,group D,some college,free/reduced,none,71,83,83
10,female,group B,associate's degree,standard,none,82,80,77
11,male,group B,associate's degree,free/reduced,none,44,41,38
12,male,group D,associate's degree,standard,completed,67,72,67
13,male,group B,high school,standard,completed,60,44,47
14,female,group C,associate's degree,free/reduced,none,60,75,74
15,male,group D,high school,standard,none,54,52,52
16,male,group E,bachelor's degree,free/reduced,completed,79,74,72
17,male,group C,some high school,free/reduced,completed,45,52,49
18,male,group C,some high school,standard,none,51,52,44
19,male,group B,associate's degree,standard,completed,81,82,82
20,male,group E,some college,standard,none,97,87,82
21,male,group B,associate's degree,free/reduced,none,67,62,60
22,female,group B,high school,standard,none,58,62,59
23,female,group D,master's degree,free/reduced,completed,85,95,100
24,female,group D,associate's degree,free/reduced,none,47,53,58
25,male,group D,some college,standard,completed,77,62,62
26,male,group D,high school,free/reduced,none,75,74,66
27,female,group D,associate's degree,standard,none,74,81,83
28,female,group C,some high school,standard,none,69,73,73
29,male,group E,some college,standard,none,84,77,71
30,male,group E,some college,standard,none,53,55,48
31,male,group D,high school,standard,none,57,50,54
32,female,group D,some college,standard,none,77,68,77
33,male,group A,some high school,standard,completed,47,49,49
34,male,group C,high school,standard,none,70,70,65
35,male,group C,some college,free/reduced,none,65,58,49
36,female,group B,associate's degree,standard,completed,59,70,66
37,male,group D,some college,standard,completed,76,83,79
38,female,group A,bachelor's degree,standard,none,51,49,51
39,female,group B,some college,standard,none,70,75,78
40,male,group D,associate's degree,free/reduced,none,66,62,64
41,male,group D,some high school,standard,completed,76,70,69
42,male,group C,some college,free/reduced,none,35,28,27
43,male,group D,associate's degree,free/reduced,none,53,54,48
44,male,group C,some college,standard,none,59,60,58
45,female,group C,some college,standard,completed,67,81,79
46,male,group B,some college,free/reduced,none,40,43,39
47,male,group C,high school,standard,none,88,89,86
48,female,group C,some college,free/reduced,completed,64,85,85
49,male,group C,high school,standard,completed,69,58,53
```

Q2

 # Create database and check if success in hive

hive

create databases lab_p1;

show databases;

```
Logging initialized using configuration in jar:file:/home/WQD7007/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties
hive> create database lab_P1;
OK
Time taken: 1.414 seconds
hive> show databases;
OK
count
default
lab_p1
Time taken: 0.289 seconds, Fetched: 3 row(s)
```

 # Create table in database and insert data from hdfs

use lab_p1;

create external table if not exists Set5(no int, gender string, raceEthnicity string, parentalLevelOfEducation string, lunch string, testPreparationCourse string, mathScore int, readingScore int, writingScore int) row format delimited fields terminated by ",";

load data inpath '/data/lab_p1/Set5.csv' into table Set5;

```
hive> use lab_p1;
OK
Time taken: 0.035 seconds
hive> load data local inpath '/data/lab_p1/Set5.csv' into table Set5;
FAILED: SemanticException [Error 10001]: Line 1:58 Table not found 'Set5'
hive> show tables;
OK
Time taken: 0.094 seconds
hive> create external table if not exists Set5(no int, gender string, raceEthnicity string, parentalLevelOfEducation string, lunch string, testPreparationCourse string, mathScore int, readingScore int, wr
ittingScore int) row format delimited fields terminated by ",";
OK
Time taken: 0.621 seconds
hive> load data local inpath '/data/lab_p1/Set5.csv' into table Set5;
FAILED: SemanticException Line 1:23 Invalid path ''/data/lab_p1/Set5.csv'': No files matching path file:/data/lab_p1/Set5.csv
hive> load data inpath '/data/lab_p1/Set5.csv' into table Set5;
Loading data to table lab_p1.set5
Table lab_p1.set5 stats: [numFiles=1, totalSize=3086]
OK
Time taken: 0.781 seconds
```

 # Check if data insert successfully, we can find the first raw is column name, so we might need to drop the first raw.

select * from Set5 limit 5;

```
hive> select * from Set5 limit 5;
OK
NULL    gender  race/ethnicity  parental level of education     1unch   test preparation course NULL    NULL    NULL
1       female  group D some high school        standard        completed       97      100     100
2       female  group D bachelor's degree       standard        completed       68      75      81
3       male    group D associate's degree      standard        none    80      68      72
4       female  group B some college    standard        none    79      86      92
Time taken: 0.375 seconds, Fetched: 5 row(s)
```

 # 5 rows of data that have the highest writing score.

select * from Set5 order by writingScore desc limit 5;

```
hive> select * from Set5 order by writingScore desc limit 5;
Query ID = student_20220611102050_c555b87a-6ed3-4d54-a689-4e5f988fdff5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1654832753295_0031, Tracking URL = http://student-VirtualBox:8088/proxy/application_1654832753295_0031/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1654832753295_0031
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-11 10:21:00,249 Stage-1 map = 0%,  reduce = 0%
2022-06-11 10:21:06,568 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.43 sec
2022-06-11 10:21:13,983 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.69 sec
MapReduce Total cumulative CPU time: 5 seconds 690 msec
Ended Job = job_1654832753295_0031
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.69 sec   HDFS Read: 11858 HDFS Write: 298 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 690 msec
OK
1       female  group D some high school        standard        completed       97      100     100
23      female  group D master's degree free/reduced    completed       85      95      100
4       female  group B some college    standard        none    79      86      92
47      male    group C high school     standard        none    88      89      86
48      female  group C some college    free/reduced    completed       64      85      85
Time taken: 25.82 seconds, Fetched: 5 row(s)
```

drop the first raw of data

insert overwrite table Set5 select * from Set5 where gender in ('male','female');

```
hive> insert overwrite table Set5 select * from Set5 where gender in ('male','female');
Query ID = student_20220611104133_a8cede4f-fbd8-4d33-b155-10fadb6386a3
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1654832753295_0034, Tracking URL = http://student-VirtualBox:8088/proxy/application_1654832753295_0034/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1654832753295_0034
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-06-11 10:41:42,438 Stage-1 map = 0%,  reduce = 0%
2022-06-11 10:41:49,830 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.71 sec
MapReduce Total cumulative CPU time: 3 seconds 710 msec
Ended Job = job_1654832753295_0034
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:9000/user/hive/warehouse/lab_p1.db/set5/.hive-staging_hive_2022-06-11_10-41-33_895_43457037976
75403939-1/-ext-10000
Loading data to table lab_p1.set5
Table lab_p1.set5 stats: [numFiles=1, numRows=50, totalSize=2912, rawDataSize=2862]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.71 sec   HDFS Read: 8345 HDFS Write: 2981 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 710 msec
OK
Time taken: 18.524 seconds
```

 # 5 rows of data that have the lowest math score.

select * from Set5 order by mathScore asc limit 5;

```
hive> select * from Set5 order by mathScore asc limit 5;
Query ID = student_20220611104159_e1527c9c-1bd9-43cf-ae27-72e9957aed2d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1654832753295_0035, Tracking URL = http://student-VirtualBox:8088/proxy/application_1654832753295_0035/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1654832753295_0035
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-11 10:42:07,763 Stage-1 map = 0%,  reduce = 0%
2022-06-11 10:42:15,162 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.65 sec
2022-06-11 10:42:22,549 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.56 sec
MapReduce Total cumulative CPU time: 5 seconds 560 msec
Ended Job = job_1654832753295_0035
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.56 sec   HDFS Read: 11786 HDFS Write: 303 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 560 msec
OK
42      male    group C some college    free/reduced    none    35      28      27
46      male    group B some college    free/reduced    none    40      43      39
11      male    group B associate's degree      free/reduced    none    44      41      38
17      male    group C some high school        free/reduced    completed       45      52      49
24      female  group D associate's degree      free/reduced    none    47      53      58
Time taken: 23.778 seconds, Fetched: 5 row(s)
```

Part 2

Q1

 # Download data from website and insert to hdfs

wget https://www.gutenberg.org/cache/epub/8993/pg8993.txt

```
student@student-VirtualBox:/home/WQD7007/hadoop$ wget https://www.gutenberg.org/cache/epub/8993/pg8993.txt
```

sudo mv pg8993.txt /home/student/

```
student@student-VirtualBox:/home/WQD7007/hadoop$ sudo mv pg8993.txt /home/student/
```

bin/hadoop fs -mkdir -p /data/lab_p2

```
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop fs -mkdir -p /data/lab_p2
```

bin/hadoop fs -put /home/student/pg8993.txt /data/lab_p2/

```
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop fs -put /home/student/pg8993.txt /data/lab_p2/
```

Q2

 # MapReduce to count the word from the text file and save in hdfs

bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar wordcount /data/lab_p2/pg8993.txt /tem

```
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar word
count /data/lab_p2/pg8993.txt /tem
```

 # check the path

bin/hadoop fs -ls /tem

```
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop fs -ls /tem
```

 # the data saved as part-r-00000 in hdfs in tem path, check if the data import successfully.

bin/hadoop fs -text /tem/part-r-00000

```
student@student-VirtualBox:/home/WQD7007/hadoop$ bin/hadoop fs -text /tem/part-r-00000
```

```
“we      1
“what    10
“when    3
“whenever        1
“where   2
“whether         2
“which   5
“while   1
“who     3
“why     1
“will    4
“with    1
“without         1
“would   4
“would,  1
“yes,    1
“you     16
“your    2
“‘A      1
“‘Heaven         1
“‘Here,  1
“‘Only,’         1
“‘Port   1
“‘Who    1
student@student-VirtualBox:/home/WQD7007/hadoop$
```

Q3

 # Create new database, create wordcount table and insert the data from hdfs

hive

create database count;

create table wordcount(word string, count int) row format delimited fields terminated by '\t';

describe wordcount;

load data inpath '/tem/part-r-00000' into table wordcount;

```
student@student-VirtualBox:/home/WQD7007/hadoop$ hive
ls: cannot access '/home/WQD7007/spark/lib/spark-assembly-*.jar': No such file or directory

Logging initialized using configuration in jar:file:/home/WQD7007/hive/lib/hive-common-1.2.2.jar!/hive-log4j.properties
hive> show databases;
OK
count
default
lab_p1
Time taken: 1.344 seconds, Fetched: 3 row(s)
hive> create database lab_p2;
OK
Time taken: 0.341 seconds
hive> show databases;
OK
count
default
lab_p1
lab_p2
Time taken: 0.018 seconds, Fetched: 4 row(s)
hive> use lab_p2;
OK
Time taken: 0.026 seconds
hive> create table wordcount(word string, count int) row format delimited fields terminated by '\t';
OK
Time taken: 0.379 seconds
hive> show tables;
OK
wordcount
Time taken: 0.044 seconds, Fetched: 1 row(s)
hive> load data inpath '/tem/part-r-00000' into table wordcount;
Loading data to table lab_p2.wordcount
Table lab_p2.wordcount stats: [numFiles=1, totalSize=234040]
OK
Time taken: 0.95 seconds
hive> select * from wordcount limit 5;
OK
"Defects,"      1
"Information    1
"Plain  2
"Project        5
"Right  1
Time taken: 0.367 seconds, Fetched: 5 row(s)
```

 # 5 words with 5 counts in ascending alphabetical order.

select * from wordcount where count = 5 order by word asc limit 5;

```
hive> select * from wordcount where count = 5 order by word asc limit 5;
Query ID = student_20220611101811_bb52a2cb-4531-4ffa-ac40-addc877c011e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1654832753295_0030, Tracking URL = http://student-VirtualBox:8088/proxy/application_1654832753295_0030/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1654832753295_0030
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-11 10:18:20,942 Stage-1 map = 0%,  reduce = 0%
2022-06-11 10:18:29,528 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.46 sec
2022-06-11 10:18:36,988 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.96 sec
MapReduce Total cumulative CPU time: 8 seconds 960 msec
Ended Job = job_1654832753295_0030
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.96 sec   HDFS Read: 240643 HDFS Write: 36 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 960 msec
OK
"Project        5
11th    5
15      5
150     5
16th    5
Time taken: 28.078 seconds, Fetched: 5 row(s)
```

 # 5 words with lowest counts in descending alphabetical order.

select * from (select * from wordcount order by count asc) as table1 order by word desc limit 5;

```
hive> select * from (select * from wordcount order by count desc) as table1 order by word asc limit 5;
Query ID = student_20220611101612_ca777ef0-3e8a-44f2-9d0f-06a6727dafd0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1654832753295_0028, Tracking URL = http://student-VirtualBox:8088/proxy/application_1654832753295_0028/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1654832753295_0028
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-11 10:16:20,863 Stage-1 map = 0%,  reduce = 0%
2022-06-11 10:16:29,326 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.1 sec
2022-06-11 10:16:37,812 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.66 sec
MapReduce Total cumulative CPU time: 7 seconds 660 msec
Ended Job = job_1654832753295_0028
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1654832753295_0029, Tracking URL = http://student-VirtualBox:8088/proxy/application_1654832753295_0029/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1654832753295_0029
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-06-11 10:16:51,847 Stage-2 map = 0%,  reduce = 0%
2022-06-11 10:16:59,276 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.86 sec
2022-06-11 10:17:07,750 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.4 sec
MapReduce Total cumulative CPU time: 7 seconds 400 msec
Ended Job = job_1654832753295_0029
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.66 sec   HDFS Read: 239401 HDFS Write: 573719 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.4 sec   HDFS Read: 578217 HDFS Write: 57 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 60 msec
OK
"Defects,"      1
"Information    1
"Plain  2
"Project        5
"Right  1
Time taken: 56.184 seconds, Fetched: 5 row(s)
hive>
```