Name: -

Roll no:-

Assignment No:-03

[1] Create Flight Information 2007 Table :

```
hive> CREATE TABLE IF NOT EXISTS FlightInfo2007
    > (
    > Year SMALLINT,
    >  Month TINYINT,
    > DayofMonth TINYINT,
    > DayOfWeek TINYINT,
    > DepTime SMALLINT,
    > CRSDepTime SMALLINT,
    > ArrTime SMALLINT,
    > CRSArrTime SMALLINT,
    > UniqueCarrier STRING,
    > FlightNum STRING,
    > TailNum STRING,
    > ActualElapsedTime SMALLINT,
    > CRSElapsedTime SMALLINT,
    > AirTime SMALLINT,
    > ArrDelay SMALLINT,
    > DepDelay SMALLINT ,
    > Origin STRING,
    > Dest STRING,
    > Distance INT,
    > TaxiIn SMALLINT,
    > TaxiOut SMALLINT,
    > Cancelled SMALLINT,
    > CancellationCode STRING,
    > Diverted SMALLINT,
    > CarrierDelay SMALLINT,
    > WeatherDelay SMALLINT,
    > NASDelay SMALLINT,
    > SecurityDelay SMALLINT,
    > LateAircraftDelay SMALLINT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE
    >
    >
    > ;
OK
Time taken: 0.225 seconds
```

[2] Load Flight information data from 2007.csv file:

```
hive> load data local inpath '/home/cloudera/2007.csv' into table FlightInfo2007;
Loading data to table ass_3_31.flightinfo2007
Table ass_3_31.flightinfo2007 stats: [numFiles=1, totalSize=421527552]
OK
Time taken: 6.209 seconds
```

[3] Create FlightInfo2008 table:

```
hive> CREATE TABLE IF NOT EXISTS FlightInfo2008 LIKE FlightInfo2007;
OK
Time taken: 0.134 seconds
```

[4] Load Flight information data from 2008.csv file into table FlightInfo2008

```
hive> load data local inpath '/home/cloudera/2008.csv' into table FlightInfo2008
    > ;
Loading data to table ass_3_31.flightinfo2008
Table ass_3_31.flightinfo2008 stats: [numFiles=1, totalSize=689413344]
OK
Time taken: 7.555 seconds
```

[5]create myflightinfo2007 table to store Year, Month, DepTime, ArrTime, FlightNum, Origin, Dest data from FlightInfo2007 table month 7 (July) and 3 (March).

[6] Write query to dispay flight information as :Year, Month, DepTime, ArrTime,FlightNum, Origin, Dest in year 2007.

```
hive> SELECT * FROM myFlightInfo2007;
OK
2007    7       700     834     5447    JFK     ORD
2007    7       1633    1812    5469    JFK     ORD
2007    7       1905    2100    5492    JFK     ORD
2007    7       1453    1624    4133    JFK     ORD
2007    7       1810    1956    4392    JFK     ORD
2007    7       643     759     903     JFK     ORD
2007    7       939     1108    907     JFK     ORD
2007    7       1313    1436    915     JFK     ORD
2007    7       1617    1755    917     JFK     ORD
2007    7       2002    2139    919     JFK     ORD
Time taken: 0.078 seconds, Fetched: 10 row(s)
```

[7] create myflightinfo2007 table to store Year, Month, DepTime, ArrTime, FlightNum, Origin, Dest data from FlightInfo2008 table for month 7 (July) and 3 (March)

```
hive> CREATE TABLE myFlightInfo2008 AS
    > SELECT Year, Month, DepTime, ArrTime, FlightNum,
    > Origin, Dest FROM FlightInfo2008
    > WHERE (Month = 7 AND DayofMonth = 3) AND
    > (Origin='JFK' AND Dest='ORD');
Query ID = cloudera_20220428053636_afe8d52d-a5e6-4f9b-9fe6-ef775e90c515
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_1651146781281_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1651146781281_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1651146781281_0003
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 0
2022-04-28 05:36:47,348 Stage-1 map = 0%,  reduce = 0%
2022-04-28 05:37:12,546 Stage-1 map = 17%,  reduce = 0%, Cumulative CPU 2.28 sec
2022-04-28 05:37:18,289 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 7.04 sec
2022-04-28 05:37:19,413 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 7.82 sec
2022-04-28 05:37:20,567 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.91 sec
MapReduce Total cumulative CPU time: 8 seconds 910 msec
Ended Job = job_1651146781281_0003
Stage-4 is filtered out by condition resolver.
Stage-3 is selected by condition resolver.
Stage-5 is filtered out by condition resolver.
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1651146781281_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1651146781281_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1651146781281_0004
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-04-28 05:37:30,506 Stage-3 map = 0%,  reduce = 0%
2022-04-28 05:37:36,865 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.66 sec
MapReduce Total cumulative CPU time: 660 msec
Ended Job = job_1651146781281_0004
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/ass_3_31.db/myflightinfo2008
Table ass_3_31.myflightinfo2008 stats: [numFiles=1, numRows=10, totalSize=290, rawDataSize=280]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3   Cumulative CPU: 8.91 sec   HDFS Read: 609453477 HDFS Write: 474 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 0.66 sec   HDFS Read: 2433 HDFS Write: 290 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 570 msec
OK
Time taken: 58.807 seconds
```

[8] **Write query to dispay flight information as :Year, Month, DepTime,ArrTime, FlightNum, Origin, Dest in year 2008.**

```
hive> SELECT * FROM myFlightInfo2008;
OK
2008    7       930     1103    5199    JFK     ORD
2008    7       705     849     5687    JFK     ORD
2008    7       1645    1914    5469    JFK     ORD
2008    7       1345    1514    4392    JFK     ORD
2008    7       1718    1907    1217    JFK     ORD
2008    7       757     929     1323    JFK     ORD
2008    7       928     1057    907     JFK     ORD
2008    7       1358    1532    915     JFK     ORD
2008    7       1646    1846    917     JFK     ORD
2008    7       2129    2341    919     JFK     ORD
Time taken: 0.058 seconds, Fetched: 10 row(s)
```

[9] **Write query to dispay flight information as :Year, Month, DepTime, ArrTime, FlightNum,Origin, Dest in year 2007 and 2008**

JOIN LEFT OUTER JOIN

```
hive> SELECT m8.Year, m8.Month, m8.FlightNum, m8.Origin, m8.Dest, m7.Year, m7.Month,
    > m7.FlightNum, m7.Origin, m7.Dest
    > FROM myFlightinfo2008 m8 LEFT OUTER JOIN myFlightinfo2007 m7
    > ON m8.FlightNum=m7.FlightNum;
Query ID = cloudera_20220428054040_d771d842-9471-4094-bc29-a8af43f9b424
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20220428054040_d771d842-9471-4094-bc29-a8af43f9b424.log
2022-04-28 05:40:13     Starting to launch local task to process map join;      maximum memory = 1013645312
2022-04-28 05:40:14     Dump the side-table for tag: 1 with group count: 10 into file: file:/tmp/cloudera/74f3dc31-156b-4175-b645-406ec7b87458/hive_2022-04-28_05-4
HashTable-Stage-3/MapJoin-mapfile11--.hashtable
2022-04-28 05:40:14     Uploaded 1 File to: file:/tmp/cloudera/74f3dc31-156b-4175-b645-406ec7b87458/hive_2022-04-28_05-40-09_644_1162630861184495934-1/-local-10003
able (598 bytes)
2022-04-28 05:40:14     End of local task; Time Taken: 0.914 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1651146781281_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1651146781281_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1651146781281_0007
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-04-28 05:40:22,765 Stage-3 map = 0%,  reduce = 0%
2022-04-28 05:40:29,201 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.79 sec
MapReduce Total cumulative CPU time: 790 msec
Ended Job = job_1651146781281_0007
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 0.79 sec   HDFS Read: 7991 HDFS Write: 372 SUCCESS
Total MapReduce CPU Time Spent: 790 msec
OK
2008    7       5199    JFK     ORD     NULL    NULL    NULL    NULL    NULL
2008    7       5687    JFK     ORD     NULL    NULL    NULL    NULL    NULL
2008    7       5469    JFK     ORD     2007    7       5469    JFK     ORD
2008    7       4392    JFK     ORD     2007    7       4392    JFK     ORD
2008    7       1217    JFK     ORD     NULL    NULL    NULL    NULL    NULL
2008    7       1323    JFK     ORD     NULL    NULL    NULL    NULL    NULL
2008    7       907     JFK     ORD     2007    7       907     JFK     ORD
2008    7       915     JFK     ORD     2007    7       915     JFK     ORD
2008    7       917     JFK     ORD     2007    7       917     JFK     ORD
2008    7       919     JFK     ORD     2007    7       919     JFK     ORD
Time taken: 20.618 seconds, Fetched: 10 row(s)
```

FULL OUTER JOIN

```
hive> SELECT m31.FlightNum,m31.Origin,m31.Dest,m7.FlightNum,m7.Origin,m7.Dest
    > FROM myFlightinfo2008 m31 FULL OUTER JOIN myFlightinfo2007 m7
    > ON m31.FlightNum = m7.FlightNum;
Query ID = cloudera_20220428055353_bb5a6fdd-17cf-476d-a8f1-96ee7255e584
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_1651146781281_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1651146781281_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1651146781281_0011
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2022-04-28 05:53:26,579 Stage-1 map = 0%,  reduce = 0%
2022-04-28 05:53:36,210 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.14 sec
2022-04-28 05:53:42,467 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.98 sec
MapReduce Total cumulative CPU time: 1 seconds 980 msec
Ended Job = job_1651146781281_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 1.98 sec   HDFS Read: 14275 HDFS Write: 323 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 980 msec
OK
1217    JFK     ORD     NULL    NULL    NULL
1323    JFK     ORD     NULL    NULL    NULL
NULL    NULL    NULL    4133    JFK     ORD
4392    JFK     ORD     4392    JFK     ORD
5199    JFK     ORD     NULL    NULL    NULL
NULL    NULL    NULL    5447    JFK     ORD
5469    JFK     ORD     5469    JFK     ORD
NULL    NULL    NULL    5492    JFK     ORD
5687    JFK     ORD     NULL    NULL    NULL
NULL    NULL    NULL    903     JFK     ORD
907     JFK     ORD     907     JFK     ORD
915     JFK     ORD     915     JFK     ORD
917     JFK     ORD     917     JFK     ORD
919     JFK     ORD     919     JFK     ORD
Time taken: 23.557 seconds, Fetched: 14 row(s)
```

[10] **Create index on origin field**

```
hive> CREATE INDEX f08_index ON TABLE flightinfo2008 (Origin) AS
    > 'COMPACT' WITH DEFERRED REBUILD;
OK
Time taken: 8.325 seconds
hive> ALTER INDEX f08_index ON flightinfo2008 REBUILD;
Query ID = cloudera_20220428054040_e72383e3-1ef2-486c-bf28-e45c4b69a939
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 3
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_1651146781281_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1651146781281_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1651146781281_0008
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3
2022-04-28 05:40:56,662 Stage-1 map = 0%,  reduce = 0%
2022-04-28 05:41:25,550 Stage-1 map = 11%,  reduce = 0%, Cumulative CPU 7.66 sec
2022-04-28 05:41:34,875 Stage-1 map = 42%,  reduce = 0%, Cumulative CPU 11.66 sec
2022-04-28 05:41:36,224 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 13.08 sec
2022-04-28 05:41:38,368 Stage-1 map = 78%,  reduce = 0%, Cumulative CPU 14.55 sec
2022-04-28 05:41:42,783 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 16.22 sec
2022-04-28 05:41:58,293 Stage-1 map = 100%,  reduce = 49%, Cumulative CPU 19.27 sec
2022-04-28 05:41:59,379 Stage-1 map = 100%,  reduce = 74%, Cumulative CPU 20.83 sec
2022-04-28 05:42:00,485 Stage-1 map = 100%,  reduce = 83%, Cumulative CPU 21.45 sec
2022-04-28 05:42:02,634 Stage-1 map = 100%,  reduce = 95%, Cumulative CPU 23.41 sec
2022-04-28 05:42:03,665 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 23.94 sec
MapReduce Total cumulative CPU time: 23 seconds 940 msec
Ended Job = job_1651146781281_0008
Loading data to table ass_3_31.ass_3_31__flightinfo2008_f08_index__
Table ass_3_31.ass_3_31__flightinfo2008_f08_index__ stats: [numFiles=3, numRows=304, totalSize=68988624, rawDataSize=68988320]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 23.94 sec   HDFS Read: 689470298 HDFS Write: 68988947 SUCCESS
Total MapReduce CPU Time Spent: 23 seconds 940 msec
OK
Time taken: 75.744 seconds
```

[11] **Find average deparature delay for flights in year 2008**

```
hive> CREATE VIEW avgdepdelay AS
    > SELECT DayOfWeek, AVG(DepDelay)
    > FROM  FlightInfo2008
    >  GROUP BY DayOfWeek;
OK
Time taken: 0.111 seconds
hive> SELECT * FROM avgdepdelay;
Query ID = cloudera_20220428054343_231b0723-433e-4d4f-9876-2d6c71ced581
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 3
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_1651146781281_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1651146781281_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1651146781281_0009
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 3
2022-04-28 05:44:01,936 Stage-1 map = 0%,  reduce = 0%
2022-04-28 05:44:25,979 Stage-1 map = 11%,  reduce = 0%, Cumulative CPU 6.46 sec
2022-04-28 05:44:28,117 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 7.45 sec
2022-04-28 05:44:29,151 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.8 sec
2022-04-28 05:44:43,268 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 9.69 sec
2022-04-28 05:44:44,296 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.48 sec
MapReduce Total cumulative CPU time: 11 seconds 480 msec
Ended Job = job_1651146781281_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 11.48 sec   HDFS Read: 689467403 HDFS Write: 148 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 480 msec
OK
NULL    NULL
3       8.289761053658728
6       8.645680904903614
1       10.269990244459473
4       9.772897177836782
7       11.560973392595312
2       8.97689712068735
5       12.158836387869656
Time taken: 51.096 seconds, Fetched: 8 row(s)
```