Hbase-Hive

1. Create table in Hbase and link with an external table in Hive
2. Create managed table in Hive, load data in the managed table,
   Create a table in Hbase and link with an external table in Hive,
   Load data from managed table in Hive into the external table in Hive

1. Commands:
   (Flight Db)

   a. Hbase:
      1. Insert

         hbase(main):002:0> create 'flight3','finfo','fsch'
         0 row(s) in 1.3340 seconds

         => Hbase::Table - flight3
         hbase(main):003:0> put 'flight3',1,'finfo:src','pune'
         0 row(s) in 0.5670 seconds

         hbase(main):004:0> put 'flight3',1,'finfo:dest','mumbai'
         0 row(s) in 0.0130 seconds

         hbase(main):005:0> put 'flight3',1,'fsch:at','10:25am'
         0 row(s) in 0.0130 seconds

         hbase(main):006:0> put 'flight3',1,'fsch:dt','10:00am'
         0 row(s) in 0.0290 seconds

         hbase(main):007:0> put 'flight3',1,'fsch:delay',5
         0 row(s) in 0.0200 seconds

         hbase(main):008:0> put 'flight3',2,'finfo:src','delhi'
         0 row(s) in 0.0130 seconds

         hbase(main):009:0> put 'flight3',2,'finfo:dest','banglore'
         0 row(s) in 0.0230 seconds

         hbase(main):010:0> put 'flight3',2,'fsch:at','8:25am'
         0 row(s) in 0.0200 seconds

         hbase(main):011:0> put 'flight3',2,'fsch:dt','7:25am'
         0 row(s) in 0.0240 seconds

```
hbase(main):012:0> put 'flight3',2,'fsch:delay',15
0 row(s) in 0.0100 seconds

hbase(main):013:0> scan 'flight3'
ROW                 COLUMN+CELL
 1                  column=finfo:dest, timestamp=1652503589195, value=mumbai
 1                  column=finfo:src, timestamp=1652503578034, value=pune
 1                  column=fsch:at, timestamp=1652503618673, value=10:25am
 1                  column=fsch:delay, timestamp=1652503651070, value=5
 1                  column=fsch:dt, timestamp=1652503631778, value=10:00am
 2                  column=finfo:dest, timestamp=1652503685757, value=banglore
 2                  column=finfo:src, timestamp=1652503669982, value=delhi
 2                  column=fsch:at, timestamp=1652503767503, value=8:25am
 2                  column=fsch:delay, timestamp=1652503793408, value=15
 2                  column=fsch:dt, timestamp=1652503780178, value=7:25am
2 row(s) in 0.0770 seconds
```

2. Create and Drop column family

```
hbase(main):014:0> alter 'flight3',NAME=>'revenue'
Updating all regions with the new schema...
0/1 regions updated.
1/1 regions updated.
Done.
0 row(s) in 2.4300 seconds

hbase(main):015:0> put 'flight3',2,'revenue:rs',10000
0 row(s) in 0.0260 seconds

hbase(main):016:0> scan 'flight3'
ROW                 COLUMN+CELL
 1                  column=finfo:dest, timestamp=1652503589195, value=mumbai
 1                  column=finfo:src, timestamp=1652503578034, value=pune
 1                  column=fsch:at, timestamp=1652503618673, value=10:25am
 1                  column=fsch:delay, timestamp=1652503651070, value=5
 1                  column=fsch:dt, timestamp=1652503631778, value=10:00am
 2                  column=finfo:dest, timestamp=1652503685757, value=banglore
 2                  column=finfo:src, timestamp=1652503669982, value=delhi
 2                  column=fsch:at, timestamp=1652503767503, value=8:25am
 2                  column=fsch:delay, timestamp=1652503793408, value=15
 2                  column=fsch:dt, timestamp=1652503780178, value=7:25am
 2                  column=revenue:rs, timestamp=1652504004617, value=10000
2 row(s) in 0.1750 seconds

hbase(main):017:0> alter 'flight3',NAME=>'revenue',METHOD=>'delete'
```

Updating all regions with the new schema...
0/1 regions updated.
1/1 regions updated.
Done.
0 row(s) in 2.4650 seconds

```
hbase(main):018:0> scan 'flight3'
ROW                    COLUMN+CELL
 1                     column=finfo:dest, timestamp=1652503589195, value=mumbai
 1                     column=finfo:src, timestamp=1652503578034, value=pune
 1                     column=fsch:at, timestamp=1652503618673, value=10:25am
 1                     column=fsch:delay, timestamp=1652503651070, value=5
 1                     column=fsch:dt, timestamp=1652503631778, value=10:00am
 2                     column=finfo:dest, timestamp=1652503685757, value=banglore
 2                     column=finfo:src, timestamp=1652503669982, value=delhi
 2                     column=fsch:at, timestamp=1652503767503, value=8:25am
 2                     column=fsch:delay, timestamp=1652503793408, value=15
 2                     column=fsch:dt, timestamp=1652503780178, value=7:25am
2 row(s) in 0.0460 seconds
```

b. Hive external Table:
1. Create
   ```
   hive> CREATE external TABLE external_table_hbase_flight(fno int, fsource string,fdest
   string,fsh_at string,fsh_dt string,fsch_delay int)
      > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
      > WITH SERDEPROPERTIES ("hbase.columns.mapping" =
      > ":key,finfo:src,finfo:dest,fsch:at,fsch:dt,fsch:delay")
      > TBLPROPERTIES ("hbase.table.name" = "flight3");
   OK
   Time taken: 4.711 seconds
   hive> select * from external_table_hbase_flight;
   OK
   1       pune    mumbai        10:25am        10:00am        5
   2       delhi   banglore      8:25am 7:25am 15
   Time taken: 1.274 seconds, Fetched: 2 row(s)
   ```

2. Avg delay
   ```
   hive> select fsh_at from external_table_hbase_flight;
   OK
   10:25am
   8:25am
   Time taken: 0.372 seconds, Fetched: 2 row(s)
   ```

hive> select avg(fsch_delay) from external_table_hbase_flight;
Query ID = cloudera_20220513220202_654ab115-829b-44b9-a12f-4dfacc9b4966
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_1652500271324_0002, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1652500271324_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1652500271324_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-05-13 22:02:34,537 Stage-1 map = 0%, reduce = 0%
2022-05-13 22:03:02,280 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.77 sec
2022-05-13 22:03:21,524 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.62 sec
MapReduce Total cumulative CPU time: 4 seconds 620 msec
Ended Job = job_1652500271324_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 4.62 sec   HDFS Read: 7962 HDFS
Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 620 msec
OK
10.0
Time taken: 79.639 seconds, Fetched: 1 row(s)

2. Commands:

a. Create Managed table in Hive and load data from local text file:

hive> create table managed_table_emp_2(eno int, ename string, esal int) row format delimited fields terminated by ',' stored as textfile;

OK

Time taken: 0.159 seconds

hive> load data local inpath '/home/cloudera/Desktop/empdbnew.txt' into table managed_table_emp_2;

Loading data to table default.managed_table_emp_2

Table default.managed_table_emp_2 stats: [numFiles=1, totalSize=65]

OK

Time taken: 0.584 seconds

hive> select * from managed_table_emp_2;

OK

1   Charles 2300

2   Max   3000

3   Checo   2200

4   Nando   2400

5   Lando   1900

Time taken: 0.125 seconds, Fetched: 5 row(s)

b..   Create External Table in Hive to connect with hbase table

    1. Table in Hbase

      hbase(main):020:0> create 'emphive3','emp'

      0 row(s) in 0.4310 seconds

      => Hbase::Table - emphive3

    2.. External Table in Hive

      hive> CREATE external TABLE external_table_hbase_emp(id int, name string, esal string)

       > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'

       > WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,emp:name,emp:esal")

       > TBLPROPERTIES ("hbase.table.name" = "emphive3");

      OK

      Time taken: 0.25 seconds

3.. Load data from Managed table in Hive(created earlier) into External table created in Hive:

hive> insert into external_table_hbase_emp select * from managed_table_emp_2;

Query ID = cloudera_20220513222222_588c0a6e-4fbe-4d12-a7ef-4d50c8d23de1

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_1652500271324_0003, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1652500271324_0003/

Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1652500271324_0003

Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0

2022-05-13 22:22:53,707 Stage-0 map = 0%, reduce = 0%

2022-05-13 22:23:11,388 Stage-0 map = 100%, reduce = 0%, Cumulative CPU 2.55 sec

MapReduce Total cumulative CPU time: 2 seconds 550 msec

Ended Job = job_1652500271324_0003

MapReduce Jobs Launched:

Stage-Stage-0: Map: 1   Cumulative CPU: 2.55 sec   HDFS Read: 3975 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 2 seconds 550 msec

OK

Time taken: 35.861 seconds

hive> select * from external_table_hbase_emp;

OK

1	Charles	2300
2	Max	3000
3	Checo	2200
4	Nando	2400
5	Lando	1900

Time taken: 0.286 seconds, Fetched: 5 row(s)

b.. Check Hbase table:

hbase(main):001:0> scan 'emphive3'

ROW   COLUMN+CELL

1   column=emp:esal, timestamp=1652505790531, value=2300

1   column=emp:name, timestamp=1652505790531, value=Charles

2   column=emp:esal, timestamp=1652505790531, value=3000

2   column=emp:name, timestamp=1652505790531, value=Max

3   column=emp:esal, timestamp=1652505790531, value=2200

3   column=emp:name, timestamp=1652505790531, value=Checo

4   column=emp:esal, timestamp=1652505790531, value=2400

4   column=emp:name, timestamp=1652505790531, value=Nando

5   column=emp:esal, timestamp=1652505790531, value=1900

5   column=emp:name, timestamp=1652505790531, value=Lando

5 row(s) in 0.7830 seconds