Assignment 3

```
[cloudera@quickstart DSBDA]$ 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.

--CREATING AND USING SPECIFIC DATABASE
hive> create database flightDB;
OK
Time taken: 0.329 seconds
hive> use flightDB;
OK
Time taken: 0.02 seconds

--CREATING AND ALTERING TABLE
hive> create table flight_info(
    > flight_id string,
    > origin string,
    > destination string,
    > delay int
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE;
OK
Time taken: 0.511 seconds
hive> alter table flight_info change column delay total_delay int;
OK
Time taken: 0.261 seconds
hive> desc flight_info;
OK
flight_id               string
origin                  string
destination             string
total_delay             int
Time taken: 0.129 seconds, Fetched: 4 row(s)
hive> select * from flight_info;
OK
Time taken: 0.455 seconds
hive> alter table flight_info add columns(
    > departure_date date
    > );
OK

Time taken: 0.122 seconds
hive> desc flight_info;
OK
flight_id               string
origin                  string
destination             string
total_delay             int
departure_date          date
Time taken: 0.085 seconds, Fetched: 5 row(s)
hive> select * from flight_info;
OK
Time taken: 0.058 seconds

--CREATING AN EXTERNAL HIVE TABLE
hive> create external table flight_info2(
    > flight_id int,
    > origin string,
    > des string,
    > delay int
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LOCATION '/home/cloudera/Desktop/DSBDA';
OK
Time taken: 0.11 seconds
hive> select * from flight_info;
OK
Time taken: 0.057 seconds

--LOADING DATA FROM EXTERNAL FILE INTO THE TABLE
hive> load data local inpath '/home/cloudera/Desktop/DSBDA/test.csv' overwrite into table flight_info;
Loading data to table flightdb.flight_info
Table flightdb.flight_info stats: [numFiles=1, numRows=0, totalSize=279, rawDataSize=0]
OK
Time taken: 0.667 seconds
--CHECKING IF WORKED
hive> select * from flight_info;
OK
F001    Mumabi  Delhi   45      2025-02-15
F002    Mumabi  Nagpur  -15     2025-02-17
F003    Mumabi  Indore  30      2025-02-18
F004    Kolkata Inodore 0       2025-02-20
F005    Kochin  Srinagar        60      2025-02-22
```

```
F006    Hyderabad       Lucknow 0       2025-02-25
F007    Mumabi  Ahemdabad       20      2025-02-26
F008    Delhi   Ahemdabad       -20     2025-02-26
        NULL    NULL    NULL    NULL
        NULL    NULL    NULL    NULL
        NULL    NULL    NULL    NULL
Time taken: 0.056 seconds, Fetched: 11 row(s)

--INSERTING DATA MANUALLY INTO THE TABLE
hive> insert into table flight_info values
    > ('F009','Mumbai','Pune',0,'2025-02-28');
Query ID = cloudera_20250205205656_91181bc3-2937-4856-95f3-dce6200c6044
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_1738815374210_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1738815374210_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1738815374210_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-02-05 20:56:24,686 Stage-1 map = 0%,   reduce = 0%
2025-02-05 20:56:31,452 Stage-1 map = 100%,   reduce = 0%, Cumulative CPU 1.07 sec
MapReduce Total cumulative CPU time: 1 seconds 70 msec
Ended Job = job_1738815374210_0001
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://quickstart.cloudera:8020/user/hive/warehouse/flightdb.db/flight_info/.hive-
staging_hive_2025-02-05_20-56-11_634_8364896290987674265-1/-ext-10000
Loading data to table flightdb.flight_info
Table flightdb.flight_info stats: [numFiles=2, numRows=1, totalSize=309, rawDataSize=29]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 1.07 sec   HDFS Read: 4906 HDFS Write: 106 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 70 msec
OK
Time taken: 21.308 seconds
hive> select * from flight_info;
OK
F009    Mumbai  Pune    0       2025-02-28
F001    Mumabi  Delhi   45      2025-02-15
F002    Mumabi  Nagpur  -15     2025-02-17
F003    Mumabi  Indore  30      2025-02-18
F004    Kolkata Inodore 0       2025-02-20
F005    Kochin  Srinagar        60      2025-02-22
F006    Hyderabad       Lucknow 0       2025-02-25
F007    Mumabi  Ahemdabad       20      2025-02-26
F008    Delhi   Ahemdabad       -20     2025-02-26
        NULL    NULL    NULL    NULL
        NULL    NULL    NULL    NULL
        NULL    NULL    NULL    NULL
Time taken: 0.056 seconds, Fetched: 12 row(s)

--CREATING INDEX ON COLUMN 'depature_date' IN TABLE
hive> create index flight_index
    > on table flight_info(departure_date)
    > as 'compact'
    > with deferred rebuild;
OK
Time taken: 0.291 seconds
--SHOWING THE INDEX
hive> show index on flight_info;
OK
flight_index            flight_info             departure_date          flightdb__flight_info_flight_index__
compact
Time taken: 0.083 seconds, Fetched: 1 row(s)
hive> describe formatted flight_info;
OK
# col_name              data_type               comment

flight_id               string
origin                  string
destination             string
total_delay             int
departure_date          date

# Detailed Table Information
Database:               flightdb
Owner:                  cloudera
CreateTime:             Wed Feb 05 20:38:48 PST 2025
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/user/hive/warehouse/flightdb.db/flight_info
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   true
        last_modified_by        cloudera
```

```
        last_modified_time       1738817192
        numFiles                 2
        numRows                  1
        rawDataSize              29
        totalSize                309
        transient_lastDdlTime    1738817792

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim              ,
        line.delim               \n
        serialization.format     ,
Time taken: 0.169 seconds, Fetched: 39 row(s)


--CALCULATING AVERAGE DELAY IN MONTH OF FEB
hive> select avg(delay) as avg_delay_feb
    > from flight_info
    > ;
FAILED: SemanticException [Error 10004]: Line 1:11 Invalid table alias or column reference 'delay': (possible
column names are: flight_id, origin, destination, total_delay, departure_date)
hive> select avg(total_delay) as avg_delay_feb
    > from flight_info;
Query ID = cloudera_20250205210404_b188b977-2bed-4ce7-a00e-4acc46f4632f
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_1738815374210_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1738815374210_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1738815374210_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2025-02-05 21:04:31,441 Stage-1 map = 0%,  reduce = 0%
2025-02-05 21:04:37,792 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
2025-02-05 21:04:45,214 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.8 sec
MapReduce Total cumulative CPU time: 1 seconds 800 msec
Ended Job = job_1738815374210_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.8 sec   HDFS Read: 9025 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 800 msec
OK
13.333333333333334
Time taken: 22.405 seconds, Fetched: 1 row(s)
hive> exit;
```