

Bigdata Assignment 4.4

Perform incremental load in Hive. Read from MySQL Table and load it in Hive table. Create hive table if it does not exist. If it exists, perform the incremental load.

Solution -

- Created a database 'acadgild' and inside it a table 'student' where 5 records were inserted

create database acadgild;

use acadgild;

**create table student (id int , name varchar(20) , city
varchar(20));**

```
mysql> create database acadgild;
Query OK, 1 row affected (0.03 sec)

mysql> use acadgild;
Database changed
mysql> create table student(id int ,name varchar(20),city varchar(20));
Query OK, 0 rows affected (0.03 sec)

mysql> insert into student values
-> (1,'ritesh','bbsr'),
-> (2,'shyam','pune'),
-> (3,'ram','agra'),
-> (4,'hari','noida'),
-> (5,'alok','delhi');
Query OK, 5 rows affected (0.00 sec)
Records: 5 Duplicates: 0 Warnings: 0

mysql> █
```

- Checked the contents of the table.

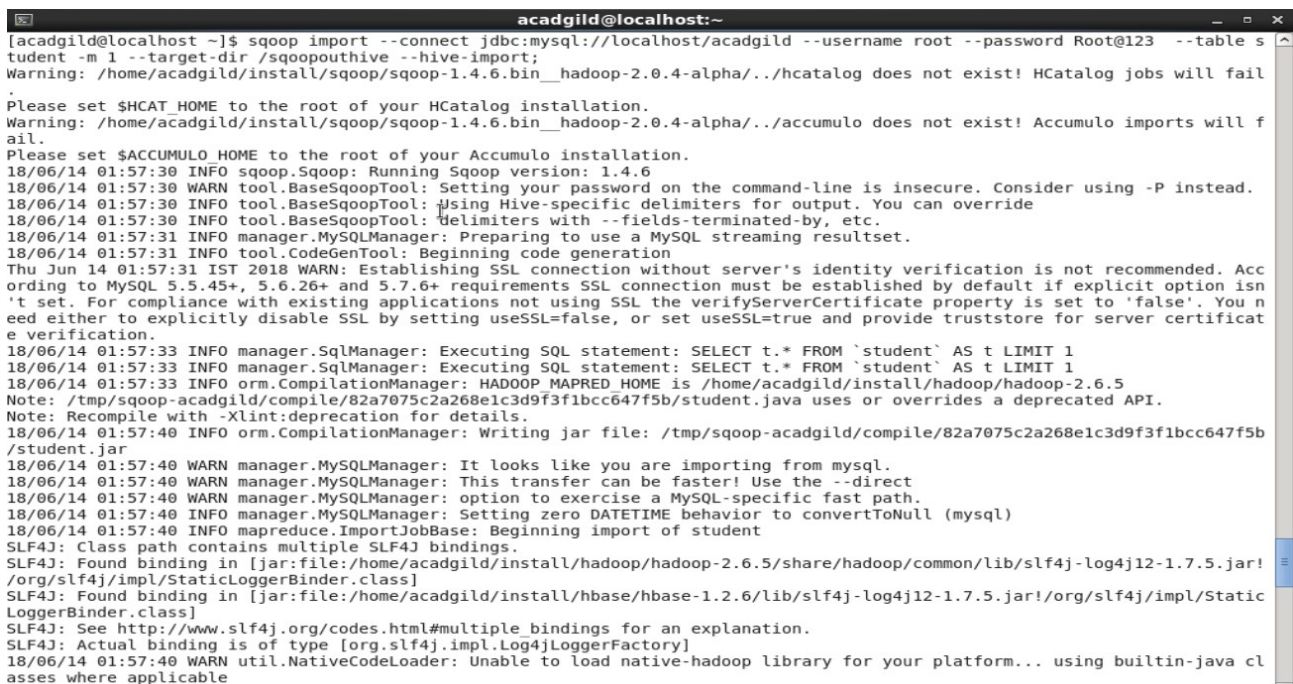
select * from student;

```
mysql> select * from student;
+----+-----+-----+
| id  | name  | city  |
+----+-----+-----+
| 1   | ritesh | bbsr  |
| 2   | shyam | pune  |
| 3   | ram   | agra  |
| 4   | hari  | noida |
| 5   | alok  | delhi |
+----+-----+-----+
5 rows in set (0.00 sec)
```

- In hive , we checked the default database , no tudent table exists.
show databases;
show tables;

```
hive> show databases;
OK
default
Time taken: 8.979 seconds, Fetched: 1 row(s)
hive> use default;
OK
Time taken: 0.05 seconds
hive> show tables;
OK
mysqlemployee
Time taken: 0.092 seconds, Fetched: 1 row(s)
hive> █
```

- We transferred the data from mysql to hive using sqoop.
sqoop import --connect jdbc:mysql://localhost/acadgild --username root --password Root@123 --table student -m 1 --target-dir /sqoopouthive --hive-import



```
acadgild@localhost:~$ sqoop import --connect jdbc:mysql://localhost/acadgild --username root --password Root@123 --table student -m 1 --target-dir /sqoopouthive --hive-import
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./hcatalog does not exist! HCatalog jobs will fail
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/06/14 01:57:30 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
18/06/14 01:57:30 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/06/14 01:57:30 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
18/06/14 01:57:30 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
18/06/14 01:57:31 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/06/14 01:57:31 INFO tool.CodeGenTool: Beginning code generation
Thu Jun 14 01:57:31 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
18/06/14 01:57:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
18/06/14 01:57:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
18/06/14 01:57:33 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/82a7075c2a268e1c3d9f3f1bcc647f5b/student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/06/14 01:57:40 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/82a7075c2a268e1c3d9f3f1bcc647f5b/student.jar
18/06/14 01:57:40 WARN manager.MySQLManager: It looks like you are importing from mysql.
18/06/14 01:57:40 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
18/06/14 01:57:40 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
18/06/14 01:57:40 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
18/06/14 01:57:40 INFO mapreduce.ImportJobBase: Beginning import of student
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/14 01:57:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
acadgild@localhost:~
18/06/14 01:58:18 INFO mapreduce.Job: map 100% reduce 0%
18/06/14 01:58:19 INFO mapreduce.Job: Job job_1528895605584_0008 completed successfully
18/06/14 01:58:20 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=127702
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=64
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=9489
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=9489
    Total vcore-milliseconds taken by all map tasks=9489
    Total megabyte-milliseconds taken by all map tasks=9716736
  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)=153
    CPU time spent (ms)=1290
    Physical memory (bytes) snapshot=120459264
    Virtual memory (bytes) snapshot=2063437824
    Total committed heap usage (bytes)=62980096
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=64
18/06/14 01:58:20 INFO mapreduce.ImportJobBase: Transferred 64 bytes in 37.8034 seconds (1.693 bytes/sec)
18/06/14 01:58:20 INFO mapreduce.ImportJobBase: Retrieved 5 records.
```

- Then , in hive we checked the content of student
select * from student;

```
acadgild
File Edit View Search Terminal Help
hive> show databases;
OK
default
Time taken: 0.052 seconds, Fetched: 1 row(s)
hive> use default;
OK
Time taken: 0.062 seconds
hive> show tables;
OK
mysqlemployee
student
Time taken: 0.045 seconds, Fetched: 2 row(s)
hive> select * from student;
OK
bbsr      1      ritesh
pune      2      shyam
agra      3      ram
noida     4      hari
delhi     5      alok
Time taken: 0.469 seconds, Fetched: 5 row(s)
hive> █
```

- Now , we added 2 records in the student table in mysql . Then checked its contents
select * from student;

```
mysql> insert into student values (6,'gopal','meerut'), (7,'karan','kanpur');
Query OK, 2 rows affected (0.08 sec)
Records: 2 Duplicates: 0 Warnings: 0

mysql> select * from student;
+----+-----+-----+
| id | name  | city  |
+----+-----+-----+
| 1  | ritesh| bbsr  |
| 2  | shyam | pune  |
| 3  | ram   | agra  |
| 4  | hari  | noida |
| 5  | alok  | delhi |
| 6  | gopal | meerut|
| 7  | karan | kanpur|
+----+-----+-----+
7 rows in set (0.00 sec)
```

- Then we performed the incremental load by using sqoop incremental
sqoop import --connect jdbc:mysql://localhost/acadgild --username root --password Root@123 --table student --check-column id --incremental append --last-value 5 -m 1 --target-dir /sqooputhive;

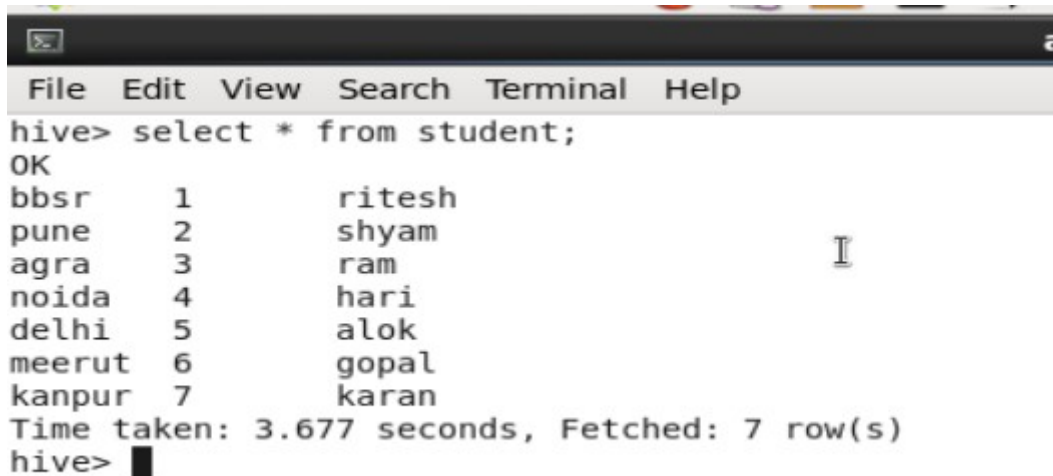
```
acadgild@localhost:~$ sqoop import --connect jdbc:mysql://localhost/acadgild --username root --password Root@123 --table student --check-column id --incremental append --last-value 5 -m 1 --target-dir /sqooputhive ;
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/06/14 02:08:03 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
18/06/14 02:08:03 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/06/14 02:08:04 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/06/14 02:08:04 INFO tool.CodeGenTool: Beginning code generation
Thu Jun 14 02:08:04 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
18/06/14 02:08:05 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
18/06/14 02:08:05 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `student` AS t LIMIT 1
18/06/14 02:08:05 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/33fb98147c7461a0b16b27aff6e6af7b/student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/06/14 02:08:08 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/33fb98147c7461a0b16b27aff6e6af7b/student.jar
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/14 02:08:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/06/14 02:08:09 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(`id`) FROM `student`
18/06/14 02:08:09 INFO tool.ImportTool: Incremental import based on column `id`
18/06/14 02:08:09 INFO tool.ImportTool: Lower bound value: 5
18/06/14 02:08:09 INFO tool.ImportTool: Upper bound value: 7
18/06/14 02:08:09 WARN manager.MySQLManager: It looks like you are importing from mysql.
18/06/14 02:08:09 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
18/06/14 02:08:09 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
```

```

acadgild@localhost:~
18/06/14 02:08:34 INFO mapreduce.Job: Job job_1528895605584_0009 completed successfully
18/06/14 02:08:34 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=127916
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=30
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=6355
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=6355
    Total vcore-milliseconds taken by all map tasks=6355
    Total megabyte-milliseconds taken by all map tasks=6507520
  Map-Reduce Framework
    Map input records=2
    Map output records=2
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=66
    CPU time spent (ms)=1460
    Physical memory (bytes) snapshot=122048512
    Virtual memory (bytes) snapshot=2063437824
    Total committed heap usage (bytes)=62980096
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=30
18/06/14 02:08:34 INFO mapreduce.ImportJobBase: Transferred 30 bytes in 24.8325 seconds (1.2081 bytes/sec)
18/06/14 02:08:34 INFO mapreduce.ImportJobBase: Retrieved 2 records.
18/06/14 02:08:34 INFO util.AppendUtils: Creating missing output directory - sqoopouthive

```

- In hive , we checked the content of the table after incremental load **select * from student;**



```

File Edit View Search Terminal Help
hive> select * from student;
OK
bbsr      1      ritesh
pune      2      shyam
agra      3      ram
noida     4      hari
delhi     5      alok
meerut    6      gopal
kanpur    7      karan
Time taken: 3.677 seconds, Fetched: 7 row(s)
hive>

```

The data is same as in the mysql after updating 2 records.