PROBLEM - 1

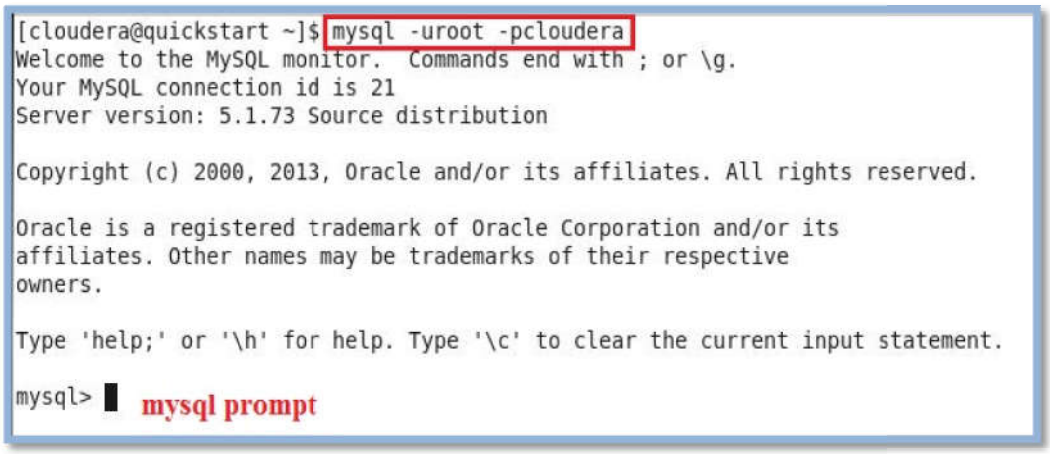
Import and Export between MYSQL and HDFS using SQOOP

Note: Cloudera quickstart VM is used to perform import and export between MYSQL and HDFS

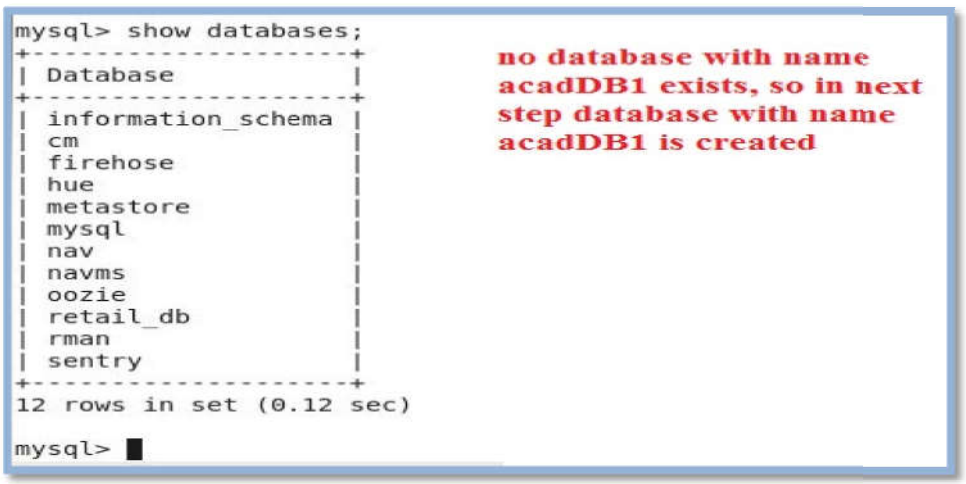
In Cloudera quickstart VM all daemons are started at the time when we start VM, so there is no need to start all the required daemons manually, like as in acadgild VM we start hadoop daemons with "start-all.sh" command and mysql service with "sudo service mysqld start" command.

To import the data from MYSQL into HDFS, below steps are followed:

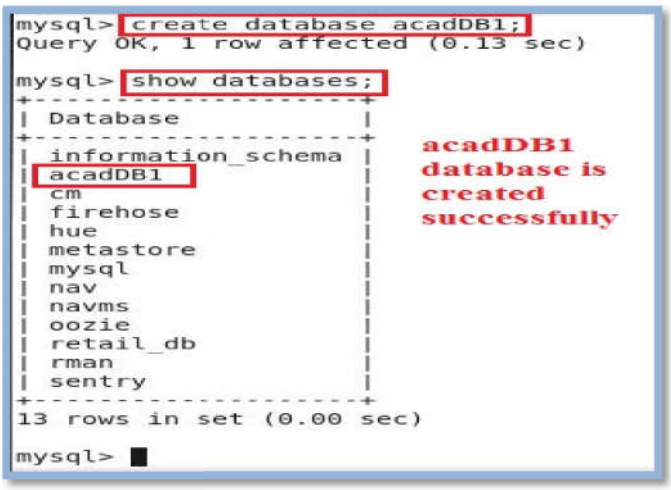
Step 1: Logged into mysql database using below command:



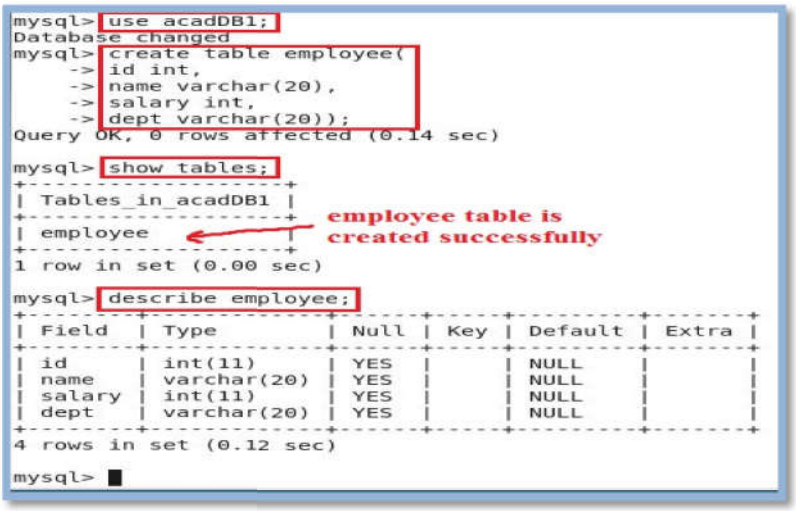
Step 2: Checked whether database with name "acadDB1" already exists using below command:



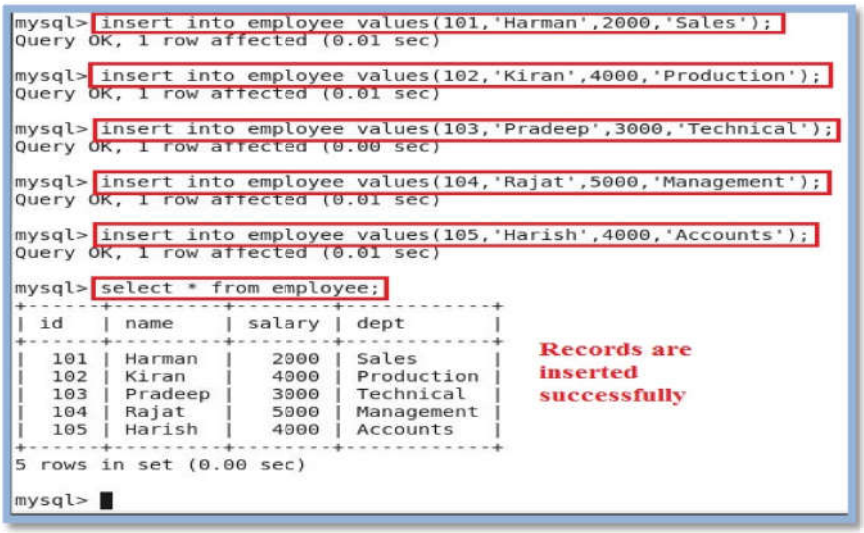
Step 3: Created Database "acadDB1" using below command:



Step 4: Created table "employee" with following fields inside acadDB1 database:



Step 5: Inserted few records inside employee table:



Step 6: Granted permission to root user to access the database over the network, followed by flushing the privileges (The reload/flush privileges command tells the server to reload the grant tables into memory), and committing all changes to database as follows:

```
mysql> grant all on *.* to 'root'@'localhost' with grant option;
Query OK, 0 rows affected (0.15 sec)

mysql> flush privileges;
Query OK, 0 rows affected (0.13 sec)

mysql> commit;
Query OK, 0 rows affected (0.00 sec)

mysql>
```

Step 7: Checked in hdfs, which directories exist inside /user, because we want data should be imported to /user/my_sqoop/sqoopout1 location, so sqoopout1 should not already exist, else during import error would have been returned if we used already existing directory name:

```
[cloudera@quickstart ~]$ hadoop fs -ls /user
Found 8 items
drwxr-xr-x   - cloudera cloudera          0 2017-08-23 08:28 /user/cloudera
drwxr-xr-x   - mapred   hadoop            0 2017-07-19 06:29 /user/history
drwxrwxrwx   - hive     supergroup        0 2017-07-19 06:31 /user/hive
drwxrwxrwx   - hue      supergroup        0 2017-07-19 06:30 /user/hue
drwxrwxrwx   - jenkins  supergroup        0 2017-07-19 06:29 /user/jenkins
drwxrwxrwx   - oozie    supergroup        0 2017-08-23 11:12 /user/oozie
drwxrwxrwx   - root     supergroup        0 2017-07-19 06:29 /user/root
drwxr-xr-x   - hdfs     supergroup        0 2017-07-19 06:31 /user/spark
[cloudera@quickstart ~]$
```

Since, my_sqoop/sqoopout1 does not exist inside /user, therefore these directories will be created automatically during import

Step 8: Using sqoop import command, imported the data from employee table of acadDB1 database inside mysql to hdfs at /user/my_sqoop/sqoopout1 location:

```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/acadDB1 -
-username root -P --table employee --target-dir /user/my_sqoop/sqoopout1 -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/08/23 12:14:33 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/08/23 12:15:51 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
17/08/23 12:15:51 INFO tool.CodeGenTool: Beginning code generation
17/08/23 12:15:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `employee` AS t LIMIT 1
17/08/23 12:15:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `employee` AS t LIMIT 1
17/08/23 12:15:53 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/ha
doop-mapreduce
```

```
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=21075
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=21075
                Total vcore-milliseconds taken by all map tasks=21075
                Total megabyte-milliseconds taken by all map tasks=21580800
        Map-Reduce Framework
                Map input records=5
                Map output records=5
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=293
                CPU time spent (ms)=2840
                Physical memory (bytes) snapshot=129871872
                Virtual memory (bytes) snapshot=1510166528
                Total committed heap usage (bytes)=60882944
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=126
17/08/23 12:17:18 INFO mapreduce.ImportJobBase: Transferred 126 bytes in 68.8346
 seconds (1.8305 bytes/sec)
17/08/23 12:17:18 INFO mapreduce.ImportJobBase: Retrieved 5 records.
[cloudera@quickstart ~]$
```

Step 9: Inside hdfs, checked using below commands whether data imported successfully or not:

```
[cloudera@quickstart ~]$ hadoop fs -ls /user
Found 9 items
drwxr-xr-x   - cloudera cloudera            0 2017-08-23 08:28 /user/cloudera
drwxr-xr-x   - mapred   hadoop              0 2017-07-19 06:29 /user/history
drwxrwxrwx   - hive     supergroup          0 2017-07-19 06:31 /user/hive
drwxrwxrwx   - hue      supergroup          0 2017-07-19 06:30 /user/hue
drwxrwxrwx   - jenkins  supergroup          0 2017-07-19 06:29 /user/jenkins
drwxr-xr-x   - cloudera supergroup          0 2017-08-23 12:16 /user/my_sqoop
drwxrwxrwx   - oozie    supergroup          0 2017-08-23 11:12 /user/oozie
drwxrwxrwx   - root     supergroup          0 2017-07-19 06:29 /user/root
drwxr-xr-x   - hdfs     supergroup          0 2017-07-19 06:31 /user/spark
[cloudera@quickstart ~]$ hadoop fs -ls /user/my_sqoop
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2017-08-23 12:17 /user/my_sqoop/sq
oopout1
[cloudera@quickstart ~]$ hadoop fs -ls /user/my_sqoop/sqoopout1
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2017-08-23 12:17 /user/my_sqoop/sq
oopout1/_SUCCESS
-rw-r--r--   1 cloudera supergroup        126 2017-08-23 12:17 /user/my_sqoop/sq
oopout1/part-m-00000
[cloudera@quickstart ~]$ hadoop fs -cat /user/my_sqoop/sqoopout1/part-m-00000
101,Harman,2000,Sales
102,Kiran,4000,Production
103,Pradeep,3000,Technical
104,Rajat,5000,Management
105,Harish,4000,Accounts
[cloudera@quickstart ~]$
```

Data has been imported successfully

To export the data from HDFS to MYSQL, below steps are followed:

Step 1: To export the data from hdfs to employee table [created during import statement], created employee.txt file [using gedit employee.txt command] in local file system first with data matching the schema of employee table:

```
[cloudera@quickstart ~]$ cat /home/cloudera/mydata/sqoop/employee.txt
106,Ajay,2500,Sales
107,Manav,4000,Production
108,Harpreet,5000,Management
109,Raman,3000,Technical
110,Aman,4500,Production
[cloudera@quickstart ~]$
```

Step 2: Put employee.txt file from local FS to hdfs at following location:

```
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/mydata/sqoop/employee.txt
 /user/my_sqoop/
[cloudera@quickstart ~]$ hadoop fs -ls /user/my_sqoop/
Found 2 items
-rw-r--r--   1 cloudera supergroup        125 2017-08-23 12:26 /user/my_sqoop/em
ployee.txt
drwxr-xr-x   - cloudera supergroup          0 2017-08-23 12:17 /user/my_sqoop/sq
oopout1
[cloudera@quickstart ~]$ hadoop fs -cat /user/my_sqoop/employee.txt
106,Ajay,2500,Sales
107,Manav,4000,Production
108,Harpreet,5000,Management
109,Raman,3000,Technical
110,Aman,4500,Production
[cloudera@quickstart ~]$
```

data is placed successfully inside hdfs from local FS

Step 3: Using export command, exported data from hdfs to mysql:

```
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/acadDB1 -
-username root -P --table employee --export-dir /user/my_sqoop/employee.txt --in
put-fields-terminated-by ',' -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/08/23 12:29:15 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/08/23 12:29:22 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
17/08/23 12:29:22 INFO tool.CodeGenTool: Beginning code generation
17/08/23 12:29:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `employee` AS t LIMIT 1
17/08/23 12:29:24 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `employee` AS t LIMIT 1
17/08/23 12:29:24 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/ha
doop-mapreduce
Note: /tmp/sqoop-cloudera/compile/9a2b8e11109268b8a86dad578c3d2630/employee.java
 uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
17/08/23 12:29:34 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-clou
dera/compile/9a2b8e11109268b8a86dad578c3d2630/employee.jar
17/08/23 12:29:34 INFO mapreduce.ExportJobBase: Beginning export of employee
17/08/23 12:29:34 INFO Configuration.deprecation: mapred.job.tracker is deprecat
ed. Instead, use mapreduce.jobtracker.address
17/08/23 12:29:36 INFO Configuration.deprecation: mapred.jar is deprecated. Inst
ead, use mapreduce.job.jar
17/08/23 12:29:42 INFO Configuration.deprecation: mapred.reduce.tasks.speculativ
e.execution is deprecated. Instead, use mapreduce.reduce.speculative
17/08/23 12:29:42 INFO Configuration.deprecation: mapred.map.tasks.speculative.e
xecution is deprecated. Instead, use mapreduce.map.speculative
17/08/23 12:29:42 INFO Configuration.deprecation: mapred.map.tasks is deprecated
```

```
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=19309
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=19309
                Total vcore-milliseconds taken by all map tasks=19309
                Total megabyte-milliseconds taken by all map tasks=19772416
        Map-Reduce Framework
                Map input records=5
                Map output records=5
                Input split bytes=139
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=449
                CPU time spent (ms)=2060
                Physical memory (bytes) snapshot=129077248
                Virtual memory (bytes) snapshot=1508073472
                Total committed heap usage (bytes)=60882944
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/08/23 12:51:47 INFO mapreduce.ExportJobBase: Transferred 267 bytes in 57.1304
 seconds (4.6735 bytes/sec)
17/08/23 12:51:47 INFO mapreduce.ExportJobBase: Exported 5 records.
[cloudera@quickstart ~]$
```

Step 4: Above screenshots show that export command ran successfully, so using select statement checked whether data in employee table inside mysql placed properly or not:

```
mysql> select * from employee;
+------+----------+--------+------------+
| id   | name     | salary | dept       |
+------+----------+--------+------------+
|  101 | Harman   |   2000 | Sales      |
|  102 | Kiran    |   4000 | Production |
|  103 | Pradeep  |   3000 | Technical  |
|  104 | Rajat    |   5000 | Management |
|  105 | Harish   |   4000 | Accounts   |
|  106 | Ajay     |   2500 | Sales      |
|  107 | Manav    |   4000 | Production |
|  108 | Harpreet |   5000 | Management |
|  109 | Raman    |   3000 | Technical  |
|  110 | Aman     |   4500 | Production |
+------+----------+--------+------------+
10 rows in set (0.00 sec)

mysql>
```

From above screenshot we can see that last 5 rows are placed from hdfs, earlier there were only first 5 rows inside employee table.

PROBLEM - 2

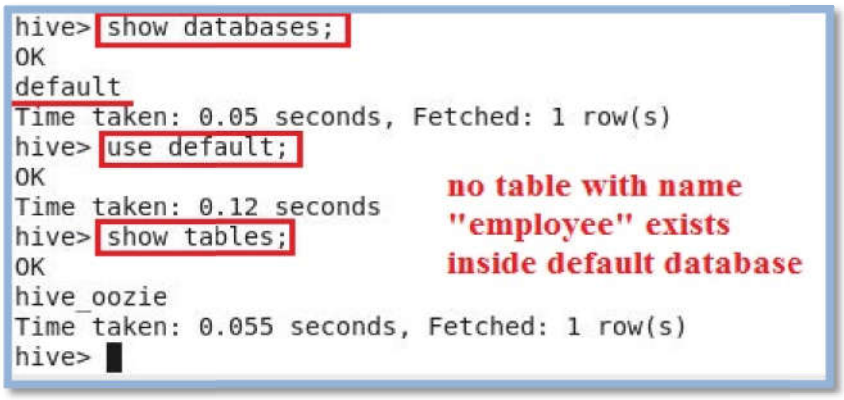
Import and Export between MYSQL and HIVE using SQOOP

NOTE: Here, we need to import and export only selected columns from mysql table to hive table and vice-versa using sqoop.

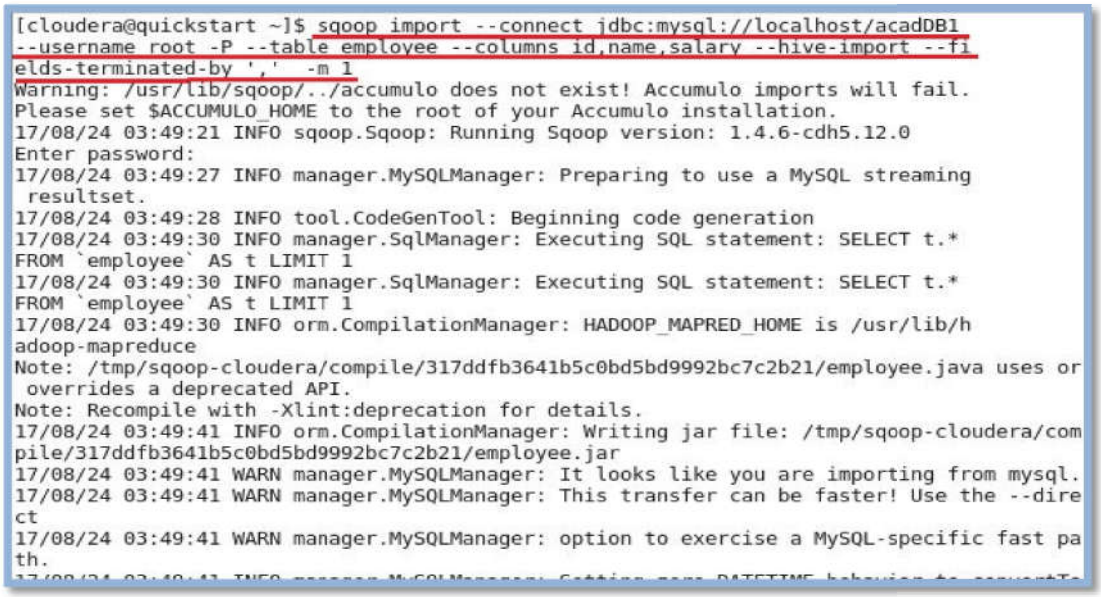
Employee table created in Problem-1 is considered to import the data from mysql to hive, and empid table is created in mysql to export only selected data from hive to mysql and it's assumed that grant and flush commands are executed in mysql.

To import the data from MYSQL into hive, below steps are followed:

Step 1: Checked whether employee table already exists inside default database in hive or not



Step 2: As employee table does not already exist inside hive, so using sqoop import command, imported selected data from mysql table to hive table [table will get created automatically with same name as in mysql]



```
        Map-Reduce Framework
                Map input records=10
                Map output records=10
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2309
                CPU time spent (ms)=3080
                Physical memory (bytes) snapshot=109101056
                Virtual memory (bytes) snapshot=1510125568
                Total committed heap usage (bytes)=60882944
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=155
17/08/24 03:53:01 INFO mapreduce.ImportJobBase: Transferred 155 bytes in 192.9478 second
s (0.8033 bytes/sec)
17/08/24 03:53:01 INFO mapreduce.ImportJobBase: Retrieved 10 records.
17/08/24 03:53:01 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `emp
loyee` AS t LIMIT 1
17/08/24 03:53:01 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-
cdh5.12.0.jar!/hive-log4j.properties
OK
Time taken: 14.045 seconds
Loading data to table default.employee
Table default.employee stats: [numFiles=1, totalSize=155]
OK
Time taken: 1.997 seconds
[cloudera@quickstart ~]$
```

Step 3: Above screenshot shows that import statement ran successfully, without any error, so checked inside hive whether table with three columns i.e. id, name, salary along with data got created or not inside default database:

```
hive> use default;
OK
Time taken: 0.042 seconds
hive> show tables;
OK
employee
hive_oozie
Time taken: 0.161 seconds, Fetched: 3 row(s)
hive> select * from employee;
OK
101     Harman  2000
102     Kiran   4000
103     Pradeep 3000
104     Rajat   5000
105     Harish  4000
106     Ajay    2500
107     Manav   4000
108     Harpreet        5000
109     Raman   3000
110     Aman    4500
Time taken: 0.388 seconds, Fetched: 10 row(s)
```

selected records imported successfully inside hive table

```
hive> dfs -ls /user/hive/warehouse;
Found 2 items
drwxrwxrwx   - cloudera supergroup          0 2017-08-24 03:53 /user/hive/wareho
use/employee
drwxrwxrwx   - oozie    supergroup          0 2017-08-23 11:15 /user/hive/wareho
use/hive_oozie
hive> dfs -ls /user/hive/warehouse/employee;
Found 1 items
-rw-r--r--   1 cloudera cloudera        155 2017-08-24 03:52 /user/hive/warehous
e/employee/part-m-00000
hive> dfs -cat /user/hive/warehouse/employee/part-m-00000;
101,Harman,2000
102,Kiran,4000
103,Pradeep,3000
104,Rajat,5000
105,Harish,4000
106,Ajay,2500
107,Manav,4000
108,Harpreet,5000
109,Raman,3000
110,Aman,4500
hive>
```

selected records imported successfully inside hive table directory

To export the data from HIVE to MYSQL, below steps are followed:

Step 1: Created "empid" table in acadGB1 database in mysql with one column i.e. id (type int), so that only id field values should get exported from hive table to mysql table [NOTE: we need to create table in mysql with the required number of fields, else export command will fail if table will not already exist in mysql]:

```
mysql> use acadDB1;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+--------------------+
| Tables_in_acadDB1 |
+--------------------+
| employee           |
+--------------------+
1 row in set (0.00 sec)

mysql> create table empid(
    -> id int);
Query OK, 0 rows affected (0.13 sec)

mysql> describe empid;
+-------+---------+------+-----+---------+-------+
| Field | Type    | Null | Key | Default | Extra |
+-------+---------+------+-----+---------+-------+
| id    | int(11) | YES  |     | NULL    |       |
+-------+---------+------+-----+---------+-------+
1 row in set (0.00 sec)

mysql> grant all on *.* to 'root'@'localhost' with grant option;
Query OK, 0 rows affected (0.17 sec)

mysql> flush privileges;
Query OK, 0 rows affected (0.04 sec)

mysql> commit;
Query OK, 0 rows affected (0.00 sec)
```

Step 2: To export only selected column i.e. id from employee table in hive to mysql empid table, used below command:

```
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/acadDB1 --usernam
e root -P --table empid --columns id --export-dir /user/hive/warehouse/employee/part-m-0
0000 --fields-terminated-by ',' -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/08/24 03:56:22 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/08/24 03:56:27 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultse
t.
17/08/24 03:56:27 INFO tool.CodeGenTool: Beginning code generation
17/08/24 03:56:30 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `emp
id` AS t LIMIT 1
17/08/24 03:56:30 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `emp
id` AS t LIMIT 1
17/08/24 03:56:30 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-map
reduce
```

```
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=40632
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=40632
                Total vcore-milliseconds taken by all map tasks=40632
                Total megabyte-milliseconds taken by all map tasks=41607168
        Map-Reduce Framework
                Map input records=10
                Map output records=10
                Input split bytes=154
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=564
                CPU time spent (ms)=2210
                Physical memory (bytes) snapshot=108822528
                Virtual memory (bytes) snapshot=1508204544
                Total committed heap usage (bytes)=60882944
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/08/24 03:58:22 INFO mapreduce.ExportJobBase: Transferred 312 bytes in 92.7731 seconds
 (3.363 bytes/sec)
17/08/24 03:58:23 INFO mapreduce.ExportJobBase: Exported 10 records.
[cloudera@quickstart ~]$
```

Step 3: Checked in mysql, whether data got successfully exported to mysql table or not:

```
mysql> select * from empid;
Empty set (0.00 sec)

mysql> select * from empid;
+------+
| id   |
+------+
|  101 |
|  102 |
|  103 |
|  104 |
|  105 |
|  106 |
|  107 |
|  108 |
|  109 |
|  110 |
+------+
10 rows in set (0.00 sec)

mysql>
```

empty set before running export command

after running sqoop export command, data got successfully exported to mysql empid table