Flight Information System using HBase and HiveQL

a. Create Flight Info HBase Table

This step involves creating an HBase table that holds flight number, source, destination, year, delay information, and other relevant flight data. The table should also consider delay and schedule data.

b. Demonstrate Creating, Dropping, and Altering Database Tables in HBase

The following HiveQL commands demonstrate creating a database, creating and altering tables, and inserting records.

```
[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> create database dsbda;
OK
Time taken: 0.676 seconds

hive> use dsbda;
OK
Time taken: 0.077 seconds

hive> create table flight(fno int, source varchar(10), year int, delay float);
OK
Time taken: 0.585 seconds

hive> alter table flight rename to air_flight;
OK
Time taken: 0.271 seconds
hive> alter table air_flight add columns(dest varchar(10));
OK
Time taken: 0.194 seconds

hive> drop table flight;
OK
Time taken: 0.042 seconds

hive> create table flight(fno int, source varchar(10), year int, delay float)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n'
```

```
   > stored as textfile;
OK
Time taken: 0.111 seconds

hive> insert into flight values(215, "pune", 2023, 15.00);
Query ID = cloudera_20250428233535_a1424e46-28ee-46e9-96f7-1a15a6d547ad
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_1745385658123_0023, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1745385658123_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745385658123_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-04-28 23:35:51,641 Stage-1 map = 0%,  reduce = 0%
2025-04-28 23:36:31,185 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.3 sec
MapReduce Total cumulative CPU time: 9 seconds 300 msec
Ended Job = job_1745385658123_0023
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/dsbda.db/flight/.hive-
staging_hive_2025-04-28_23-35-28_935_6785641868858050669-1/-ext-10000
Loading data to table dsbda.flight
Table dsbda.flight stats: [numFiles=1, numRows=1, totalSize=19, rawDataSize=18]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 9.3 sec   HDFS Read: 4162 HDFS Write: 87 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 300 msec
OK
Time taken: 69.789 seconds

hive> insert into flight values(216, "nagpur", 2024, 10.00);
Query ID = cloudera_20250428233939_015a7267-5f51-4080-a72f-9cf867e2f277
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_1745385658123_0024, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1745385658123_0024/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745385658123_0024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-04-28 23:40:03,835 Stage-1 map = 0%,  reduce = 0%
2025-04-28 23:40:41,931 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
MapReduce Total cumulative CPU time: 7 seconds 170 msec
```

```
Ended Job = job_1745385658123_0024
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/dsbda.db/flight/.hive-
staging_hive_2025-04-28_23-39-52_054_2152873723054585347-1/-ext-10000
Loading data to table dsbda.flight
Table dsbda.flight stats: [numFiles=2, numRows=2, totalSize=40, rawDataSize=38]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 9.98 sec   HDFS Read: 4258 HDFS Write: 89
SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 980 msec
OK
Time taken: 55.407 seconds

hive> select * from flight;
OK
215     pune    2023    15.0
216     nagpur 2024    10.0
Time taken: 0.296 seconds, Fetched: 2 row(s)

hive> load data local inpath "ipp.txt"
   > overwrite into table flight;
Loading data to table dsbda.flight
Table dsbda.flight stats: [numFiles=1, numRows=0, totalSize=128, rawDataSize=0]
OK
Time taken: 0.55 seconds

hive> select * from flight;
OK
215     pune    NULL    10.0
216     nagpur          NULL    15.0
217     amravati        NULL    10.0
218     mumbai          NULL    15.0
219     bangalore       NULL    12.0
Time taken: 0.131 seconds, Fetched: 5 row(s)

hive> drop table flight;
OK
Time taken: 0.174 seconds

hive> create table flight(fno int, source string, year int, delay float)
```

```
   > row format delimited
   > fields terminated by ','
   > lines terminated by '\n'
   > stored as textfile;
OK
Time taken: 0.153 seconds

hive> insert into flight values(215, "pune", 2023, 15.00);
Query ID = cloudera_20250429000000_418a06e8-10a2-41bf-95c9-514f7d66f220
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_1745385658123_0025, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1745385658123_0025/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745385658123_0025
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-04-29 00:00:15,637 Stage-1 map = 0%,  reduce = 0%
2025-04-29 00:00:25,047 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
MapReduce Total cumulative CPU time: 2 seconds 510 msec
Ended Job = job_1745385658123_0025
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/dsbda.db/flight/.hive-
staging_hive_2025-04-29_00-00-05_015_1629466954543588115-1/-ext-10000
Loading data to table dsbda.flight
Table dsbda.flight stats: [numFiles=1, numRows=1, totalSize=19, rawDataSize=18]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.51 sec   HDFS Read: 3991 HDFS Write: 87
SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 510 msec
OK
Time taken: 22.523 seconds

hive> select * from flight;
OK
215     pune    2023    15.0
Time taken: 0.153 seconds, Fetched: 1 row(s)

hive> load data local inpath "ipp.txt"
   > overwrite into table flight;
Loading data to table dsbda.flight
```

```
Table dsbda.flight stats: [numFiles=1, numRows=0, totalSize=128, rawDataSize=0]
OK
Time taken: 0.389 seconds

hive> select * from flight;
OK
215     pune    NULL    10.0
216     nagpur          NULL    15.0
217     amravati        NULL    10.0
218     mumbai          NULL    15.0
219     bangalore       NULL    12.0
Time taken: 0.104 seconds, Fetched: 5 row(s)

hive> load data local inpath "ipp.txt"
    > overwrite into table flight;
Loading data to table dsbda.flight
Table dsbda.flight stats: [numFiles=1, numRows=0, totalSize=110, rawDataSize=0]
OK
Time taken: 0.403 seconds


hive> select * from flight;
OK
215     pune    2023    10.0
216     nagpur 2024     15.0
217     amaravati       2024    10.0
218     mumbai          2025    15.0
219     bangalore       2025    12.0
NULL    NULL    NULL    NULL
Time taken: 0.081 seconds, Fetched: 6 row(s)

hive> load data local inpath "ipp.txt"
    > overwrite into table flight;
Loading data to table dsbda.flight
Table dsbda.flight stats: [numFiles=1, numRows=0, totalSize=109, rawDataSize=0]
OK
Time taken: 0.443 seconds
hive> select * from flight;
OK
215     pune    2023    10.0
216     nagpur 2024     15.0
217     amaravati       2024    10.0
218     mumbai          2025    15.0
```

```
219     bangalore       2025    12.0
Time taken: 0.106 seconds, Fetched: 5 row(s)

hive> create table nflight(fno int, dest string, year int)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n'
    > stored as textfile;
OK
Time taken: 0.163 seconds

hive> insert into nflight values(215,"agra",2023);
Query ID = cloudera_20250429000505_c4357e77-350b-4822-b0da-5f3eb8fbc197
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_1745385658123_0026, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1745385658123_0026/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745385658123_0026
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-04-29 00:07:21,469 Stage-1 map = 0%,  reduce = 0%
2025-04-29 00:08:21,731 Stage-1 map = 0%,  reduce = 0%
2025-04-29 00:09:05,171 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.03 sec
MapReduce Total cumulative CPU time: 13 seconds 30 msec
Ended Job = job_1745385658123_0026
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/dsbda.db/nflight/.hive-
staging_hive_2025-04-29_00-05-45_262_7416856106102140297-1/-ext-10000
Loading data to table dsbda.nflight
Table dsbda.nflight stats: [numFiles=1, numRows=1, totalSize=14, rawDataSize=13]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 13.03 sec   HDFS Read: 3719 HDFS Write: 83
SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 30 msec
OK
Time taken: 210.439 seconds

hive> select * from nflight;
OK
215     agra    2023
```

```
Time taken: 0.209 seconds, Fetched: 1 row(s)

hive> select a.fno,a.source,a.year,a.delay,b.dest
    > from flight a join nflight b
    > on(a.fno=b.fno);
Query ID = cloudera_20250429001010_9d54e1ac-7e9d-4e53-8f8b-a35a2fcf8683
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20250429001010_9d54e1ac-7e9d-4e53-8f8b-a35a2fcf8683.log
2025-04-29 12:10:51   Starting to launch local task to process map join;          maximum memory = 932184064
2025-04-29 12:11:24   Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/63806aeb-452a-4150-af97-3f96b55130cd/hive_2025-04-29_00-10-34_456_3690174021450178404-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable
2025-04-29 12:11:24   Uploaded 1 File to: file:/tmp/cloudera/63806aeb-452a-4150-af97-3f96b55130cd/hive_2025-04-29_00-10-34_456_3690174021450178404-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (285 bytes)
2025-04-29 12:11:24   End of local task; Time Taken: 33.628 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_1745385658123_0027, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1745385658123_0027/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745385658123_0027
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2025-04-29 00:13:31,653 Stage-3 map = 0%,  reduce = 0%
2025-04-29 00:14:32,349 Stage-3 map = 0%,  reduce = 0%
2025-04-29 00:15:16,169 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 14.41 sec
MapReduce Total cumulative CPU time: 14 seconds 410 msec
Ended Job = job_1745385658123_0027
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 15.25 sec   HDFS Read: 6518 HDFS Write: 24 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 250 msec
OK
215     pune    2023    10.0    agra
Time taken: 288.519 seconds, Fetched: 1 row(s)



hive> create table hive_int(id int, name string, salary float)
```

```
   > row format delimited
   > fields terminated by ','
   > lines terminated by '\n'
   > stored as textfile;
OK
Time taken: 5.005 seconds

hive>  load data local inpath 'file.txt' into table hive_int;
Loading data to table default.hive_int
Table default.hive_int stats: [numFiles=1, totalSize=44]
OK
Time taken: 5.429 seconds

hive> select *from hive_int
   > ;
OK
101     shravani        500000.0
102     sampada         100000.65
Time taken: 1.567 seconds, Fetched: 2 row(s)

hive> create external table hive_ext(id int, name string, sal float)
   > row format delimited
   > fields terminated by ','
   > lines terminated by '\n'
   > stored as textfile;
OK
Time taken: 0.683 seconds

hive> insert into hive_ext select * from hive_int;
Query ID = cloudera_20250429024343_392ed7e1-4e22-412a-82a3-7452c3898ee6
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_1745385658123_0028, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1745385658123_0028/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745385658123_0028
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-04-29 02:43:42,505 Stage-1 map = 0%,  reduce = 0%
2025-04-29 02:44:04,378 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.44 sec
MapReduce Total cumulative CPU time: 3 seconds 440 msec
Ended Job = job_1745385658123_0028
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/hive_ext/.hive-staging_hive_2025-04-29_02-43-02_770_1408128689746765651-1/-ext-10000
Loading data to table default.hive_ext
Table default.hive_ext stats: [numFiles=1, numRows=2, totalSize=44, rawDataSize=42]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.44 sec   HDFS Read: 3246 HDFS Write: 116 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 440 msec
OK
Time taken: 64.954 seconds

hive> select * from hive_ext;
OK
101     shravani        500000.0
102     sampada         100000.65
Time taken: 0.327 seconds, Fetched: 2 row(s)

hive> show tables;
OK
hive_ext
hive_int
Time taken: 0.362 seconds, Fetched: 2 row(s)



hive> create external table hive_ext1(id int, name string, sal float)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES("hbase.columns.mapping"=":key,cf:name,cf:sal")
    > TBLPROPERTIES("hbase.table.name"="hive_table");
OK
Time taken: 0.945 seconds

hive> INSERT OVERWRITE TABLE hive_ext1
    > SELECT id, name, sal FROM hive_ext1;
Query ID = cloudera_20250429084545_f761988f-16a6-4046-ae93-5dbbddf6c7b5
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_1745385658123_0029, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1745385658123_0029/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745385658123_0029
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
```

```
2025-04-29 08:47:12,486 Stage-0 map = 0%,  reduce = 0%
2025-04-29 08:48:13,561 Stage-0 map = 0%,  reduce = 0%
2025-04-29 08:49:15,511 Stage-0 map = 0%,  reduce = 0%
2025-04-29 08:49:25,820 Stage-0 map = 100%,  reduce = 0%, Cumulative CPU 15.22 sec
MapReduce Total cumulative CPU time: 15 seconds 220 msec
Ended Job = job_1745385658123_0029
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1   Cumulative CPU: 15.22 sec   HDFS Read: 10839 HDFS Write: 0
SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 220 msec
OK
Time taken: 237.997 seconds

hive> SELECT * FROM hive_ext1;
OK
1       Alice   65000.0
2       Bob     70000.0
Time taken: 1.589 seconds, Fetched: 2 row(s)
hive>
```

c. Creating an External Hive Table Connected to HBase

An external Hive table can be created using HiveQL to interface with an HBase table. This allows querying HBase data using Hive syntax.

d. Find the Total Departure Delay in Hive

Use the following HiveQL command to calculate the total departure delay:

```
SELECT SUM(delay) FROM flight;
```

e. Find the Average Departure Delay in Hive

Use the following HiveQL command to calculate the average departure delay:

```
SELECT AVG(delay) FROM flight;
```

f. Create Index on Flight Information Table

Creating an index on the flight information table helps in optimizing queries. For example:

```
CREATE INDEX idx_delay ON TABLE flight (delay) AS 'COMPACT' WITH DEFERRED
REBUILD;
```