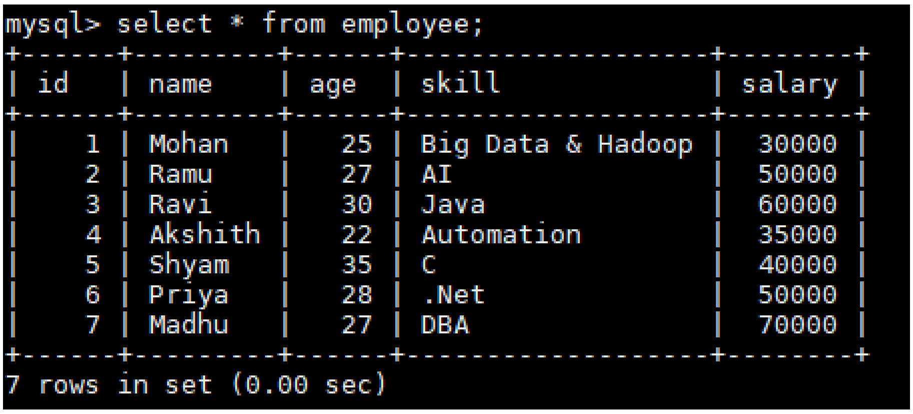
Assignment 11.2:

Problem Statement:

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.

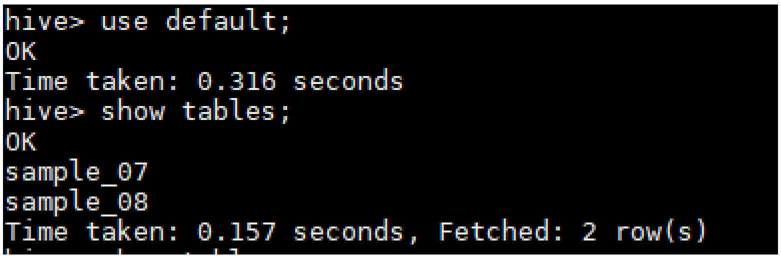
Steps:

- Below ‘employee’ table is used for the solution of the problem statement.
- **Initially loaded with 7 records as shown below:**



```
mysql> select * from employee;
+------+---------+------+-------------------+--------+
| id   | name    | age  | skill             | salary |
+------+---------+------+-------------------+--------+
|    1 | Mohan   |   25 | Big Data & Hadoop |  30000 |
|    2 | Ramu    |   27 | AI                |  50000 |
|    3 | Ravi    |   30 | Java              |  60000 |
|    4 | Akshith |   22 | Automation        |  35000 |
|    5 | Shyam   |   35 | C                 |  40000 |
|    6 | Priya   |   28 | .Net              |  50000 |
|    7 | Madhu   |   27 | DBA               |  70000 |
+------+---------+------+-------------------+--------+
7 rows in set (0.00 sec)
```

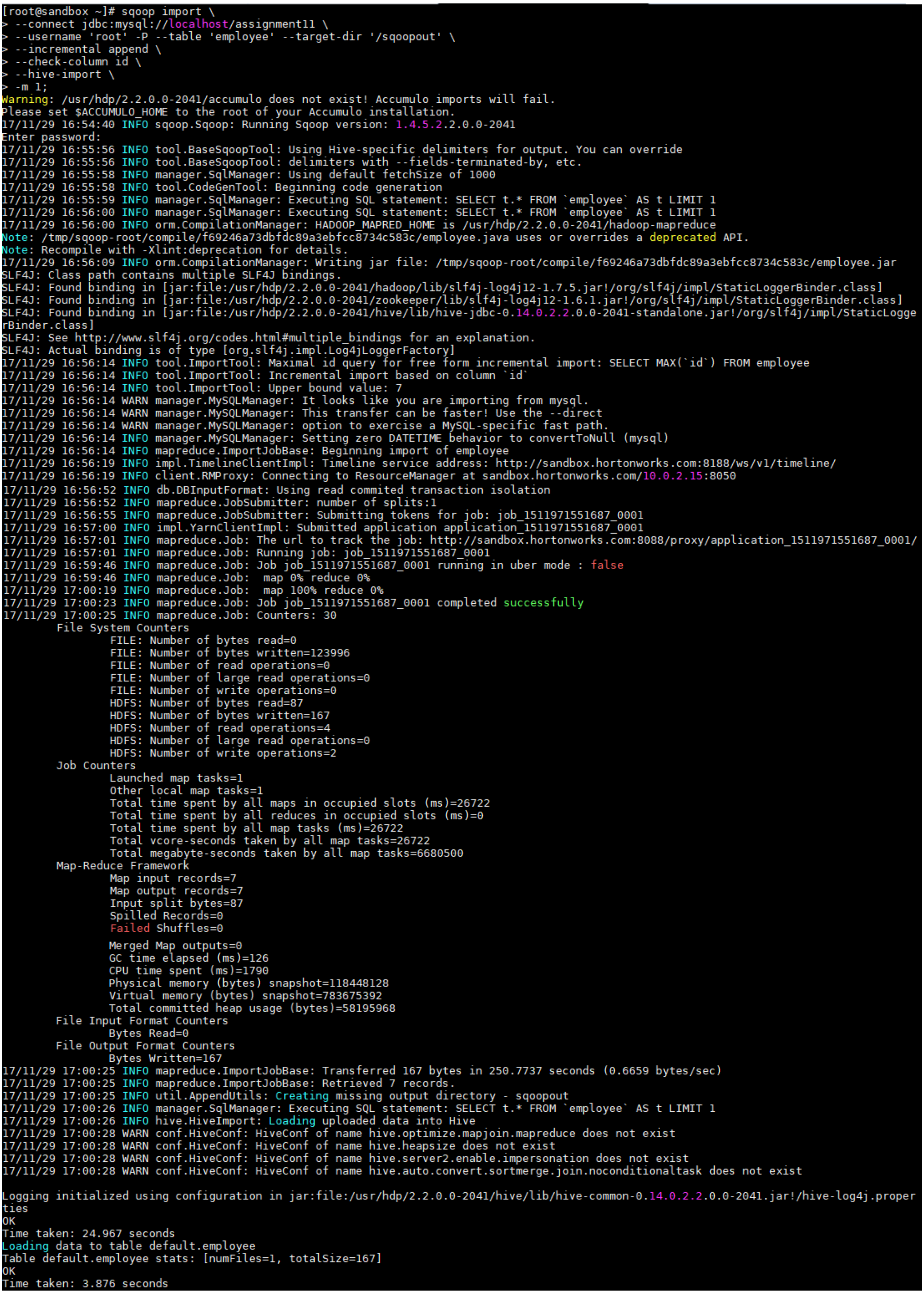
- No ‘employee’ table is present in hive, as shown below:



```
hive> use default;
OK
Time taken: 0.316 seconds
hive> show tables;
OK
sample_07
sample_08
Time taken: 0.157 seconds, Fetched: 2 row(s)
```

- **Run the sqoop import query to import data from ‘employee’ table in MySQL as shown below:**

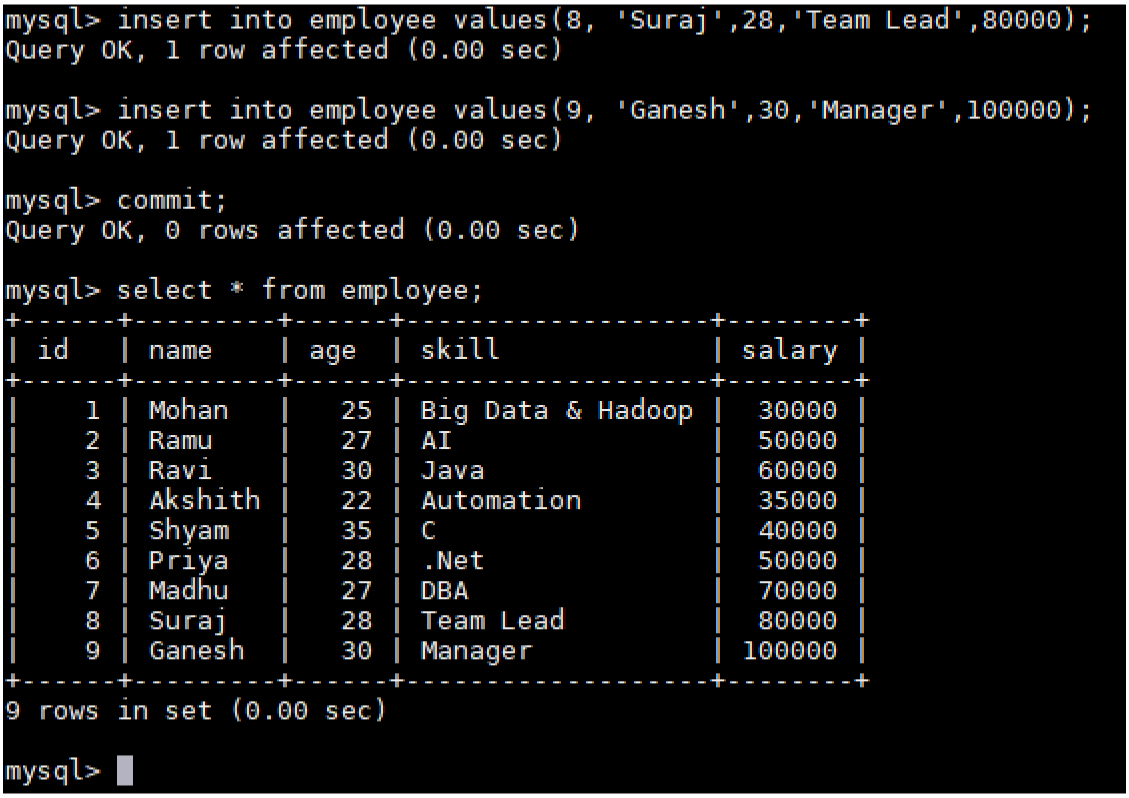
sqoop import \
--connect jdbc:mysql://localhost/assignment11 \
--username 'root' -P --table 'employee' --target-dir '/sqoopout' \
--incremental append \
--check-column id \
--hive-import \
-m 1;



```
[root@sandbox ~]# sqoop import \
> --connect jdbc:mysql://localhost/assignment11 \
> --username 'root' -P --table 'employee' --target-dir '/sqoopout' \
> --incremental append \
> --check-column id \
> --hive-import \
> -m 1;
Warning: /usr/hdp/2.2.0.0-2041/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/29 16:54:40 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5.2.2.0.0-2041
Enter password:
17/11/29 16:55:56 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
17/11/29 16:55:56 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
17/11/29 16:55:58 INFO manager.SqlManager: Using default fetchSize of 1000
17/11/29 16:55:58 INFO tool.CodeGenTool: Beginning code generation
17/11/29 16:55:59 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/29 16:56:00 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/29 16:56:00 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/hdp/2.2.0.0-2041/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/f69246a73dbfdc89a3ebfcc8734c583c/employee.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
17/11/29 16:56:09 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/f69246a73dbfdc89a3ebfcc8734c583c/employee.jar
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hadoop/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/zookeeper/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hive/lib/hive-jdbc-0.14.0.2.2.0.0-2041-standalone.jar!/org/slf4j/impl/StaticLogge
rBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
17/11/29 16:56:14 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(`id`) FROM employee
17/11/29 16:56:14 INFO tool.ImportTool: Incremental import based on column `id`
17/11/29 16:56:14 INFO tool.ImportTool: Upper bound value: 7
17/11/29 16:56:14 WARN manager.MySQLManager: It looks like you are importing from mysql.
17/11/29 16:56:14 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
17/11/29 16:56:14 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
17/11/29 16:56:14 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
17/11/29 16:56:14 INFO mapreduce.ImportJobBase: Beginning import of employee
17/11/29 16:56:19 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
17/11/29 16:56:19 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
17/11/29 16:56:52 INFO db.DBInputFormat: Using read commited transaction isolation
17/11/29 16:56:52 INFO mapreduce.JobSubmitter: number of splits:1
17/11/29 16:56:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1511971551687_0001
17/11/29 16:57:00 INFO impl.YarnClientImpl: Submitted application application_1511971551687_0001
17/11/29 16:57:01 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1511971551687_0001/
17/11/29 16:57:01 INFO mapreduce.Job: Running job: job_1511971551687_0001
17/11/29 16:59:46 INFO mapreduce.Job: Job job_1511971551687_0001 running in uber mode : false
17/11/29 16:59:46 INFO mapreduce.Job:  map 0% reduce 0%
17/11/29 17:00:19 INFO mapreduce.Job:  map 100% reduce 0%
17/11/29 17:00:23 INFO mapreduce.Job: Job job_1511971551687_0001 completed successfully
17/11/29 17:00:25 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=123996
                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=167
                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)=26722
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=26722
                Total vcore-seconds taken by all map tasks=26722
                Total megabyte-seconds taken by all map tasks=6680500
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=126
                CPU time spent (ms)=1790
                Physical memory (bytes) snapshot=118448128
                Virtual memory (bytes) snapshot=783675392
                Total committed heap usage (bytes)=58195968
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=167
17/11/29 17:00:25 INFO mapreduce.ImportJobBase: Transferred 167 bytes in 250.7737 seconds (0.6659 bytes/sec)
17/11/29 17:00:25 INFO mapreduce.ImportJobBase: Retrieved 7 records.
17/11/29 17:00:25 INFO util.AppendUtils: Creating missing output directory - sqoopout
17/11/29 17:00:26 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/29 17:00:26 INFO hive.HiveImport: Loading uploaded data into Hive
17/11/29 17:00:28 WARN conf.HiveConf: HiveConf of name hive.optimize.mapjoin.mapreduce does not exist
17/11/29 17:00:28 WARN conf.HiveConf: HiveConf of name hive.heapsize does not exist
17/11/29 17:00:28 WARN conf.HiveConf: HiveConf of name hive.server2.enable.impersonation does not exist
17/11/29 17:00:28 WARN conf.HiveConf: HiveConf of name hive.auto.convert.sortmerge.join.noconditionaltask does not exist

Logging initialized using configuration in jar:file:/usr/hdp/2.2.0.0-2041/hive/lib/hive-common-0.14.0.2.2.0.0-2041.jar!/hive-log4j.proper
ties
OK
Time taken: 24.967 seconds
Loading data to table default.employee
Table default.employee stats: [numFiles=1, totalSize=167]
OK
Time taken: 3.876 seconds
```

- **Now load data again into ‘employee’ table in MySQL, as shown below:**

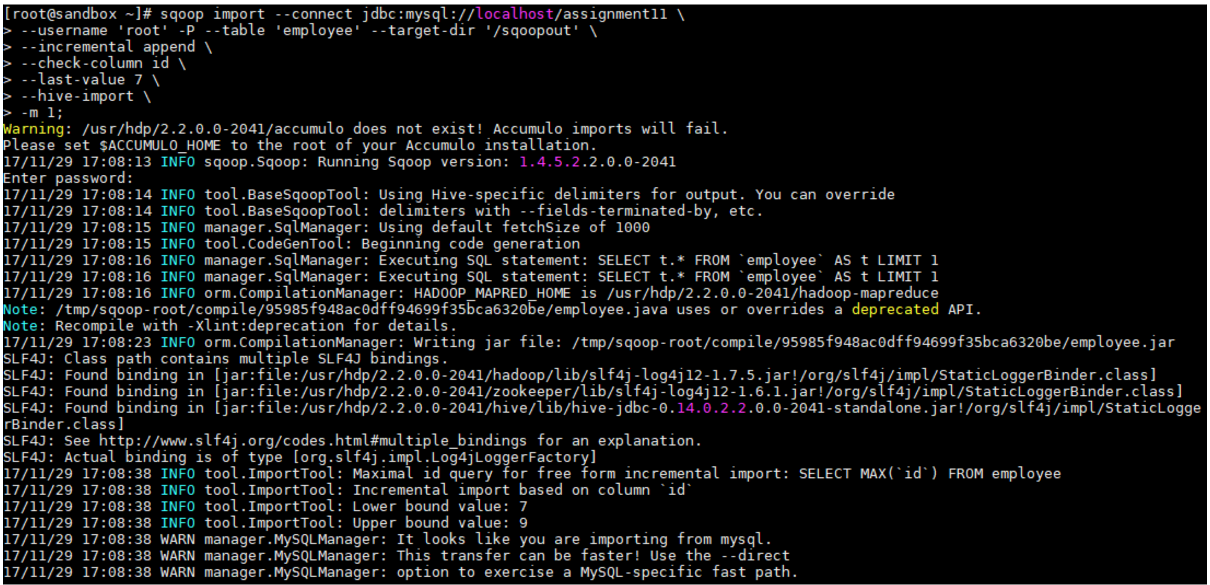
  insert into employee values(8, 'Suraj',28,'Team Lead',80000);
  insert into employee values(9, 'Ganesh',30,'Manager',100000);
  commit;



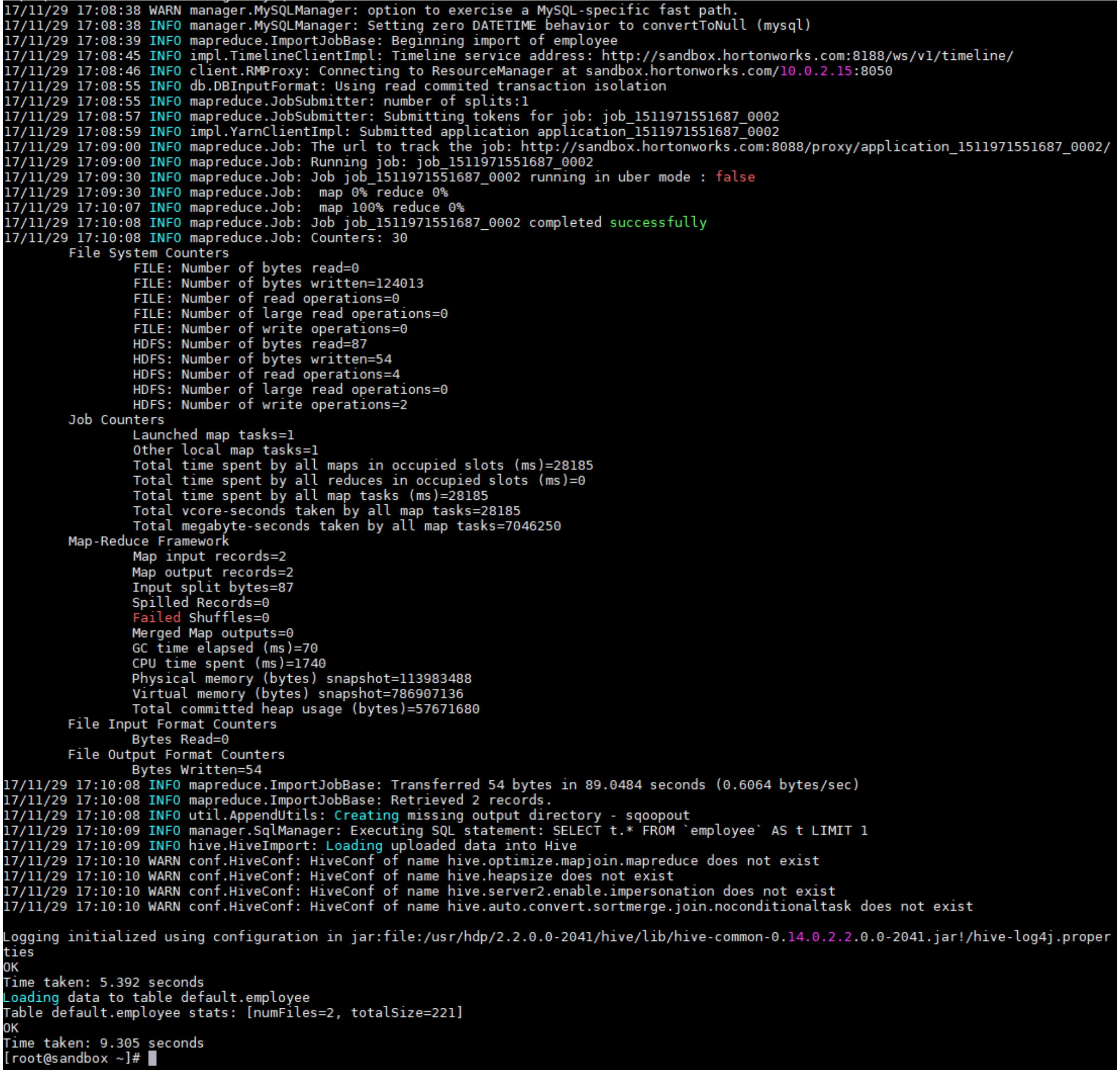
```
mysql> insert into employee values(8, 'Suraj',28,'Team Lead',80000);
Query OK, 1 row affected (0.00 sec)

mysql> insert into employee values(9, 'Ganesh',30,'Manager',100000);
Query OK, 1 row affected (0.00 sec)

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

mysql> select * from employee;
+------+---------+------+-------------------+--------+
| id   | name    | age  | skill             | salary |
+------+---------+------+-------------------+--------+
|    1 | Mohan   |   25 | Big Data & Hadoop |  30000 |
|    2 | Ramu    |   27 | AI                |  50000 |
|    3 | Ravi    |   30 | Java              |  60000 |
|    4 | Akshith |   22 | Automation        |  35000 |
|    5 | Shyam   |   35 | C                 |  40000 |
|    6 | Priya   |   28 | .Net              |  50000 |
|    7 | Madhu   |   27 | DBA               |  70000 |
|    8 | Suraj   |   28 | Team Lead         |  80000 |
|    9 | Ganesh  |   30 | Manager           | 100000 |
+------+---------+------+-------------------+--------+
9 rows in set (0.00 sec)

mysql> 
```

- **Now again run the Sqoop import query mentioning the last updated column value in the query as shown below:**

```
sqoop import --connect jdbc:mysql://localhost/assignment11 \
--username 'root' -P --table 'employee' --target-dir '/sqoopout' \
--incremental append \
--check-column id \
--last-value 7 \
--hive-import \
-m 1;
```

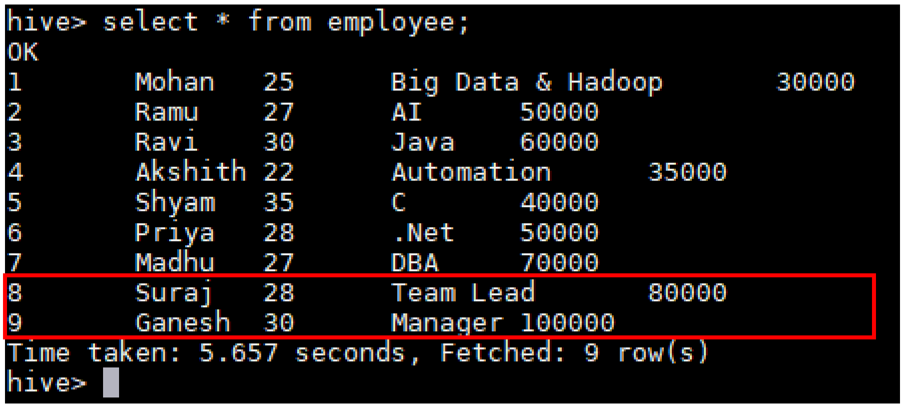


```
[root@sandbox ~]# sqoop import --connect jdbc:mysql://localhost/assignment11 \
> --username 'root' -P --table 'employee' --target-dir '/sqoopout' \
> --incremental append \
> --check-column id \
> --last-value 7 \
> --hive-import \
> -m 1;
Warning: /usr/hdp/2.2.0.0-2041/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/29 17:08:13 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5.2.2.0.0-2041
Enter password:
17/11/29 17:08:14 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
17/11/29 17:08:14 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
17/11/29 17:08:15 INFO manager.SqlManager: Using default fetchSize of 1000
17/11/29 17:08:15 INFO tool.CodeGenTool: Beginning code generation
17/11/29 17:08:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/29 17:08:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/29 17:08:16 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/hdp/2.2.0.0-2041/hadoop-mapreduce
Note: /tmp/sqoop-root/compile/95985f948ac0dff94699f35bca6320be/employee.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
17/11/29 17:08:23 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/95985f948ac0dff94699f35bca6320be/employee.jar
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hadoop/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/zookeeper/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.2.0.0-2041/hive/lib/hive-jdbc-0.14.0.2.2.0.0-2041-standalone.jar!/org/slf4j/impl/StaticLogge
rBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
17/11/29 17:08:38 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(`id`) FROM employee
17/11/29 17:08:38 INFO tool.ImportTool: Incremental import based on column `id`
17/11/29 17:08:38 INFO tool.ImportTool: Lower bound value: 7
17/11/29 17:08:38 INFO tool.ImportTool: Upper bound value: 9
17/11/29 17:08:38 WARN manager.MySQLManager: It looks like you are importing from mysql.
17/11/29 17:08:38 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
17/11/29 17:08:38 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
```



```
17/11/29 17:08:38 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
17/11/29 17:08:38 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
17/11/29 17:08:39 INFO mapreduce.ImportJobBase: Beginning import of employee
17/11/29 17:08:45 INFO impl.TimelineClientImpl: Timeline service address: http://sandbox.hortonworks.com:8188/ws/v1/timeline/
17/11/29 17:08:46 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
17/11/29 17:08:55 INFO db.DBInputFormat: Using read commited transaction isolation
17/11/29 17:08:55 INFO mapreduce.JobSubmitter: number of splits:1
17/11/29 17:08:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1511971551687_0002
17/11/29 17:08:59 INFO impl.YarnClientImpl: Submitted application application_1511971551687_0002
17/11/29 17:09:00 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1511971551687_0002/
17/11/29 17:09:00 INFO mapreduce.Job: Running job: job_1511971551687_0002
17/11/29 17:09:30 INFO mapreduce.Job: Job job_1511971551687_0002 running in uber mode : false
17/11/29 17:09:30 INFO mapreduce.Job:  map 0% reduce 0%
17/11/29 17:10:07 INFO mapreduce.Job:  map 100% reduce 0%
17/11/29 17:10:08 INFO mapreduce.Job: Job job_1511971551687_0002 completed successfully
17/11/29 17:10:08 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=124013
                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=54
                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)=28185
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=28185
                Total vcore-seconds taken by all map tasks=28185
                Total megabyte-seconds taken by all map tasks=7046250
        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)=70
                CPU time spent (ms)=1740
                Physical memory (bytes) snapshot=113983488
                Virtual memory (bytes) snapshot=786907136
                Total committed heap usage (bytes)=57671680
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=54
17/11/29 17:10:08 INFO mapreduce.ImportJobBase: Transferred 54 bytes in 89.0484 seconds (0.6064 bytes/sec)
17/11/29 17:10:08 INFO mapreduce.ImportJobBase: Retrieved 2 records.
17/11/29 17:10:08 INFO util.AppendUtils: Creating missing output directory - sqoopout
17/11/29 17:10:09 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `employee` AS t LIMIT 1
17/11/29 17:10:09 INFO hive.HiveImport: Loading uploaded data into Hive
17/11/29 17:10:10 WARN conf.HiveConf: HiveConf of name hive.optimize.mapjoin.mapreduce does not exist
17/11/29 17:10:10 WARN conf.HiveConf: HiveConf of name hive.heapsize does not exist
17/11/29 17:10:10 WARN conf.HiveConf: HiveConf of name hive.server2.enable.impersonation does not exist
17/11/29 17:10:10 WARN conf.HiveConf: HiveConf of name hive.auto.convert.sortmerge.join.noconditionaltask does not exist

Logging initialized using configuration in jar:file:/usr/hdp/2.2.0.0-2041/hive/lib/hive-common-0.14.0.2.2.0.0-2041.jar!/hive-log4j.proper
ties
OK
Time taken: 5.392 seconds
Loading data to table default.employee
Table default.employee stats: [numFiles=2, totalSize=221]
OK
Time taken: 9.305 seconds
[root@sandbox ~]#
```

- **Finally chek the 'employee' table in hive to verify whether new data from 'employee' table from MySQL loaded or not.**



```
hive> select * from employee;
OK
1       Mohan   25      Big Data & Hadoop       30000
2       Ramu    27      AI      50000
3       Ravi    30      Java    60000
4       Akshith 22      Automation      35000
5       Shyam   35      C       40000
6       Priya   28      .Net    50000
7       Madhu   27      DBA     70000
8       Suraj   28      Team Lead       80000
9       Ganesh  30      Manager 100000
Time taken: 5.657 seconds, Fetched: 9 row(s)
hive>
```