Name: Alvyn Abranches

Class: FY MSc Data Science and Big Data Analytics

Roll No 1

Exercise 1

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
Time taken: 5.675 seconds, Fetched: 1 row(s)
hive> create database alvyn;
OK
Time taken: 1.16 seconds
hive> use alvyn;
OK
Time taken: 0.079 seconds
```

```
hive> create table hospitals(DRGDefinition string, ProviderId int, ProviderName string, ProviderStreetAddress string, ProviderCity string, ProviderState string, ProviderZipCode int, HospitalReferralRegionDescription string, TotalDischarg
es int, AverageCoveredCharges double, AverageTotalPayments double, AverageMedicarePayments double) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.782 seconds
```

```
hive> load data local inpath 'hospital_data.csv' into table hospitals;
Loading data to table alvyn.hospitals
Table alvyn.hospitals stats: [numFiles=1, totalSize=26841576]
OK
Time taken: 1.205 seconds
```

```
hive> select * from hospitals limit 5;
OK
DRGDefinition   NULL    ProviderName    ProviderStreetAddress   ProviderCity    ProviderState   NULL    HospitalReferralRegionDescription       NULL    NULL    NULL    NULL
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC        10001   SOUTHEAST ALABAMA MEDICAL CENTER        1108 ROSS CLARK CIRCLE  DOTHAN  AL      36301   AL - Dothan     91      32963.07        5777.24 4763.73
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC        10005   MARSHALL MEDICAL CENTER SOUTH   2505 U S HIGHWAY 431 NORTH      BOAZ    AL      35957   AL - Birmingham 14      15131.85        5787.57 4976.71
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC        10006   ELIZA COFFEE MEMORIAL HOSPITAL  205 MARENGO STREET      FLORENCE        AL      35631   AL - Birmingham 24      37560.37        5434.95 4453.79
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC        10011   ST VINCENT'S EAST       50 MEDICAL PARK EAST DRIVE      BIRMINGHAM      AL      35235   AL - Birmingham 25      13998.28        5417.56 4129.16
Time taken: 0.063 seconds, Fetched: 5 row(s)
```

Exercise 2

```
hive> select AverageTotalPayments, ProviderState from hosptials order by AverageTotalPayments desc, ProviderState asc limit 5;
FAILED: SemanticException [Error 10001]: Line 1:48 Table not found 'hosptials'
hive> select AverageTotalPayments, ProviderState from hospitals order by AverageTotalPayments desc, ProviderState asc limit 5;
Query ID = cloudera_20191118195454_e3c1a9c0-f3c9-449a-a23b-9383194f10b9
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_1574131451414_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1574131451414_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1574131451414_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-11-18 19:54:30,615 Stage-1 map = 0%,  reduce = 0%
2019-11-18 19:54:37,972 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.94 sec
2019-11-18 19:54:44,204 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.93 sec
MapReduce Total cumulative CPU time: 5 seconds 930 msec
Ended Job = job_1574131451414_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.93 sec   HDFS Read: 26850270 HDFS Write: 111 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 930 msec
OK
447714.88       SAN FRANCISCO
414555.91       SAN FRANCISCO
400675.86       GREENBRAE
391446.0        GREENBRAE
381799.81       SAN FRANCISCO
Time taken: 25.371 seconds, Fetched: 5 row(s)
```

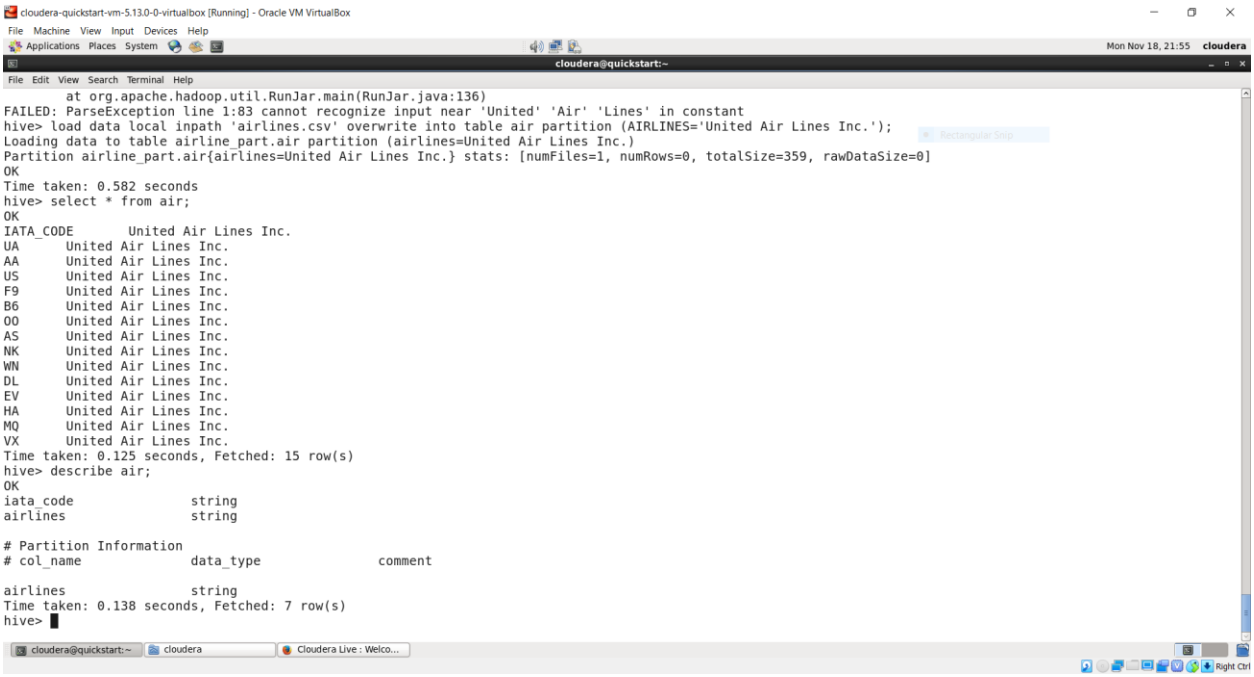
Exercise 3

```
hive> select ProviderState, avg(AverageMedicarePayments) from hospitals group by ProviderState limit 5;
Query ID = cloudera_20191118195656_e377defc-c730-427e-a320-64f3577a56c4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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_1574131451414_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1574131451414_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1574131451414_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-11-18 19:57:01,749 Stage-1 map = 0%,  reduce = 0%
2019-11-18 19:57:06,957 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.64 sec
2019-11-18 19:57:12,143 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.3 sec
MapReduce Total cumulative CPU time: 4 seconds 300 msec
Ended Job = job_1574131451414_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.3 sec   HDFS Read: 26851870 HDFS Write: 117 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 300 msec
OK
  2ND FLOOR"    41004.6
  INC"  18.0
  P O BOX 280"  8306.91
  PO BOX 788"   12882.036666666667
 10 EAST 31ST ST"       24371.192
Time taken: 17.135 seconds, Fetched: 5 row(s)
```

Exercise 4

```
hive> CREATE TABLE airline (YEAR int, MONTH int, DAY int, DAY_OF_WEEK int, AIRLINE string, FLIGHT_NUMBER int, TAIL_NUMBER string, ORIGIN_AIRPORT string, DESTINATION_AIRPORT string, SCHEDULED_DEPARTURE int, DEPARTURE_TIME int, DEPARTURE_D
ELAY int, TAXI_OUT int, WHEELS_OFF int, SCHEDULED_TIME int, ELAPSED_TIME int, AIR_TIME int, DISTANCE int, WHEELS_ON int, TAXI_N int, SCHEDULED_ARRIVAL int, ARRIVAL_TIME int, ARRIVAL_DELAY int, DIVERTED int, CANCELLED int) row format deli
mited fields terminated by ',';
OK
Time taken: 0.129 seconds
```

```
hive> load data local inpath 'airline_data.csv' into table airline;
Loading data to table alvyn.airline
Table alvyn.airline stats: [numFiles=1, totalSize=97368314]
OK
Time taken: 0.724 seconds
```



Exercise 5

```
hive> create table realestate(street varchar(40),city string,zip int,state string,beds int,baths int,sq_ft int,type string,price int) c
3 buckets row format delimited fields terminated by ',';
OK
Time taken: 0.094 seconds
```

Exercise 6

Map Side Join