Name : Prajakta Keer
Roll No : 33231
Batch : L10

ASSIGNMENT 4

Write an application using HBase and HiveQL for flight information system which will include :

- Creating, Dropping, and altering Database tables
- Creating an external Hive table to connect to the HBase for Customer Information Table
- Load table with data, insert new values and field in the table, Join tables with Hive
- Create index on Flight information Table
- Find the average departure delay per day in 2008.

Creating, Dropping, and altering Database tables

Opening HBASE shell

```
[cloudera@quickstart ~]$ hbase shell
2021-03-16 06:54:38,739 INFO  [main] Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.0.0-cdh5.4.2, rUnknown, Tue May 19 17:07:29 PDT 2015
```

Creating a table and listing the created table

```
hbase(main):007:0> create 'flight','finfo','fsch'
0 row(s) in 0.9080 seconds

=> Hbase::Table – flight

hbase(main):008:0> list
TABLE
flight
student
2 row(s) in 0.0140 seconds

=> ["flight", "student"]
```

Adding data to the table

```
hbase(main):003:0> put 'flight',1,'finfo:source','pune'
0 row(s) in 0.5620 seconds

hbase(main):004:0> put 'flight',1,'finfo:dest','kolkata'
0 row(s) in 0.0250 seconds

hbase(main):005:0> put 'flight',1,'fsch:at','10.00a.m.'
0 row(s) in 0.0100 seconds
```

```
hbase(main):006:0> put 'flight',1,'fsch:dt','11.25 a.m.'
0 row(s) in 0.0320 seconds

hbase(main):007:0> put 'flight',1,'fsch:delay','5min'
0 row(s) in 0.0420 seconds

hbase(main):008:0> put 'flight',2,'finfo:source','kokata'
0 row(s) in 0.0200 seconds

hbase(main):009:0> put 'flight',2,'finfo:dest','pune'
0 row(s) in 0.0110 seconds

hbase(main):010:0> put 'flight',2,'fsch:at','7.00a.m.'
0 row(s) in 0.0230 seconds

hbase(main):011:0> put 'flight',2,'fsch:dt','7.30a.m.'
0 row(s) in 0.0200 seconds

hbase(main):012:0> put 'flight',2,'fsch:delay','2 min'
0 row(s) in 0.0220 seconds

hbase(main):013:0> put 'flight',3,'finfo:source','pune'
0 row(s) in 0.0090 seconds

hbase(main):014:0> put 'flight',3,'finfo:dest','goa'
0 row(s) in 0.0170 seconds

hbase(main):015:0> put 'flight',3,'fsch:at','12.30p.m.'
0 row(s) in 0.0220 seconds

hbase(main):016:0> put 'flight',3,'fsch:dt','12.45p.m.'
0 row(s) in 0.0200 seconds

hbase(main):017:0> put 'flight',3,'fsch:delay','1 min'
0 row(s) in 0.0200 seconds

hbase(main):018:0> put 'flight',4,'finfo:source','goa'
0 row(s) in 0.0830 seconds

hbase(main):019:0> put 'flight',4,'finfo:dest','pune'
0 row(s) in 0.0110 seconds

hbase(main):020:0> put 'flight',4,'fsch:at','2.00p.m.'
0 row(s) in 0.0200 seconds

hbase(main):021:0> put 'flight',4,'fsch:dt','2.45p.m.'
0 row(s) in 0.0180 seconds

hbase(main):022:0> put 'flight',4,'fsch:delay','10 min'
0 row(s) in 0.0200 seconds
```

Display contents of the table

```
hbase(main):030:0> scan 'flight'
ROW                         COLUMN+CELL
 1                          column=finfo:dest, timestamp=1615903581369, value=kolkata
 1                          column=finfo:source, timestamp=1615903574736, value=pune
 1                          column=fsch:at, timestamp=1615903592692, value=10.00a.m.
 1                          column=fsch:delay, timestamp=1615903608703, value=5min
 1                          column=fsch:dt, timestamp=1615903601645, value=11.25 a.m.
 2                          column=finfo:dest, timestamp=1615903624504, value=pune
 2                          column=finfo:source, timestamp=1615903616101, value=kokata
 2                          column=fsch:at, timestamp=1615903631225, value=7.00a.m.
 2                          column=fsch:delay, timestamp=1615903647586, value=2 min
 2                          column=fsch:dt, timestamp=1615903639890, value=7.30a.m.
 3                          column=finfo:dest, timestamp=1615903661209, value=goa
 3                          column=finfo:source, timestamp=1615903654771, value=pune
 3                          column=fsch:at, timestamp=1615903666312, value=12.30p.m.
 3                          column=fsch:delay, timestamp=1615903680337, value=1 min
 3                          column=fsch:dt, timestamp=1615903673848, value=12.45p.m.
 4                          column=finfo:dest, timestamp=1615903699462, value=pune
 4                          column=finfo:source, timestamp=1615903689829, value=goa
 4                          column=fsch:at, timestamp=1615903705357, value=2.00p.m.
 4                          column=fsch:delay, timestamp=1615903723874, value=10 min
 4                          column=fsch:dt, timestamp=1615903715620, value=2.45p.m.
4 row(s) in 0.2500 seconds
```

Adding a column to the existing table and assigning values to this new column

```
hbase(main):031:0> alter 'flight',NAME=>'fare'
Updating all regions with the new schema...
0/1 regions updated.
1/1 regions updated.
Done.
0 row(s) in 2.2930 seconds

hbase(main):033:0> put 'flight',1,'fare:rs','20000'
0 row(s) in 0.0190 seconds

hbase(main):034:0> put 'flight',2,'fare:rs','50000'
0 row(s) in 0.0150 seconds

hbase(main):035:0> put 'flight',3,'fare:rs','30000'
0 row(s) in 0.0130 seconds

hbase(main):036:0> put 'flight',4,'fare:rs','10000'
0 row(s) in 0.0150 seconds

hbase(main):037:0> scan 'flight'
ROW                         COLUMN+CELL
 1                          column=fare:rs, timestamp=1615904086813, value=20000
 1                          column=finfo:dest, timestamp=1615903581369, value=kolkata
 1                          column=finfo:source, timestamp=1615903574736, value=pune
```

```
1                       column=fsch:at, timestamp=1615903592692, value=10.00a.m.
1                       column=fsch:delay, timestamp=1615903608703, value=5min
1                       column=fsch:dt, timestamp=1615903601645, value=11.25 a.m.
2                       column=fare:rs, timestamp=1615904092110, value=50000
2                       column=finfo:dest, timestamp=1615903624504, value=pune
2                       column=finfo:source, timestamp=1615903616101, value=kokata
2                       column=fsch:at, timestamp=1615903631225, value=7.00a.m.
2                       column=fsch:delay, timestamp=1615903647586, value=2 min
2                       column=fsch:dt, timestamp=1615903639890, value=7.30a.m.
3                       column=fare:rs, timestamp=1615904096398, value=30000
3                       column=finfo:dest, timestamp=1615903661209, value=goa
3                       column=finfo:source, timestamp=1615903654771, value=pune
3                       column=fsch:at, timestamp=1615903666312, value=12.30p.m.
3                       column=fsch:delay, timestamp=1615903680337, value=1 min
3                       column=fsch:dt, timestamp=1615903673848, value=12.45p.m.
4                       column=fare:rs, timestamp=1615904100692, value=10000
4                       column=finfo:dest, timestamp=1615903699462, value=pune
4                       column=finfo:source, timestamp=1615903689829, value=goa
4                       column=fsch:at, timestamp=1615903705357, value=2.00p.m.
4                       column=fsch:delay, timestamp=1615903723874, value=10 min
4                       column=fsch:dt, timestamp=1615903715620, value=2.45p.m.
4 row(s) in 0.0850 seconds
```

Deleting the newly added column

```
hbase(main):037:0> alter 'flight',NAME=>'fare',METHOD=>'delete'
Updating all regions with the new schema...
0/1 regions updated.
1/1 regions updated.
Done.
0 row(s) in 2.2640 seconds

hbase(main):030:0> scan 'flight'
ROW                     COLUMN+CELL
1                       column=finfo:dest, timestamp=1615903581369, value=kolkata
1                       column=finfo:source, timestamp=1615903574736, value=pune
1                       column=fsch:at, timestamp=1615903592692, value=10.00a.m.
1                       column=fsch:delay, timestamp=1615903608703, value=5min
1                       column=fsch:dt, timestamp=1615903601645, value=11.25 a.m.
2                       column=finfo:dest, timestamp=1615903624504, value=pune
2                       column=finfo:source, timestamp=1615903616101, value=kokata
2                       column=fsch:at, timestamp=1615903631225, value=7.00a.m.
2                       column=fsch:delay, timestamp=1615903647586, value=2 min
2                       column=fsch:dt, timestamp=1615903639890, value=7.30a.m.
3                       column=finfo:dest, timestamp=1615903661209, value=goa
3                       column=finfo:source, timestamp=1615903654771, value=pune
3                       column=fsch:at, timestamp=1615903666312, value=12.30p.m.
3                       column=fsch:delay, timestamp=1615903680337, value=1 min
3                       column=fsch:dt, timestamp=1615903673848, value=12.45p.m.
4                       column=finfo:dest, timestamp=1615903699462, value=pune
4                       column=finfo:source, timestamp=1615903689829, value=goa
4                       column=fsch:at, timestamp=1615903705357, value=2.00p.m.
```

```
 4                              column=fsch:delay, timestamp=1615903723874, value=10 min
 4                              column=fsch:dt, timestamp=1615903715620, value=2.45p.m.
4 row(s) in 0.2500 seconds
```

Retrieve a particular record by key

```
hbase(main):038:0> get 'flight',3
COLUMN                          CELL
 finfo:dest                     timestamp=1615903661209, value=goa
 finfo:source                   timestamp=1615903654771, value=pune
 fsch:at                        timestamp=1615903666312, value=12.30p.m.
 fsch:delay                     timestamp=1615903680337, value=1 min
 fsch:dt                        timestamp=1615903673848, value=12.45p.m.
6 row(s) in 0.0620 seconds
```

Retrieving the value of a particular column

```
hbase(main):039:0> get 'flight','2',COLUMN=>'finfo:source'
COLUMN                          CELL
 finfo:source                   timestamp=1615903616101, value=kokata
1 row(s) in 0.0190 seconds
```

Retrieving the value from multiple columns

```
hbase(main):040:0> get 'flight','4',COLUMN=>['finfo:source','finfo:dest']
COLUMN                          CELL
 finfo:dest                     timestamp=1615903699462, value=pune
 finfo:source                   timestamp=1615903689829, value=goa
2 row(s) in 0.0140 seconds

hbase(main):042:0> scan 'flight',COLUMNS=>'fsch:delay'
ROW                             COLUMN+CELL
 1                              column=fsch:delay, timestamp=1615903608703, value=5min
 2                              column=fsch:delay, timestamp=1615903647586, value=2 min
 3                              column=fsch:delay, timestamp=1615903680337, value=1 min
 4                              column=fsch:delay, timestamp=1615903723874, value=10 min
4 row(s) in 0.0230 seconds
```

Disabling and Deleting a table

```
hbase(main):004:0> disable 'flight'
0 row(s) in 1.4050 seconds

hbase(main):005:0> drop 'flight'
0 row(s) in 0.5080 seconds

hbase(main):006:0> list
TABLE
student
1 row(s) in 0.0160 seconds
```

Creating an external Hive table to connect to the HBase for Customer Information Table

Create external table and Load data from studentdb.txt to studentdata table

```
hive> create external table studentdata ( name string, roll int)
    > row format delimited fields terminated by "," stored as textfile;
OK
Time taken: 0.559 seconds

hive> load data local inpath '/home/cloudera/Desktop/studentdb.txt' into table studentdata;
Loading data to table default.studentdata
Table default.studentdata stats: [numFiles=5, numRows=0, totalSize=255, rawDataSize=0]
OK
Time taken: 0.859 seconds

hive> select * from studentdata;
OK
prajakta        31
rishita 15
riya    17
sakshi  22
priya   20
Time taken: 0.561 seconds, Fetched: 25 row(s)
```

Create a table in Hbase

```
hbase(main):046:0> create 'student', 'cf'
0 row(s) in 0.4960 seconds

=> Hbase::Table – student
```

Creating external hive table

```
hive> CREATE external TABLE hive_table_student(name string, roll int)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,cf:roll")
    > TBLPROPERTIES ("hbase.table.name" = "student");
OK
Time taken: 1.761 seconds
```

Creating manager table

```
hive> create table studentnew(name string, roll int) row format delimited fields terminated by ','
stored as textfile;
OK
Time taken: 1.148 seconds

hive> load data local inpath '/home/cloudera/Desktop/studentdb.txt' into table studentnew;
Loading data to table default.studentnew
Table default.studentnew stats: [numFiles=1, totalSize=50]
OK
```

Time taken: 0.658 seconds

Inserting data into external table

```
hive> insert into hive_table_student select * from studentnew;
Query ID = cloudera_20210316075252_6b4e3995-a793-4913-b4ba-85f323c7750f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1615902595627_0003, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1615902595627_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1615902595627_0003
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
2021-03-16 07:52:57,872 Stage-0 map = 0%,  reduce = 0%
2021-03-16 07:53:11,775 Stage-0 map = 100%,  reduce = 0%, Cumulative CPU 2.12 sec
MapReduce Total cumulative CPU time: 2 seconds 120 msec
Ended Job = job_1615902595627_0003
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1   Cumulative CPU: 2.12 sec   HDFS Read: 10515 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 120 msec
OK
Time taken: 37.559 seconds
```

Displaying the data

```
hive> select * from hive_table_student;
OK
prajakta        31
priya   20
rishita 15
riya    17
sakshi  22
Time taken: 0.186 seconds, Fetched: 5 row(s)

hbase(main):047:0> scan 'student'
ROW                             COLUMN+CELL
 prajakta                       column=cf:roll, timestamp=1615906559987, value=31
 priya                          column=cf:roll, timestamp=1615906559987, value=20
 rishita                        column=cf:roll, timestamp=1615906559987, value=15
 riya                           column=cf:roll, timestamp=1615906559987, value=17
 sakshi                         column=cf:roll, timestamp=1615906559987, value=22
5 row(s) in 0.0340 seconds
```

Load table with data, insert new values and field in the table, Join tables with Hive

```
hive> create table studentgrades(roll int, grade string) row format delimited fields terminated by ','
stored as textfile;
OK
Time taken: 0.131 seconds
```

```
hive> load data local inpath '/home/cloudera/Desktop/grades.txt' into table studentgrades;
Loading data to table default.studentgrades
Table default.studentgrades stats: [numFiles=1, totalSize=31]
OK
Time taken: 0.343 seconds

hive> select studentnew.roll, studentnew.name, studentgrades.grade from studentnew join
studentgrades on studentnew.roll = studentgrades.roll;
Query ID = cloudera_20210316080909_e4da0a3a-4031-4970-b617-54c64d5eaf4b
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20210316080909_e4da0a3a-4031-4970-b617-
54c64d5eaf4b.log
2021-03-16 08:10:06  Starting to launch local task to process map join;     maximum memory =
1013645312
2021-03-16 08:10:09  Dump the side-table for tag: 1 with group count: 5 into file:
file:/tmp/cloudera/a5372133-f90e-42e5-b08f-69ca2bbed2d0/hive_2021-03-16_08-09-
57_858_5063287824997492492-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable
2021-03-16 08:10:09  Uploaded 1 File to: file:/tmp/cloudera/a5372133-f90e-42e5-b08f-
69ca2bbed2d0/hive_2021-03-16_08-09-57_858_5063287824997492492-1/-local-10003/
HashTable-Stage-3/MapJoin-mapfile01--.hashtable (370 bytes)
2021-03-16 08:10:09  End of local task; Time Taken: 2.602 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_1615902595627_0005, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1615902595627_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1615902595627_0005
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-03-16 08:10:24,530 Stage-3 map = 0%,  reduce = 0%
2021-03-16 08:10:37,178 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.62 sec
MapReduce Total cumulative CPU time: 2 seconds 620 msec
Ended Job = job_1615902595627_0005
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.62 sec   HDFS Read: 5994 HDFS Write: 65 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 620 msec
OK
31      prajakta        A
15      rishita B
17      riya    C
22      sakshi  B
20      priya   A
Time taken: 40.446 seconds, Fetched: 5 row(s)
```

Create index on Flight information Table

```
hive> CREATE INDEX hbasefltnew_index ON TABLE hbase_flight_new (fsch_delay)
AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler' WITH
DEFERRED REBUILD;
OK
Time taken: 0.431 seconds
```

```
hive> SHOW INDEX ON hbase_flight_new;
OK
hbasefltnew_index
hbase_flight_new
fsch_delay
default__hbase_flight_new_hbasefltnew_index__ compact
Time taken: 0.493 seconds, Fetched: 1 row(s)
hive> show tables;
OK
default__hbase_flight_new_hbasefltnew_index__
hbase_flight_newhive_table_student
studentinfo
studentnew
Time taken: 0.033 seconds, Fetched: 5 row(s)
```

Find the average departure delay per day in 2008.

```
hive> select sum(delay) from hbase_flight_new;
Query ID = cloudera_20210314042525_f5043e76-bc43-4117-a7b4-6b0dfd781b76
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_1615717830781_0001, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1615717830781_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1615717830781_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers:
1
2021-03-14 04:26:34,111 Stage-1 map = 0%, reduce = 0%
2021-03-14 04:27:01,703 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.02
sec
2021-03-14 04:27:26,888 Stage-1 map = 100%, reduce = 100%, Cumulative CPU
5.62 secMapReduce Total cumulative CPU time: 5 seconds 620 msec
Ended Job = job_1615717830781_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.62 sec HDFS Read: 7177
HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 620 msec
OK
33
Time taken: 96.731 seconds, Fetched: 1 row(s)
```
