**Incremental Import from MYSQL to HIVE Using Sqoop**

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

**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.**

**Sqoop supports two types of incremental imports:**

* **append**
* **lastmodified**

We can use the --**incremental** argument to specify the type of incremental import we want to perform.

**When to use append mode?**

We should specify the append mode when importing a table, where new rows are continually added with increasing row id values. We must specify the column containing the row’s id with --**check-column**. Sqoop imports rows where the check column has a value greater than the one specified with --**last-value**.

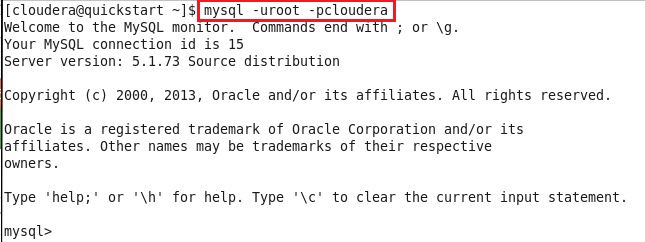
**When to use lastmodified mode?**

An alternate table **update** strategy supported by Sqoop is called lastmodified mode. This should be used when rows of the source table are updated, and each such update will set the value of a last-modified column to the current timestamp. Rows where the check column holds a timestamp more recent than the timestamp specified with --**last-value** are imported.

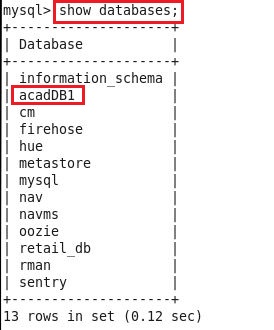
At the end of an incremental import, the value which should be specified as --**last-value** for a subsequent import is printed to the screen. When running a subsequent import, we should specify***--*last-value** in such a way to ensure we import only the new or updated data. This is handled automatically by creating an incremental import as a saved job known as **sqoop job**, which is the preferred mechanism for performing a recurring incremental import.

**To perform incremental import in hive from mysql using sqoop below steps are followed:**

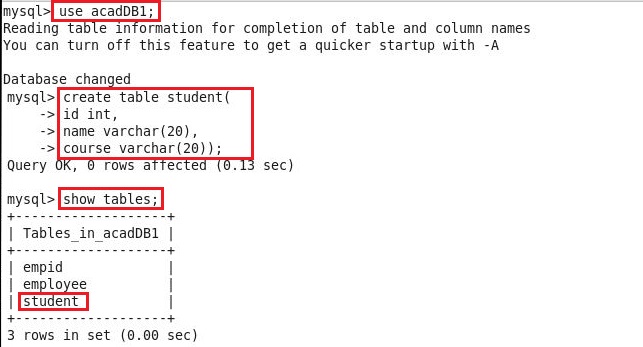
**Step 1: Logged into mysql using below command:**

****

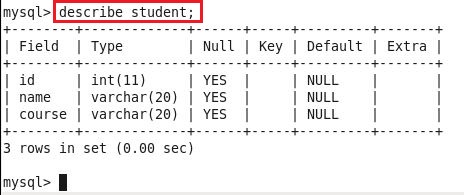
**Step 2: Checked which databases are present in mysql:**



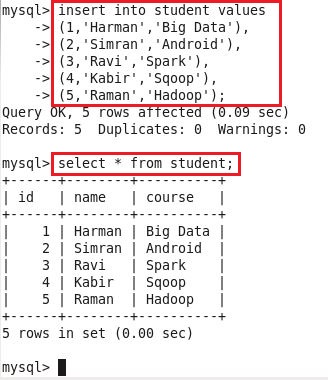
**Step 3: Inside acadDB1 database, created new student table with three fields:**

****

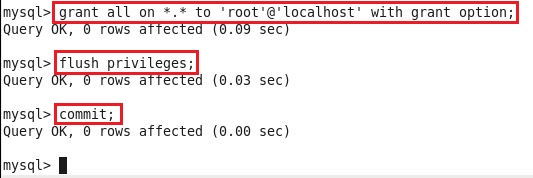
**Below command displays schema of student table:**

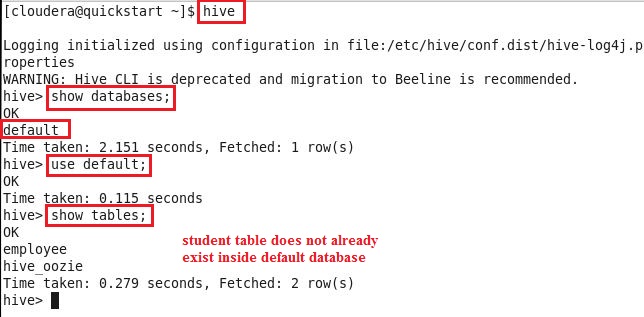
****

**Step 4: Inserted five records in student table using below command, and using select statement checked whether records got inserted successfully inside table or not:**

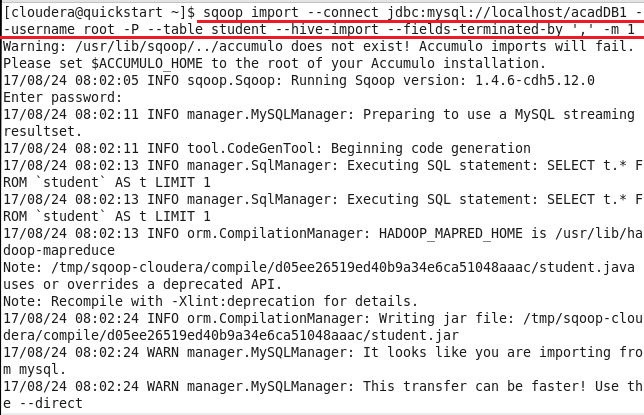
****

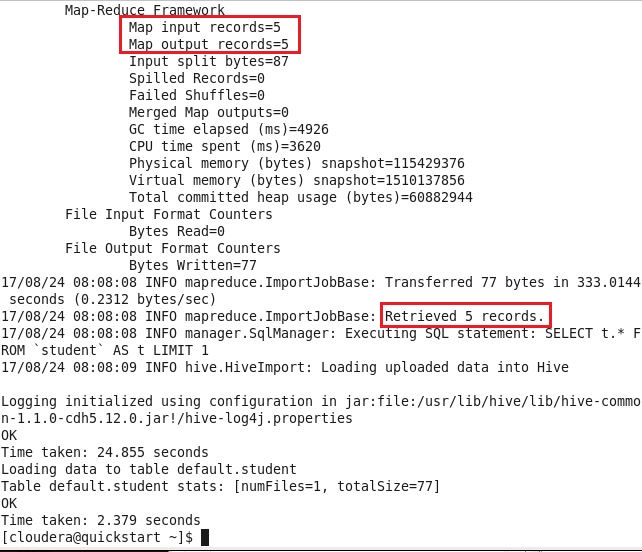
**Step 5: Granted permission to root user to access the database over the network, followed by flushing the privileges (The reload/flush privileges command tells the server to reload the grant tables into memory), and committing all changes to database as follows:**

****

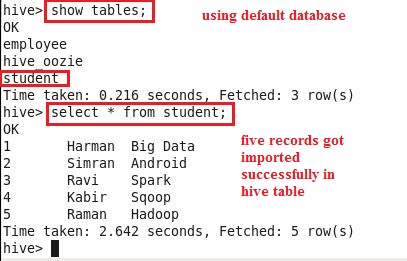
**Step 6: In hive, checked whether “student” table inside default database already exist or not:**

**Step 7: Since “student” table does not already exist in hive, therefore ran “sqoop import” command to import the data in hive table:**

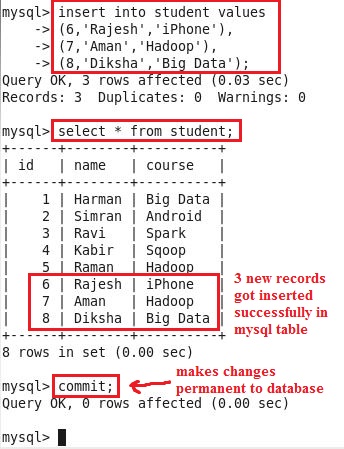
****

****

**Step 8: Above screenshots show that data got imported successfully inside hive table, so using below command checked in hive whether table with same name i.e. “student” inside default database created or not:**

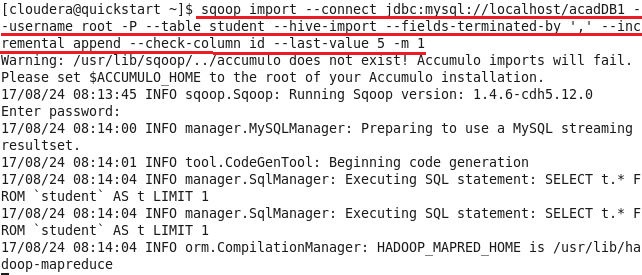
****

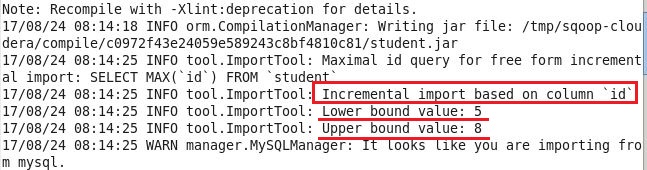
**Step 9: As table has been created inside hive, so again inserted few records in mysql table to show incremental import:**

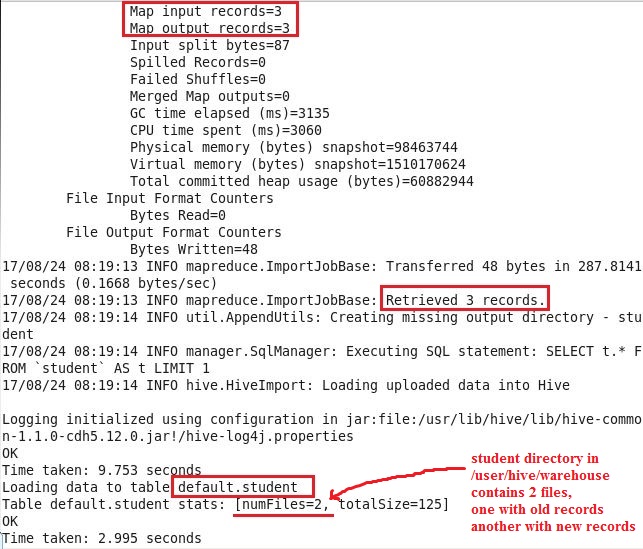
****

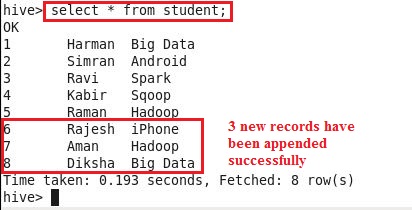
**There is no need to grant and flush privileges again, because same session of mysql is used, new session of mysql is not started.**

**Step 10: In this step, incremental import is performed with append mode, where last value of column “id” is checked, and only those records having greater value than last value of “id” are imported:**

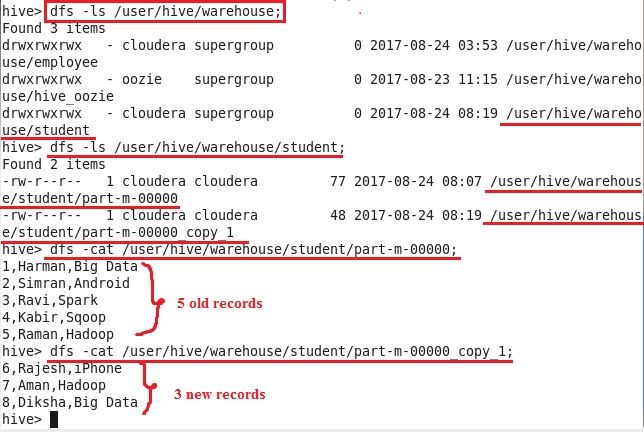
****

****

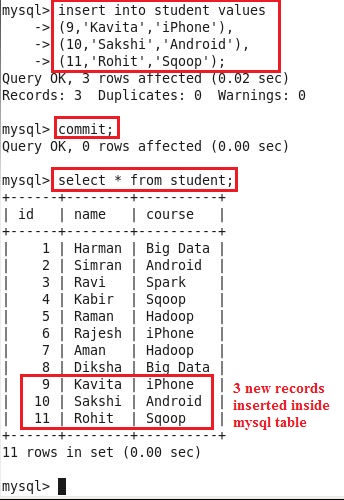
****

**Step 11: Below command shows that latest records got appended successfully inside hive table “student”:**

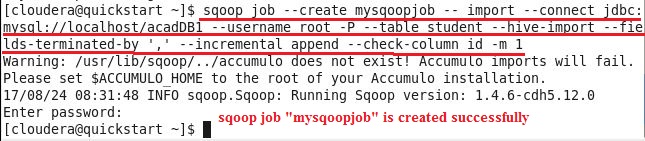
**Step 12: Below screenshot shows that inside HDFS, two copies of records are maintained in student table directory, one with old records and other with new records:**

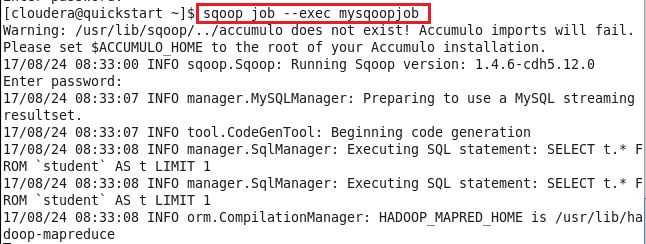
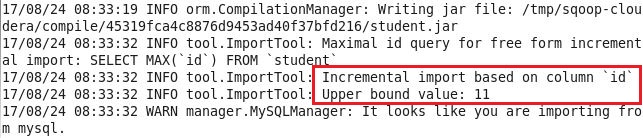
****

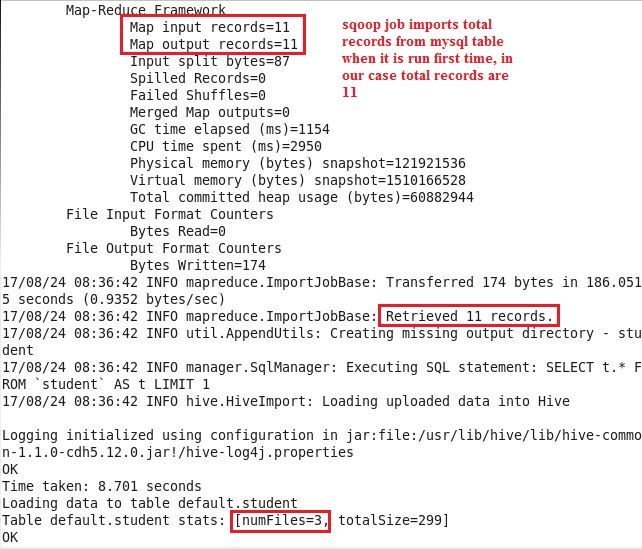
**Step 13: To show how sqoop job works, inserted few records inside mysql table:**

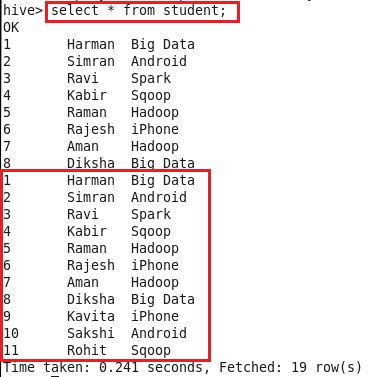
****

**Step 14: Created sqoop job “mysqoopjob” which does incremental import by checking itself the last value of --check-column argument:**

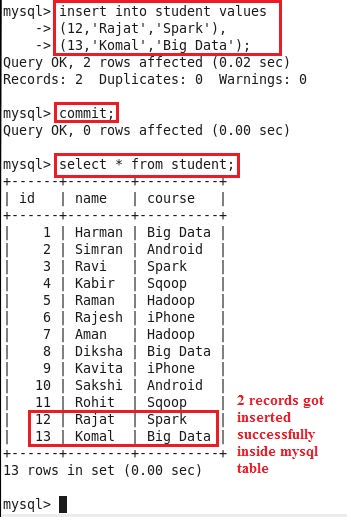
****

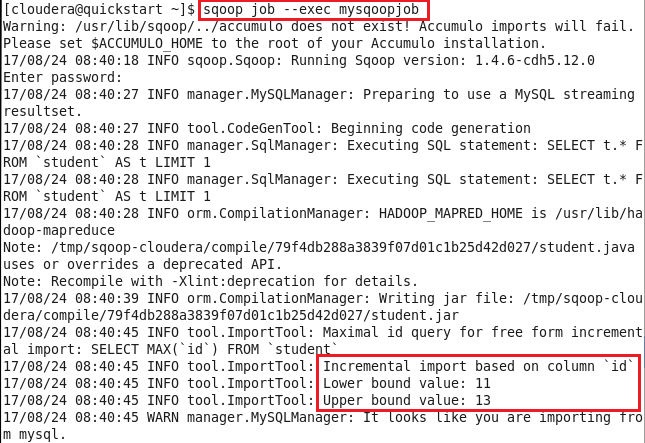
**Step 15: Executed sqoop job using below command:**

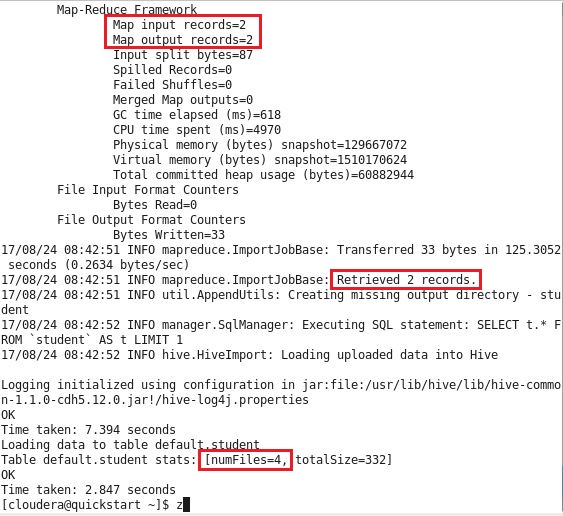
****

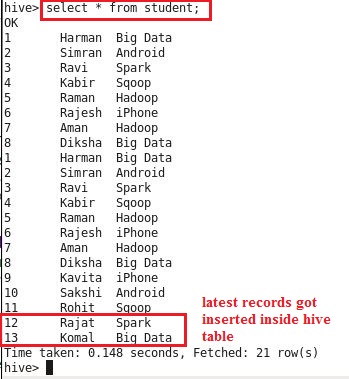
**Step 16: Below screenshot shows that, latest 3 records are not imported, however, all records are imported in hive table when sqoop job is run first time:**

**Step 17: Again inserted 2 new records inside mysql table, to show how sqoop job worked when it was executed second time:**

****

**Step 18: Executed sqoop job “mysqoopjob”:**

****

**Step 19: Below command shows that only latest records are appended in hive table, when sqoop job was run second time, we do not need to provide –last-value of –check-column argument as sqoop job automatically picks latest value of –check-column:**