Exporting data from Hive tables to HBase tables

1) Create a table in HBase and insert data into it:

```
hbase(main):002:0> create 'Flights','finfo','fsch'
0 row(s) in 2.2850 seconds

=> Hbase::Table - Flights
hbase(main):003:0> put 'Flights',1,'finfo:source','Mumbai'
0 row(s) in 0.3900 seconds

hbase(main):004:0> put 'Flights',1,'finfo:dest','Pune'
0 row(s) in 0.0120 seconds

hbase(main):005:0> put 'Flights',1,'finfo:fno','12128'
0 row(s) in 0.0100 seconds

hbase(main):006:0> put 'Flights',2,'finfo:source','Navi Mumbai'
0 row(s) in 0.0180 seconds

hbase(main):007:0> put 'Flights',2,'finfo:dest','Pune'
0 row(s) in 0.0170 seconds

hbase(main):008:0> put 'Flights',2,'finfo:fno','12126'
0 row(s) in 0.0190 seconds

hbase(main):009:0> scan 'Flights'
ROW                                                    COLUMN+CELL
 1                                                     column=finfo:dest, timestamp=1715340615106, value=Pune
 1                                                     column=finfo:fno, timestamp=1715340632848, value=12128
 1                                                     column=finfo:source, timestamp=1715340601707, value=Mumbai
 2                                                     column=finfo:dest, timestamp=1715340669108, value=Pune
 2                                                     column=finfo:fno, timestamp=1715340680901, value=12126
 2                                                     column=finfo:source, timestamp=1715340656975, value=Navi Mumbai
2 row(s) in 0.0600 seconds
```

```
hbase(main):001:0> put 'Flights',1,'fsch:arrival','10:00 am'
0 row(s) in 1.0160 seconds

hbase(main):002:0> put 'Flights',2,'fsch:arrival','07:00 am'
0 row(s) in 0.0090 seconds

hbase(main):003:0> put 'Flights',1,'fsch:departure','10:00 pm'
0 row(s) in 0.0090 seconds

hbase(main):004:0> put 'Flights',2,'fsch:departure','06:00 pm'
0 row(s) in 0.0080 seconds
```

```
hbase(main):001:0> put 'Flights',1,'fsch:arrival','10:00 am'
0 row(s) in 1.0160 seconds

hbase(main):002:0> put 'Flights',2,'fsch:arrival','07:00 am'
0 row(s) in 0.0090 seconds

hbase(main):003:0> put 'Flights',1,'fsch:departure','10:00 pm'
0 row(s) in 0.0090 seconds

hbase(main):004:0> put 'Flights',2,'fsch:departure','06:00 pm'
0 row(s) in 0.0080 seconds

hbase(main):005:0> scan 'Flights'
ROW                                                   COLUMN+CELL
 1                                                    column=finfo:dest, timestamp=1715340615106, value=Pune
 1                                                    column=finfo:fno, timestamp=1715340632848, value=12128
 1                                                    column=finfo:source, timestamp=1715340601707, value=Mumbai
 1                                                    column=fsch:arrival, timestamp=1715340992527, value=10:00 am
 1                                                    column=fsch:departure, timestamp=1715341035641, value=10:00 pm
 2                                                    column=finfo:dest, timestamp=1715340669108, value=Pune
 2                                                    column=finfo:fno, timestamp=1715340680901, value=12126
 2                                                    column=finfo:source, timestamp=1715340656975, value=Navi Mumbai
 2                                                    column=fsch:arrival, timestamp=1715341004208, value=07:00 am
 2                                                    column=fsch:departure, timestamp=1715341058473, value=06:00 pm
2 row(s) in 0.2050 seconds
```

```
hbase(main):006:0> put 'Flights',2,'fsch:delay',40
0 row(s) in 0.0150 seconds

hbase(main):007:0> put 'Flights',1,'fsch:delay',20
0 row(s) in 0.0110 seconds

hbase(main):008:0> scan 'Flights'
ROW                                                   COLUMN+CELL
 1                                                    column=finfo:dest, timestamp=1715340615106, value=Pune
 1                                                    column=finfo:fno, timestamp=1715340632848, value=12128
 1                                                    column=finfo:source, timestamp=1715340601707, value=Mumbai
 1                                                    column=fsch:arrival, timestamp=1715340992527, value=10:00 am
 1                                                    column=fsch:delay, timestamp=1715341152973, value=20
 1                                                    column=fsch:departure, timestamp=1715341035641, value=10:00 pm
 2                                                    column=finfo:dest, timestamp=1715340669108, value=Pune
 2                                                    column=finfo:fno, timestamp=1715340680901, value=12126
 2                                                    column=finfo:source, timestamp=1715340656975, value=Navi Mumbai
 2                                                    column=fsch:arrival, timestamp=1715341004208, value=07:00 am
 2                                                    column=fsch:delay, timestamp=1715341145145, value=40
 2                                                    column=fsch:departure, timestamp=1715341058473, value=06:00 pm
2 row(s) in 0.0550 seconds
```

2) Create an external table in Hive to connect to HBase

```
hive> create external table hbase_flights(
    >      fno int,
    >      fsource string,
    >      fdest string,
    >      farrival string,
    >      fdeparture string,
    >      delay int)
    >      stored by 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    >      with serdeproperties(
    >      "hbase.columns.mapping" = ":key,finfo:source,finfo:dest,fsch:arrival,fsch:departure,fsch:delay")
    >      tblproperties("hbase.table.name" = "Flights");
OK
Time taken: 7.631 seconds
```

3) Display the data from Hive table

```
hive> select * from hbase_flights;
OK
1       Mumbai  Pune    10:00 am        10:00 pm        20
2       Navi Mumbai     Pune    07:00 am        06:00 pm        40
Time taken: 2.278 seconds, Fetched: 2 row(s)
```

4) Total and Average Delay

```
hive> select sum(delay) as Total_delay, avg(delay) as Avg_delay from hbase_flights;
Query ID = cloudera_20240510045151_824d7b52-247b-4bbd-91e7-f82bfd495013
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_1715338982966_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1715338982966_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1715338982966_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-05-10 04:51:52,266 Stage-1 map = 0%,  reduce = 0%
2024-05-10 04:52:18,635 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.32 sec
2024-05-10 04:52:37,757 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.75 sec
MapReduce Total cumulative CPU time: 9 seconds 750 msec
Ended Job = job_1715338982966_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.75 sec   HDFS Read: 15226 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 750 msec
OK
60      30.0
Time taken: 79.938 seconds, Fetched: 1 row(s)
```