IIT- M Advanced Certificate Program in Machine Learning and Cloud- upGrad Capstone Project

User Demographics Prediction using Telecom dataset

Data Ingestion Commands

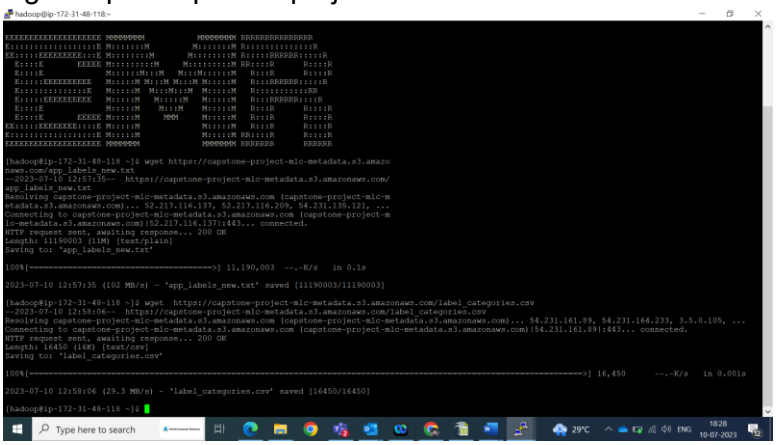
Authors :

Mukul Pahawa

Mitesh

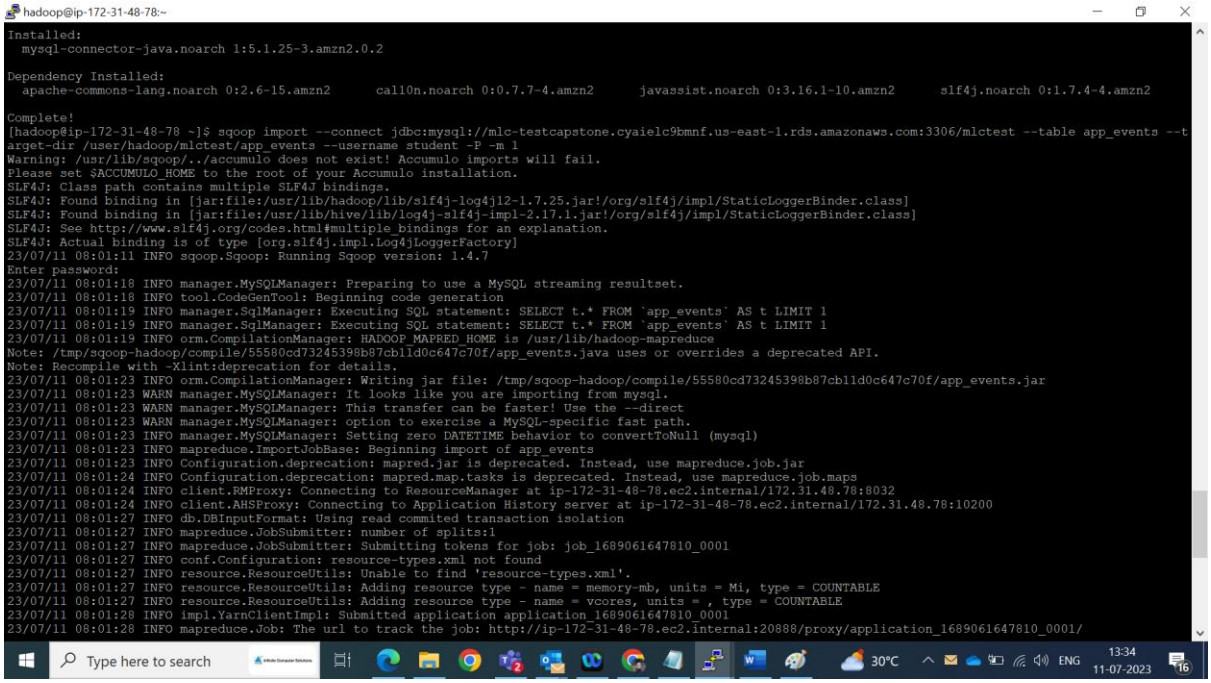
Fetching the source files containing Data from S3

wget https://capstone-project-mlc-metadata.s3.amazonaws.com/app_labels_new.txt
wget https://capstone-project-mlc-metadata.s3.amazonaws.com/label_categories.csv



The scoop command to load tables from mysql to HDFS

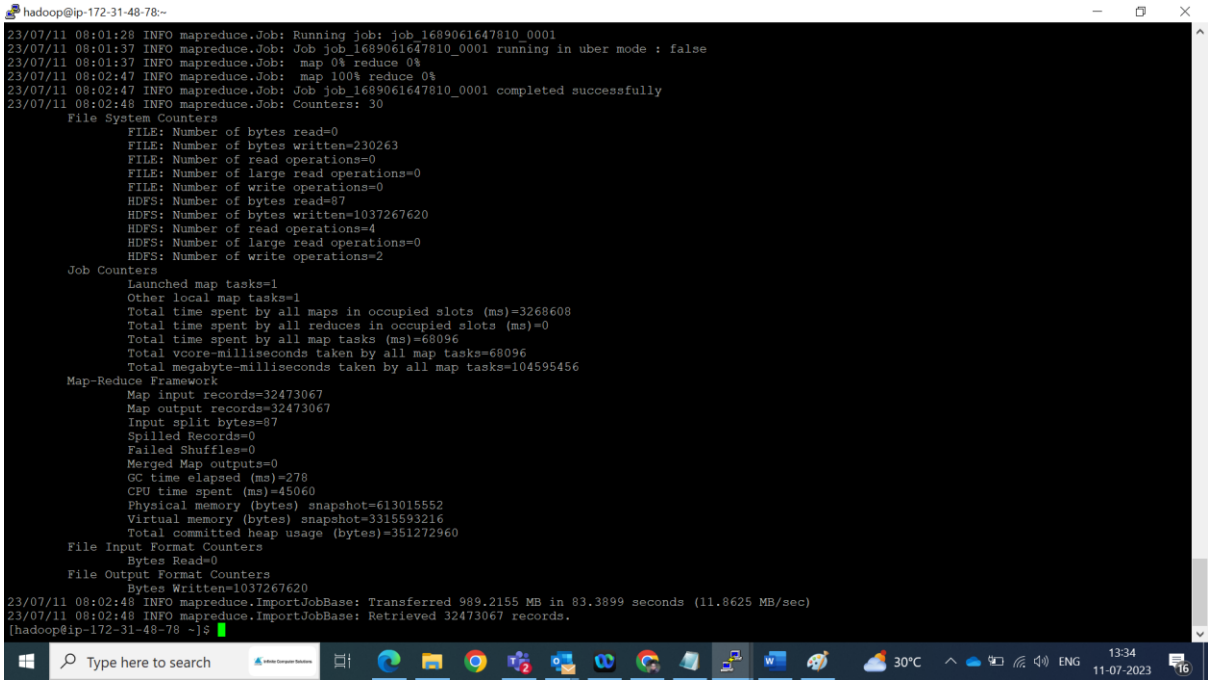
sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table app_events --target-dir /user/hadoop/mlctest/app_events --username student -P -m 1



```
hadoop@ip-172-31-48-78:~
Installed:
  mysql-connector-java.noarch 1:5.1.25-3.amzn2.0.2

Dependency Installed:
  apache-commons-lang.noarch 0:2.6-15.amzn2      cal10n.noarch 0:0.7.7-4.amzn2      javassist.noarch 0:3.16.1-10.amzn2      slf4j.noarch 0:1.7.4-4.amzn2

Complete!
[hadoop@ip-172-31-48-78 ~]$ sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table app_events --t
arget-dir /user/hadoop/mlctest/app_events --username student -P -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.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.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]
23/07/11 08:01:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
23/07/11 08:01:18 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/07/11 08:01:18 INFO tool.CodeGenTool: Beginning code generation
23/07/11 08:01:19 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `app_events` AS t LIMIT 1
23/07/11 08:01:19 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `app_events` AS t LIMIT 1
23/07/11 08:01:19 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/55580cd73245398b87cb11d0c647c70f/app_events.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/07/11 08:01:23 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/55580cd73245398b87cb11d0c647c70f/app_events.jar
23/07/11 08:01:23 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/07/11 08:01:23 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/07/11 08:01:23 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/07/11 08:01:23 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/07/11 08:01:23 INFO mapreduce.ImportJobBase: Beginning import of app_events
23/07/11 08:01:23 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/07/11 08:01:24 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/07/11 08:01:24 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-78.ec2.internal/172.31.48.78:8032
23/07/11 08:01:24 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-48-78.ec2.internal/172.31.48.78:10200
23/07/11 08:01:27 INFO db.DBInputFormat: Using read commited transaction isolation
23/07/11 08:01:27 INFO mapreduce.JobSubmitter: number of splits:1
23/07/11 08:01:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1689061647810_0001
23/07/11 08:01:27 INFO conf.Configuration: resource-types.xml not found
23/07/11 08:01:27 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/07/11 08:01:27 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
23/07/11 08:01:27 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
23/07/11 08:01:28 INFO impl.YarnClientImpl: Submitted application application_1689061647810_0001
23/07/11 08:01:28 INFO mapreduce.Job: The url to track the job: http://ip-172-31-48-78.ec2.internal:20888/proxy/application_1689061647810_0001/
```



```
hadoop@ip-172-31-48-78:~
23/07/11 08:01:28 INFO mapreduce.Job: Running job: job_1689061647810_0001
23/07/11 08:01:37 INFO mapreduce.Job: Job job_1689061647810_0001 running in uber mode : false
23/07/11 08:01:37 INFO mapreduce.Job:  map 0% reduce 0%
23/07/11 08:02:47 INFO mapreduce.Job:  map 100% reduce 0%
23/07/11 08:02:47 INFO mapreduce.Job: Job job_1689061647810_0001 completed successfully
23/07/11 08:02:48 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=230263
                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=1037267620
                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)=3268608
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=68096
                Total vcore-milliseconds taken by all map tasks=68096
                Total megabyte-milliseconds taken by all map tasks=104595456
        Map-Reduce Framework
                Map input records=32473067
                Map output records=32473067
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=278
                CPU time spent (ms)=45060
                Physical memory (bytes) snapshot=613015552
                Virtual memory (bytes) snapshot=3315593216
                Total committed heap usage (bytes)=351272960
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=1037267620
23/07/11 08:02:48 INFO mapreduce.ImportJobBase: Transferred 989.2155 MB in 83.3899 seconds (11.8625 MB/sec)
23/07/11 08:02:48 INFO mapreduce.ImportJobBase: Retrieved 32473067 records.
[hadoop@ip-172-31-48-78 ~]$
```

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table brand_device --target-dir /user/hadoop/mlctest/brand_device --username student -P -m 1

```
hadoop@ip-172-31-48-78:~
[hadoop@ip-172-31-48-78 ~]$ sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table brand_device -
-target-dir /user/hadoop/mlctest/brand_device --username student -P -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.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.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]
23/07/11 08:10:10 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
23/07/11 08:10:15 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/07/11 08:10:15 INFO tool.CodeGenTool: Beginning code generation
23/07/11 08:10:15 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `brand_device` AS t LIMIT 1
23/07/11 08:10:15 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `brand_device` AS t LIMIT 1
23/07/11 08:10:15 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/e6e06418a94293ca53119c2d7de5bbd4/brand_device.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/07/11 08:10:18 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/e6e06418a94293ca53119c2d7de5bbd4/brand_device.jar
23/07/11 08:10:18 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/07/11 08:10:18 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/07/11 08:10:18 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/07/11 08:10:18 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/07/11 08:10:18 INFO mapreduce.ImportJobBase: Beginning import of brand_device
23/07/11 08:10:18 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/07/11 08:10:19 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/07/11 08:10:19 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-78.ec2.internal/172.31.48.78:8032
23/07/11 08:10:19 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-48-78.ec2.internal/172.31.48.78:10200
23/07/11 08:10:24 INFO db.DBInputFormat: Using read commited transaction isolation
23/07/11 08:10:24 INFO mapreduce.JobSubmitter: number of splits:1
23/07/11 08:10:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1689061647810_0002
23/07/11 08:10:24 INFO conf.Configuration: resource-types.xml not found
23/07/11 08:10:24 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/07/11 08:10:24 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
23/07/11 08:10:24 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
23/07/11 08:10:24 INFO impl.YarnClientImpl: Submitted application application_1689061647810_0002
23/07/11 08:10:24 INFO mapreduce.Job: The url to track the job: http://ip-172-31-48-78.ec2.internal:20888/proxy/application_1689061647810_0002/
23/07/11 08:10:24 INFO mapreduce.Job: Running job: job_1689061647810_0002
23/07/11 08:10:32 INFO mapreduce.Job: Job job_1689061647810_0002 running in uber mode : false
23/07/11 08:10:32 INFO mapreduce.Job:  map 0% reduce 0%
23/07/11 08:10:40 INFO mapreduce.Job:  map 100% reduce 0%
23/07/11 08:10:40 INFO mapreduce.Job: Job job_1689061647810_0002 completed successfully
23/07/11 08:10:40 INFO mapreduce.Job: Counters: 30
```

```
hadoop@ip-172-31-48-78:~
23/07/11 08:10:24 INFO mapreduce.Job: Running job: job_1689061647810_0002
23/07/11 08:10:32 INFO mapreduce.Job: Job job_1689061647810_0002 running in uber mode : false
23/07/11 08:10:32 INFO mapreduce.Job:  map 0% reduce 0%
23/07/11 08:10:40 INFO mapreduce.Job:  map 100% reduce 0%
23/07/11 08:10:40 INFO mapreduce.Job: Job job_1689061647810_0002 completed successfully
23/07/11 08:10:40 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=230265
                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=6996440
                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)=203856
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4247
                Total vcore-milliseconds taken by all map tasks=4247
                Total megabyte-milliseconds taken by all map tasks=6523392
        Map-Reduce Framework
                Map input records=187245
                Map output records=187245
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=83
                CPU time spent (ms)=3360
                Physical memory (bytes) snapshot=336379904
                Virtual memory (bytes) snapshot=3316310016
                Total committed heap usage (bytes)=304087040
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=6996440
23/07/11 08:10:40 INFO mapreduce.ImportJobBase: Transferred 6.6723 MB in 20.5924 seconds (331.7951 KB/sec)
23/07/11 08:10:40 INFO mapreduce.ImportJobBase: Retrieved 187245 records.
[hadoop@ip-172-31-48-78 ~]$
```

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table events --target-dir /user/hadoop/mlctest/events --username student -P -m 1

```
hadoop@ip-172-31-48-78:~
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-48-78 ~]$ sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table events --targe
t-dir /user/hadoop/mlctest/events --username student -P -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.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.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]
23/07/11 08:27:54 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
23/07/11 08:27:59 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/07/11 08:27:59 INFO tool.CodeGenTool: Beginning code generation
23/07/11 08:28:00 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `events` AS t LIMIT 1
23/07/11 08:28:00 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `events` AS t LIMIT 1
23/07/11 08:28:00 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/96b5db6515e8b11ce1f73ec2fbfdf0b9/events.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/07/11 08:28:03 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/96b5db6515e8b11ce1f73ec2fbfdf0b9/events.jar
23/07/11 08:28:03 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/07/11 08:28:03 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/07/11 08:28:03 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/07/11 08:28:03 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/07/11 08:28:03 INFO mapreduce.ImportJobBase: Beginning import of events
23/07/11 08:28:03 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/07/11 08:28:04 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/07/11 08:28:04 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-78.ec2.internal/172.31.48.78:8032
23/07/11 08:28:04 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-48-78.ec2.internal/172.31.48.78:10200
23/07/11 08:28:05 INFO db.DBInputFormat: Using read commited transaction isolation
23/07/11 08:28:05 INFO mapreduce.JobSubmitter: number of splits:1
23/07/11 08:28:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1689061647810_0003
23/07/11 08:28:06 INFO conf.Configuration: resource-types.xml not found
23/07/11 08:28:06 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/07/11 08:28:06 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
23/07/11 08:28:06 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
23/07/11 08:28:06 INFO impl.YarnClientImpl: Submitted application application_1689061647810_0003
23/07/11 08:28:06 INFO mapreduce.Job: The url to track the job: http://ip-172-31-48-78.ec2.internal:20888/proxy/application_1689061647810_0003/
23/07/11 08:28:06 INFO mapreduce.Job: Running job: job_1689061647810_0003
```

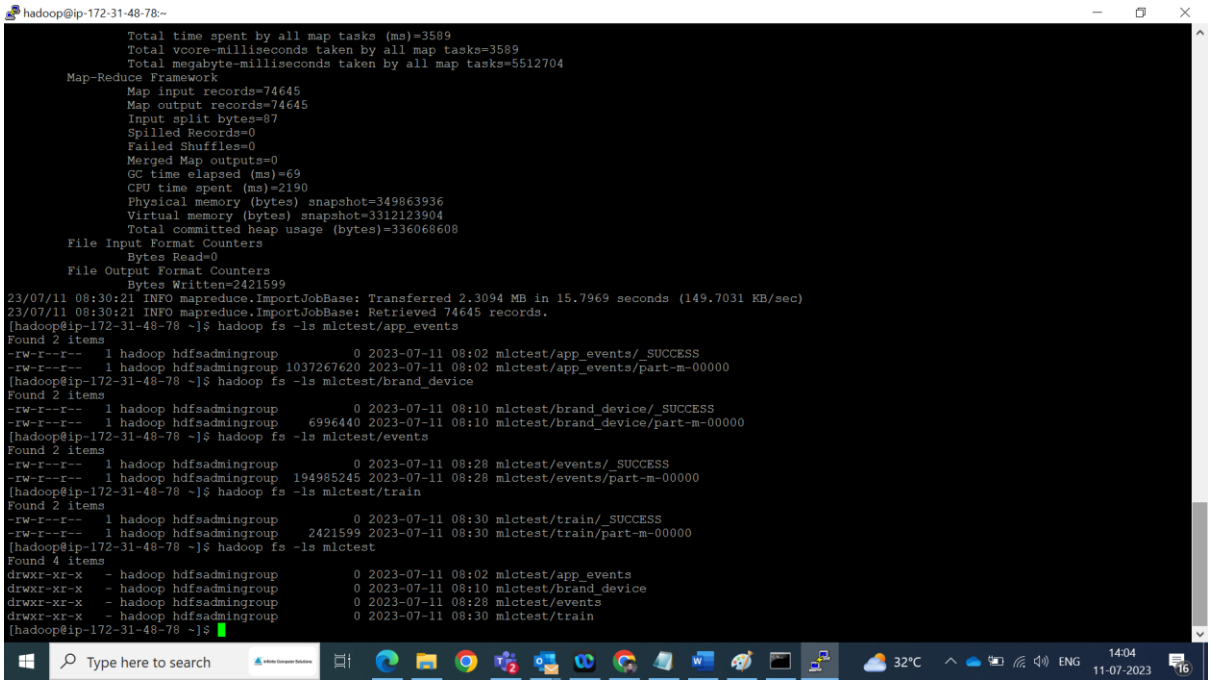
```
hadoop@ip-172-31-48-78:~
23/07/11 08:28:06 INFO mapreduce.Job: Running job: job_1689061647810_0003
23/07/11 08:28:13 INFO mapreduce.Job: Job job_1689061647810_0003 running in uber mode : false
23/07/11 08:28:13 INFO mapreduce.Job:  map 0% reduce 0%
23/07/11 08:28:30 INFO mapreduce.Job:  map 100% reduce 0%
23/07/11 08:28:30 INFO mapreduce.Job: Job job_1689061647810_0003 completed successfully
23/07/11 08:28:30 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=230258
                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=194985245
                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)=693744
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=14453
                Total vcore-milliseconds taken by all map tasks=14453
                Total megabyte-milliseconds taken by all map tasks=22199808
        Map-Reduce Framework
                Map input records=3252950
                Map output records=3252950
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=129
                CPU time spent (ms)=17260
                Physical memory (bytes) snapshot=632926208
                Virtual memory (bytes) snapshot=3322384384
                Total committed heap usage (bytes)=510132224
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=194985245
23/07/11 08:28:30 INFO mapreduce.ImportJobBase: Transferred 185.9524 MB in 25.8926 seconds (7.1817 MB/sec)
23/07/11 08:28:30 INFO mapreduce.ImportJobBase: Retrieved 3252950 records.
[hadoop@ip-172-31-48-78 ~]$
```

sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table train --target-dir /user/hadoop/mlctest/train --username student -P -m 1

```
hadoop@ip-172-31-48-78:~
[hadoop@ip-172-31-48-78 ~]$ sqoop import --connect jdbc:mysql://mlc-testcapstone.cyaielc9bmnf.us-east-1.rds.amazonaws.com:3306/mlctest --table train --target
-dir /user/hadoop/mlctest/train --username student -P -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.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.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]
23/07/11 08:29:55 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
23/07/11 08:30:01 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/07/11 08:30:01 INFO tool.CodeGenTool: Beginning code generation
23/07/11 08:30:01 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `train` AS t LIMIT 1
23/07/11 08:30:01 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `train` AS t LIMIT 1
23/07/11 08:30:01 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/b8d655dfc9f5458adbf54a648444fad8/train.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/07/11 08:30:04 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/b8d655dfc9f5458adbf54a648444fad8/train.jar
23/07/11 08:30:04 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/07/11 08:30:04 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/07/11 08:30:04 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/07/11 08:30:04 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/07/11 08:30:04 INFO mapreduce.ImportJobBase: Beginning import of train
23/07/11 08:30:04 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
23/07/11 08:30:05 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
23/07/11 08:30:05 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-78.ec2.internal/172.31.48.78:8032
23/07/11 08:30:05 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-48-78.ec2.internal/172.31.48.78:10200
23/07/11 08:30:06 INFO db.DBInputFormat: Using read commited transaction isolation
23/07/11 08:30:07 INFO mapreduce.JobSubmitter: number of splits:1
23/07/11 08:30:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1689061647810_0004
23/07/11 08:30:07 INFO conf.Configuration: resource-types.xml not found
23/07/11 08:30:07 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/07/11 08:30:07 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
23/07/11 08:30:07 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
23/07/11 08:30:07 INFO impl.YarnClientImpl: Submitted application application_1689061647810_0004
23/07/11 08:30:08 INFO mapreduce.Job: The url to track the job: http://ip-172-31-48-78.ec2.internal:20888/proxy/application_1689061647810_0004/
23/07/11 08:30:08 INFO mapreduce.Job: Running job: job_1689061647810_0004
23/07/11 08:30:15 INFO mapreduce.Job: Job job_1689061647810_0004 running in uber mode : false
23/07/11 08:30:15 INFO mapreduce.Job:  map 0% reduce 0%
23/07/11 08:30:21 INFO mapreduce.Job:  map 100% reduce 0%
23/07/11 08:30:21 INFO mapreduce.Job: Job job_1689061647810_0004 completed successfully
23/07/11 08:30:21 INFO mapreduce.Job: Counters: 30
        File System Counters
```

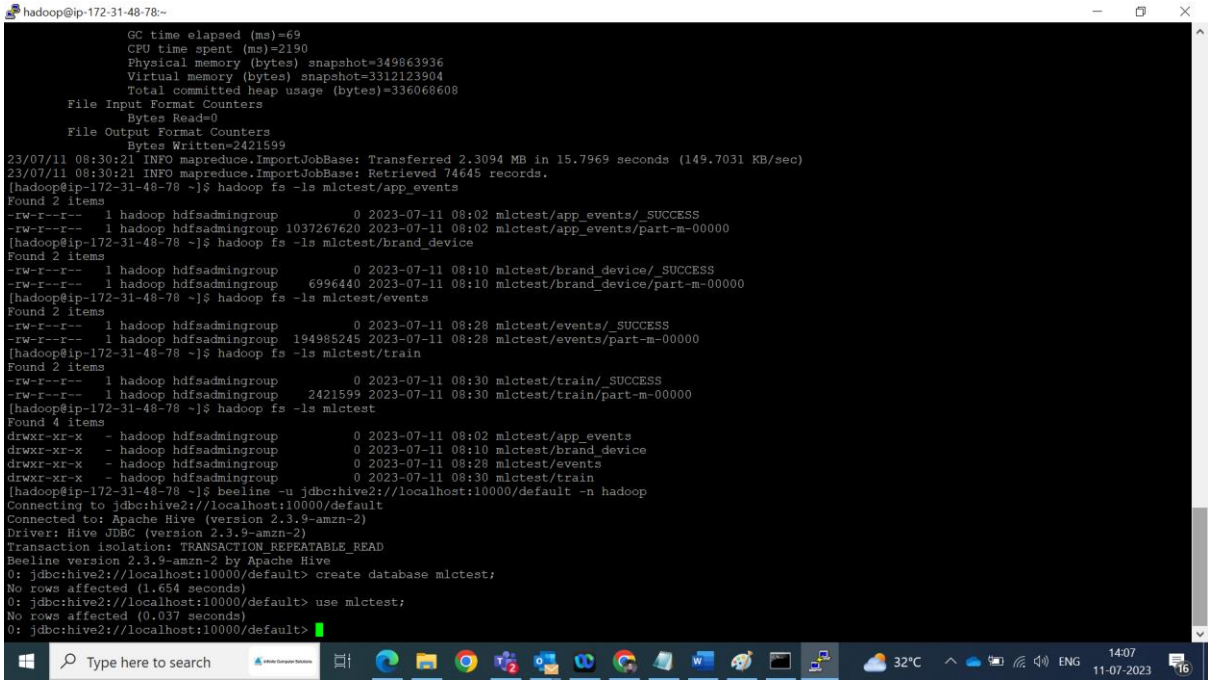
```
hadoop@ip-172-31-48-78:~
23/07/11 08:30:08 INFO mapreduce.Job: Running job: job_1689061647810_0004
23/07/11 08:30:15 INFO mapreduce.Job: Job job_1689061647810_0004 running in uber mode : false
23/07/11 08:30:15 INFO mapreduce.Job:  map 0% reduce 0%
23/07/11 08:30:21 INFO mapreduce.Job:  map 100% reduce 0%
23/07/11 08:30:21 INFO mapreduce.Job: Job job_1689061647810_0004 completed successfully
23/07/11 08:30:21 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=230237
                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=2421599
                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)=172272
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3589
                Total vcore-milliseconds taken by all map tasks=3589
                Total megabyte-milliseconds taken by all map tasks=5512704
        Map-Reduce Framework
                Map input records=74645
                Map output records=74645
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=69
                CPU time spent (ms)=2190
                Physical memory (bytes) snapshot=349863936
                Virtual memory (bytes) snapshot=3312123904
                Total committed heap usage (bytes)=336068608
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=2421599
23/07/11 08:30:21 INFO mapreduce.ImportJobBase: Transferred 2.3094 MB in 15.7969 seconds (149.7031 KB/sec)
23/07/11 08:30:21 INFO mapreduce.ImportJobBase: Retrieved 74645 records.
[hadoop@ip-172-31-48-78 ~]$
```

Listing hadoop filesystem
==========
hadoop fs -ls mlctest/app_events
hadoop fs -ls mlctest/brand_device
hadoop fs -ls mlctest/events
hadoop fs -ls mlctest/train
hadoop fs -ls mlctest



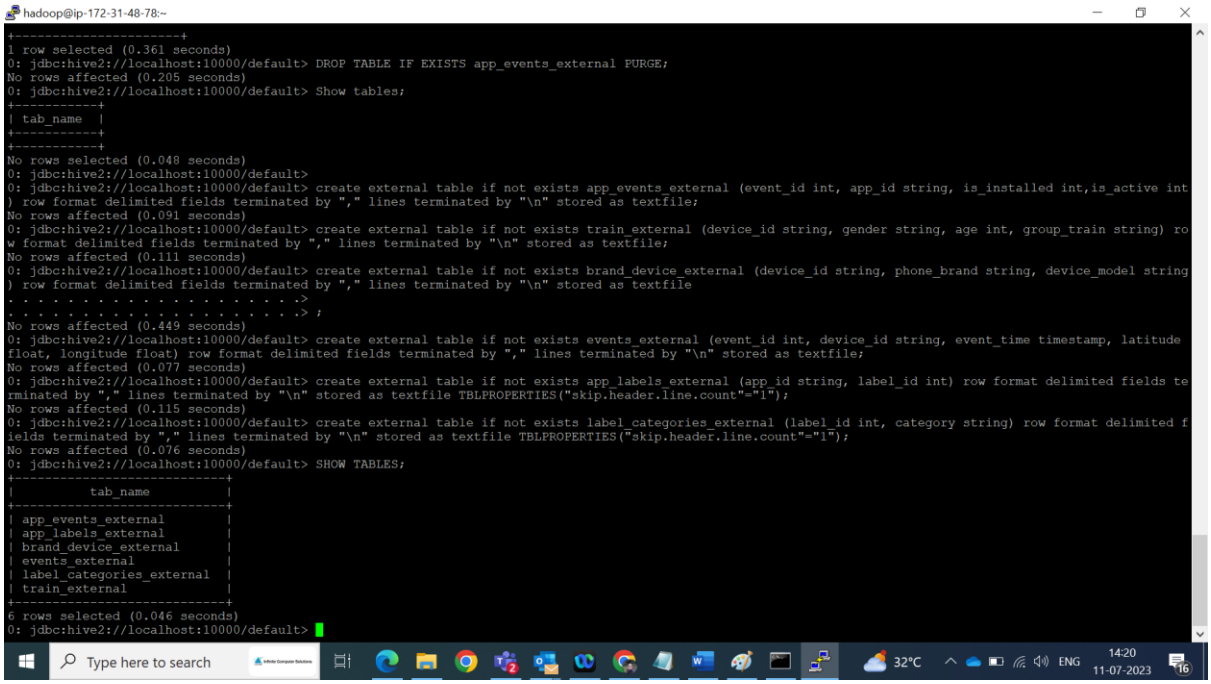
Connecting to HIVE, creating database and using the same to create HIVE tables

beeline -u jdbc:hive2://localhost:10000/default -n hadoop
create database mlctest;
use mlctest;



```
hadoop@ip-172-31-48-78:~
                GC time elapsed (ms)=69
                CPU time spent (ms)=2190
                Physical memory (bytes) snapshot=349863936
                Virtual memory (bytes) snapshot=3312123904
                Total committed heap usage (bytes)=336068608
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=2421599
23/07/11 08:30:21 INFO mapreduce.ImportJobBase: Transferred 2.3094 MB in 15.7969 seconds (149.7031 KB/sec)
23/07/11 08:30:21 INFO mapreduce.ImportJobBase: Retrieved 74645 records.
[hadoop@ip-172-31-48-78 ~]$ hadoop fs -ls mlctest/app_events
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-07-11 08:02 mlctest/app_events/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup 1037267620 2023-07-11 08:02 mlctest/app_events/part-m-00000
[hadoop@ip-172-31-48-78 ~]$ hadoop fs -ls mlctest/brand_device
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-07-11 08:10 mlctest/brand_device/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup    6996440 2023-07-11 08:10 mlctest/brand_device/part-m-00000
[hadoop@ip-172-31-48-78 ~]$ hadoop fs -ls mlctest/events
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-07-11 08:28 mlctest/events/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup  194985245 2023-07-11 08:28 mlctest/events/part-m-00000
[hadoop@ip-172-31-48-78 ~]$ hadoop fs -ls mlctest/train
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-07-11 08:30 mlctest/train/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup    2421599 2023-07-11 08:30 mlctest/train/part-m-00000
[hadoop@ip-172-31-48-78 ~]$ hadoop fs -ls mlctest
Found 4 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-07-11 08:02 mlctest/app_events
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-07-11 08:10 mlctest/brand_device
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-07-11 08:28 mlctest/events
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-07-11 08:30 mlctest/train
[hadoop@ip-172-31-48-78 ~]$ beeline -u jdbc:hive2://localhost:10000/default -n hadoop
Connecting to jdbc:hive2://localhost:10000/default
Connected to: Apache Hive (version 2.3.9-amzn-2)
Driver: Hive JDBC (version 2.3.9-amzn-2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.9-amzn-2 by Apache Hive
0: jdbc:hive2://localhost:10000/default> create database mlctest;
No rows affected (1.654 seconds)
0: jdbc:hive2://localhost:10000/default> use mlctest;
No rows affected (0.037 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Creation of HIVE external Tables
==========
create external table if not exists app_events_external (event_id int, app_id string, is_installed int,is_active int) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
create external table if not exists train_external (device_id string, gender string, age int, group_train string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
create external table if not exists brand_device_external (device_id string, phone_brand string, device_model string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
create external table if not exists events_external (event_id int, device_id string, event_time timestamp, latitude float, longitude float) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
create external table if not exists app_labels_external (app_id string, label_id int) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");
create external table if not exists label_categories_external (label_id int, category string) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");



```
hadoop@ip-172-31-48-78:~
+----------------------+
1 row selected (0.361 seconds)
0: jdbc:hive2://localhost:10000/default> DROP TABLE IF EXISTS app_events_external PURGE;
No rows affected (0.205 seconds)
0: jdbc:hive2://localhost:10000/default> Show tables;
+-----------+
| tab_name  |
+-----------+
+-----------+
No rows selected (0.048 seconds)
0: jdbc:hive2://localhost:10000/default>
0: jdbc:hive2://localhost:10000/default> create external table if not exists app_events_external (event_id int, app_id string, is_installed int,is_active int
) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
No rows affected (0.091 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists train_external (device_id string, gender string, age int, group_train string) ro
w format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
No rows affected (0.111 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists brand_device_external (device_id string, phone_brand string, device_model string
) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile
. . . . . .>
. . . . . .> ;
No rows affected (0.449 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists events_external (event_id int, device_id string, event_time timestamp, latitude
float, longitude float) row format delimited fields terminated by "," lines terminated by "\n" stored as textfile;
No rows affected (0.077 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists app_labels_external (app_id string, label_id int) row format delimited fields te
rminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");
No rows affected (0.115 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists label_categories_external (label_id int, category string) row format delimited f
ields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");
No rows affected (0.076 seconds)
0: jdbc:hive2://localhost:10000/default> SHOW TABLES;
+----------------------------+
|          tab_name          |
+----------------------------+
| app_events_external        |
| app_labels_external        |
| brand_device_external      |
| events_external            |
| label_categories_external  |
| train_external             |
+----------------------------+
6 rows selected (0.046 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Load into Hive tables from HDFS and validate Data in the tables
==========
load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external;
select * from app_events_external limit 5;
load data inpath '/user/hadoop/mlctest/brand_device' into table brand_device_external;
select * from brand_device_external limit 5;
load data inpath '/user/hadoop/mlctest/events' into table events_external;
select * from events_external limit 5;
load data inpath '/user/hadoop/mlctest/train' into table train_external;
select * from train_external limit 5

```
hadoop@ip-172-31-48-78:~
No rows affected (0.077 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists app_labels_external (app_id string, label_id int) row format delimited fields te
rminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");
No rows affected (0.115 seconds)
0: jdbc:hive2://localhost:10000/default> create external table if not exists label_categories_external (label_id int, category string) row format delimited f
ields terminated by "," lines terminated by "\n" stored as textfile TBLPROPERTIES("skip.header.line.count"="1");
No rows affected (0.076 seconds)
0: jdbc:hive2://localhost:10000/default> SHOW TABLES;
+----------------------------+
|          tab_name          |
+----------------------------+
| app_events_external        |
| app_labels_external        |
| brand_device_external      |
| events_external            |
| label_categories_external  |
| train_external             |
+----------------------------+
6 rows selected (0.046 seconds)
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/app_events' into table app_events_external;
No rows affected (0.968 seconds)
0: jdbc:hive2://localhost:10000/default> select * from app_events_external limit 5;
+-------------------------------+-------------------------------+-----------------------------------+-------------------------------+
| app_events_external.event_id  | app_events_external.app_id    | app_events_external.is_installed  | app_events_external.is_active |
+-------------------------------+-------------------------------+-----------------------------------+-------------------------------+
| 2                             | 5927333115845830913           | 1                                 | 1                             |
| 2                             | -5720078949152207372          | 1                                 | 0                             |
| 2                             | -1633887856876571208          | 1                                 | 0                             |
| 2                             | -653184325010919369           | 1                                 | 1                             |
| 2                             | 8693964245073640147           | 1                                 | 1                             |
+-------------------------------+-------------------------------+-----------------------------------+-------------------------------+
5 rows selected (2.224 seconds)
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/brand_device' into table brand_device_external;
No rows affected (0.488 seconds)
0: jdbc:hive2://localhost:10000/default> t/events' into table events_external;
Error: Error while compiling statement: FAILED: ParseException line 1:0 cannot recognize input near 't' '/' 'events' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/default> t/events' into table events_external;
Error: Error while compiling statement: FAILED: ParseException line 1:0 cannot recognize input near 't' '/' 'events' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/default> select * from events_external limit 5;
+-------------------------+---------------------------+---------------------------+---------------------------+---------------------------+
| events_external.event_id | events_external.device_id | events_external.event_time | events_external.latitude | events_external.longitude |
+-------------------------+---------------------------+---------------------------+---------------------------+---------------------------+
+-------------------------+---------------------------+---------------------------+---------------------------+---------------------------+
No rows selected (0.233 seconds)
```

```
hadoop@ip-172-31-48-78:~
+-------------------------+---------------------------+---------------------------+---------------------------+---------------------------+
No rows selected (0.233 seconds)
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/brand_device' into table brand_device_external;
Error: Error while compiling statement: FAILED: SemanticException Line 1:17 Invalid path ''/user/hadoop/mlctest/brand_device'': No files matching path hdfs:/
/ip-172-31-48-78.ec2.internal:8020/user/hadoop/mlctest/brand_device (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/default> select * from brand_device_external limit 5;
+-----------------------------+-----------------------------------+-----------------------------------+
| brand_device_external.device_id | brand_device_external.phone_brand | brand_device_external.device_model |
+-----------------------------+-----------------------------------+-----------------------------------+
| 1845358998536310000         | meitu                             | 2                                 |
| 3126957642374570000         | meitu                             | 2                                 |
| -3051457881268070000        | meitu                             | 2                                 |
| 4608241502940040000         | meitu                             | 2                                 |
| 6005031767544890000         | meitu                             | 2                                 |
+-----------------------------+-----------------------------------+-----------------------------------+
5 rows selected (0.173 seconds)
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/events' into table events_external;
No rows affected (0.352 seconds)
0: jdbc:hive2://localhost:10000/default> select * from events_external limit 5;
+-------------------------+---------------------------+---------------------------+---------------------------+---------------------------+
| events_external.event_id | events_external.device_id | events_external.event_time | events_external.latitude | events_external.longitude |
+-------------------------+---------------------------+---------------------------+---------------------------+---------------------------+
| 1                       | 29182687948017100         | 2016-05-01 00:55:25.0     | 121.38                    | 31.24                     |
| 2                       | -6401643145415150000      | 2016-05-01 00:54:12.0     | 103.65                    | 30.97                     |
| 3                       | -4833982096941400000      | 2016-05-01 00:08:05.0     | 106.6                     | 29.7                      |
| 4                       | -6815121365017310000      | 2016-05-01 00:06:40.0     | 104.27                    | 23.28                     |
| 5                       | -5373797595892510000      | 2016-05-01 00:07:18.0     | 115.88                    | 28.66                     |
+-------------------------+---------------------------+---------------------------+---------------------------+---------------------------+
5 rows selected (0.188 seconds)
0: jdbc:hive2://localhost:10000/default> load data inpath '/user/hadoop/mlctest/train' into table train_external;
No rows affected (0.495 seconds)
0: jdbc:hive2://localhost:10000/default> select * from train_external 5;
Error: Error while compiling statement: FAILED: ParseException line 1:29 cannot recognize input near 'train_external' '5' '<EOF>' in joinSource (state=42000,
code=40000)
0: jdbc:hive2://localhost:10000/default> select * from train_external limit 5;
+-------------------------+-----------------------+-------------------+---------------------------+
| train_external.device_id | train_external.gender | train_external.age | train_external.group_train |
+-------------------------+-----------------------+-------------------+---------------------------+
| -7548291590301750000    | M                     | 33                | M32+                      |
| 6943568600617760000     | M                     | 37                | M32+                      |
| 5441349705980020000     | M                     | 40                | M32+                      |
| -5393876656119450000    | M                     | 33                | M32+                      |
| 4543988487649880000     | M                     | 53                | M32+                      |
+-------------------------+-----------------------+-------------------+---------------------------+
```

Load Data into Hive tables from local files
==========
load data local inpath '/home/hadoop/app_labels_new.txt' into table app_labels_external;
select * from app_lables_external limit 5;

load data local inpath '/home/hadoop/label_categories.csv' into table label_categories_external;
select * from label_categories_external limit 5;

```
hadoop@ip-172-31-48-78:~
code=40000)
0: jdbc:hive2://localhost:10000/default> select * from train_external limit 5;
+-------------------------+-----------------------+--------------------+---------------------------+
| train_external.device_id | train_external.gender | train_external.age | train_external.group_train |
+-------------------------+-----------------------+--------------------+---------------------------+
| -7548291590301750000    | M                     | 33                 | M32+                      |
| 6943568600617760000     | M                     | 37                 | M32+                      |
| 5441349705980020000     | M                     | 40                 | M32+                      |
| -5393876656119450000    | M                     | 33                 | M32+                      |
| 4543988487649880000     | M                     | 53                 | M32+                      |
+-------------------------+-----------------------+--------------------+---------------------------+
5 rows selected (0.176 seconds)
0: jdbc:hive2://localhost:10000/default> load data local inpath '/home/hadoop/app_labels_new.txt' into table app_labels_external;
No rows affected (0.378 seconds)
0: jdbc:hive2://localhost:10000/default> select * from app_lables_external limit 5;
Error: Error while compiling statement: FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'app_lables_external' (state=42S02,code=10001)
0: jdbc:hive2://localhost:10000/default> select * from app_lables_external limit 5;
Error: Error while compiling statement: FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'app_lables_external' (state=42S02,code=10001)
0: jdbc:hive2://localhost:10000/default> select * from app_labels_external limit 5;
+---------------------------+-----------------------------+
| app_labels_external.app_id | app_labels_external.label_id |
+---------------------------+-----------------------------+
| 7324884708820027918       | 251                         |
| -4494216993218550286      | 251                         |
| 6058196446775239644       | 406                         |
| 6058196446775239644       | 407                         |
| 8694625920731541625       | 406                         |
+---------------------------+-----------------------------+
5 rows selected (0.149 seconds)
0: jdbc:hive2://localhost:10000/default> load data local inpath '/home/hadoop/label_categories.csv' into table label_categories_external;
No rows affected (0.298 seconds)
0: jdbc:hive2://localhost:10000/default> Display all 560 possibilities? (y or n)
0: jdbc:hive2://localhost:10000/default> select * from label_categories_external limit 5;
+-------------------------------------+-----------------------------------+
| label_categories_external.label_id  | label_categories_external.category |
+-------------------------------------+-----------------------------------+
| 1                                   |                                   |
| 2                                   | game-game type                    |
| 3                                   | game-Game themes                  |
| 4                                   | game-Art Style                    |
| 5                                   | game-Leisure time                 |
+-------------------------------------+-----------------------------------+
5 rows selected (0.188 seconds)
0: jdbc:hive2://localhost:10000/default>
```