TRENDYTECH
UPLIFT YOUR CAREER!

Apache Sqoop

By Sumit Mittal

Apache Sqoop Exercise 1

TRENDY TECH

UPLIFT YOUR CAREER!

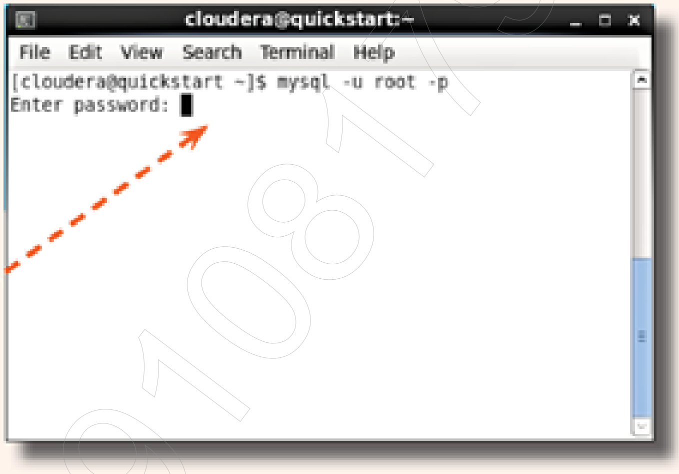
Sqoop Basics

To enter into MySQL:

```
mysql -u root -p
```

Note:

Enter password: cloudera



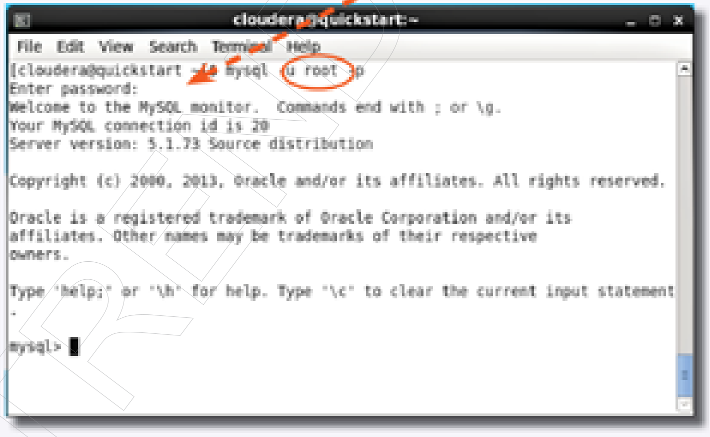
```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password: █
```

MySQL root user:

root user has acess to all the databases.

```
mysql -u root -p
```

(Enter password: cloudera)



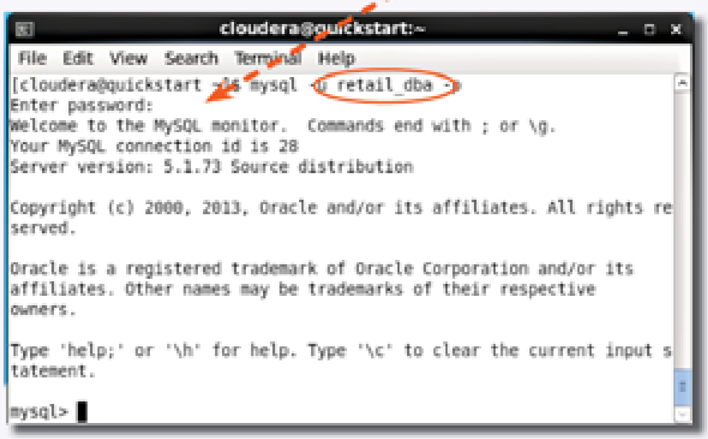
```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 20
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement
.

mysql> █
```

MySQL reatil_dba user:

retail_dba user has acess to limited databases.

```
mysql -u retail_dba -p
```

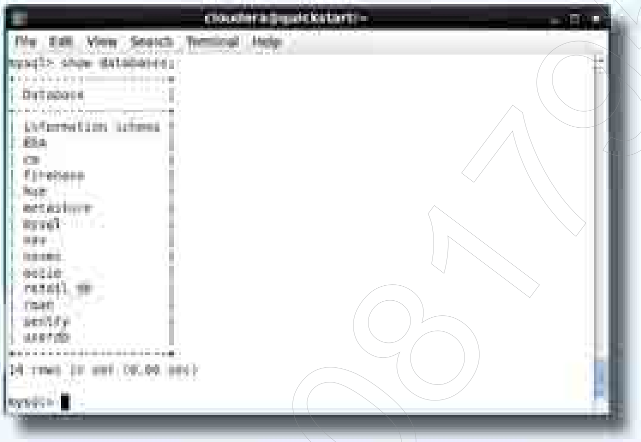
(Enter password: cloudera)



```
[cloudera@quickstart ~]$ mysql -u retail_dba -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 28
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights re
served.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input s
tatement.

mysql> █
```

To display databases in MySQL:

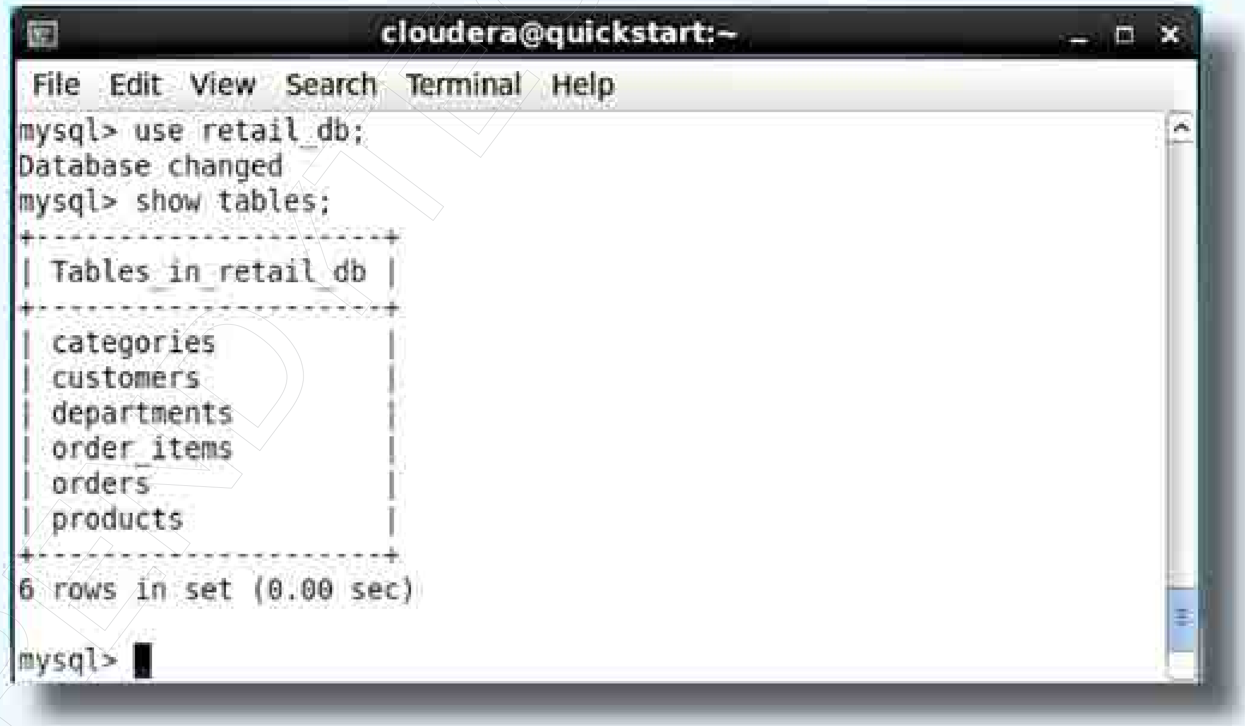
```
show databases;
```



```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| amon               |
| cm                 |
| firehose           |
| hue                |
| metastore          |
| mysql              |
| nav                |
| navms              |
| oozie              |
| retail_db          |
| rman               |
| sentry             |
| userdb             |
+--------------------+
14 rows in set (0.00 sec)

mysql> 
```

Use databases and display tables:

```
use retail_db;


show tables;
```

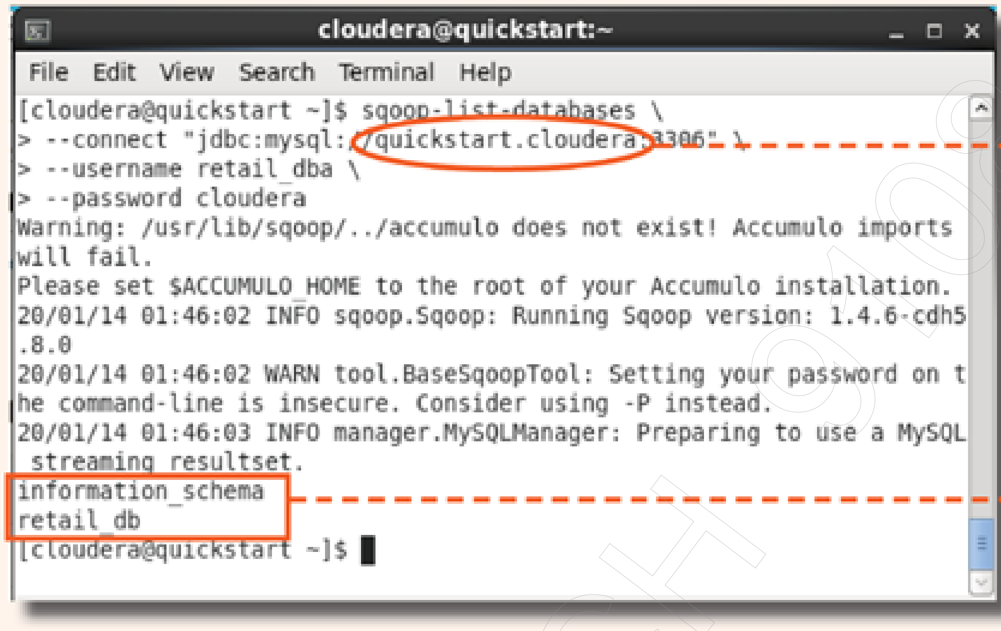


```
mysql> use retail_db;
Database changed
mysql> show tables;
+---------------------+
| Tables_in_retail_db |
+---------------------+
| categories          |
| customers           |
| departments         |
| order_items         |
| orders              |
| products            |
+---------------------+
6 rows in set (0.00 sec)

mysql> 
```

Acessenig MySQL databses from Hadoop using Sqoop:

```
sqoop-list-databases \
--connect "jdbc:mysql://quickstart.cloudera:3306" \
--username retail_dba \
--password cloudera
```

Space with backslash (\) Indicates continuation of line



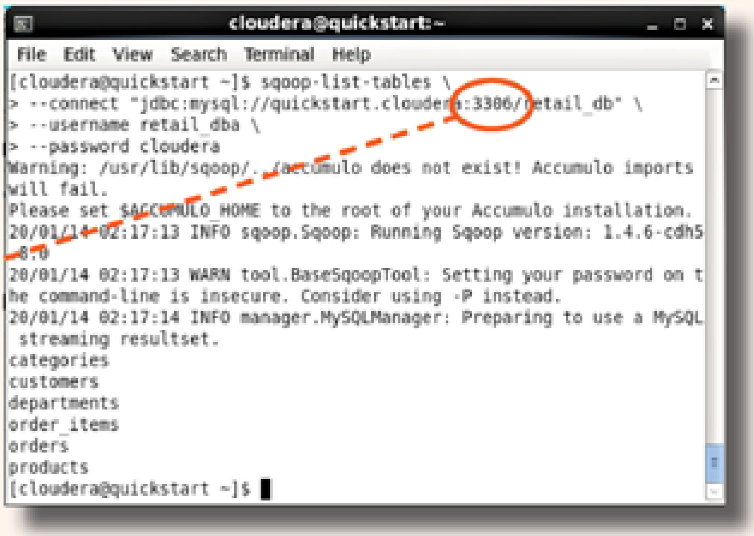
```
[cloudera@quickstart ~]$ sqoop-list-databases \
> --connect "jdbc:mysql://quickstart.cloudera:3306" \
> --username retail_dba \
> --password cloudera
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports
will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/14 01:46:02 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5
.8.0
20/01/14 01:46:02 WARN tool.BaseSqoopTool: Setting your password on t
he command-line is insecure. Consider using -P instead.
20/01/14 01:46:03 INFO manager.MySQLManager: Preparing to use a MySQL
 streaming resultset.
information_schema
retail_db
[cloudera@quickstart ~]$
```

Local host name

List of databese

Acessenig MySQL tables using the *root* user:

```
sqoop-list-tables \
--connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" \
--username retail_dba \
--password cloudera
```

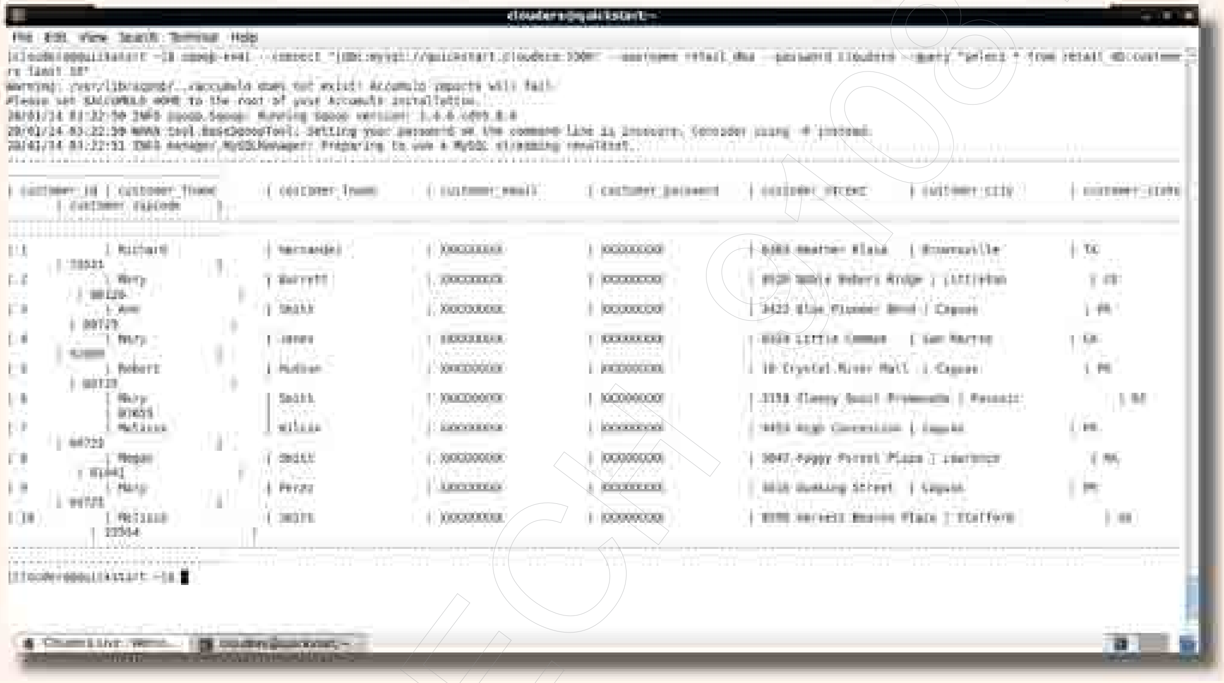


```
[cloudera@quickstart ~]$ sqoop-list-tables \
> --connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" \
> --username retail_dba \
> --password cloudera
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports
will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/14 02:17:13 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5
.8.0
20/01/14 02:17:13 WARN tool.BaseSqoopTool: Setting your password on t
he command-line is insecure. Consider using -P instead.
20/01/14 02:17:14 INFO manager.MySQLManager: Preparing to use a MySQL
 streaming resultset.
categories
customers
departments
order_items
orders
products
[cloudera@quickstart ~]$
```

Local port no. where MySQL runs

Displaying table data using sqoop-eval:

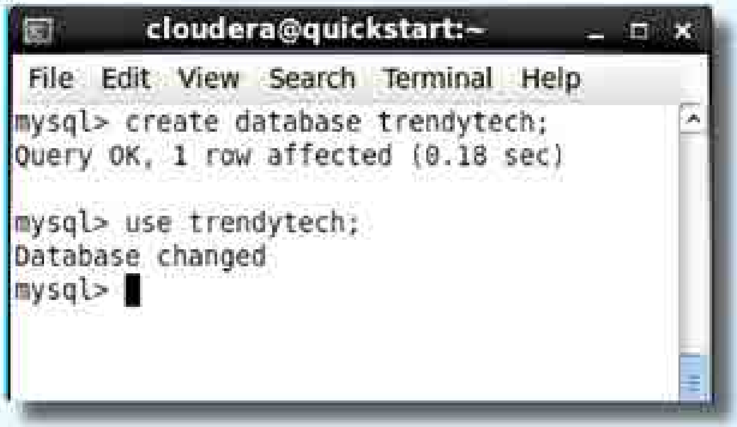
```
sqoop-eval \
--connect "jdbc:mysql://quickstart.cloudera:3306" \
--username retail_dba \
--password cloudera \
--query "select * from retail_db.customers limit 10"
```



```
[cloudera@quickstart ~]$ sqoop-eval --connect "jdbc:mysql://quickstart.cloudera:3306" --username retail_dba --password cloudera --query "select * from retail_db.custome
rs limit 10"
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/14 03:22:30 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
20/01/14 03:22:30 WARN tool.BaseSqoopTool: Setting your password on the command line is insecure. Consider using -P instead.
20/01/14 03:22:31 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
-----------------------------------------------------------------------------------------------------------------------------------------------
| customer_id | customer_fname       | customer_lname       | customer_email       | customer_password    | customer_street      | customer_city        | customer_state
      | customer_zipcode     |
-----------------------------------------------------------------------------------------------------------------------------------------------
| 1           | Richard              | Hernandez            | XXXXXXXXX            | XXXXXXXXX            | 6303 Heather Plaza   | Brownsville          | TX
      | 78521                |
| 2           | Mary                 | Barrett              | XXXXXXXXX            | XXXXXXXXX            | 9526 Noble Embers Ridge | Littleton         | CO
      | 80126                |
| 3           | Ann                  | Smith                | XXXXXXXXX            | XXXXXXXXX            | 3422 Blue Pioneer Bend | Caguas             | PR
      | 00725                |
| 4           | Mary                 | Jones                | XXXXXXXXX            | XXXXXXXXX            | 8324 Little Common   | San Marcos           | CA
      | 92069                |
| 5           | Robert               | Hudson               | XXXXXXXXX            | XXXXXXXXX            | 10 Crystal River Mall | Caguas              | PR
      | 00725                |
| 6           | Mary                 | Smith                | XXXXXXXXX            | XXXXXXXXX            | 3151 Sleepy Quail Promenade | Passaic       | NJ
      | 07055                |
| 7           | Melissa              | Wilcox               | XXXXXXXXX            | XXXXXXXXX            | 9453 High Concession | Caguas               | PR
      | 00725                |
| 8           | Megan                | Smith                | XXXXXXXXX            | XXXXXXXXX            | 3047 Foggy Forest Plaza | Lawrence          | MA
      | 01841                |
| 9           | Mary                 | Perez                | XXXXXXXXX            | XXXXXXXXX            | 3616 Quaking Street  | Caguas               | PR
      | 00725                |
| 10          | Melissa              | Smith                | XXXXXXXXX            | XXXXXXXXX            | 8598 Harvest Beacon Plaza | Stafford        | VA
      | 22554                |
-----------------------------------------------------------------------------------------------------------------------------------------------
[cloudera@quickstart ~]$
```

Create and use a database in MySQL:

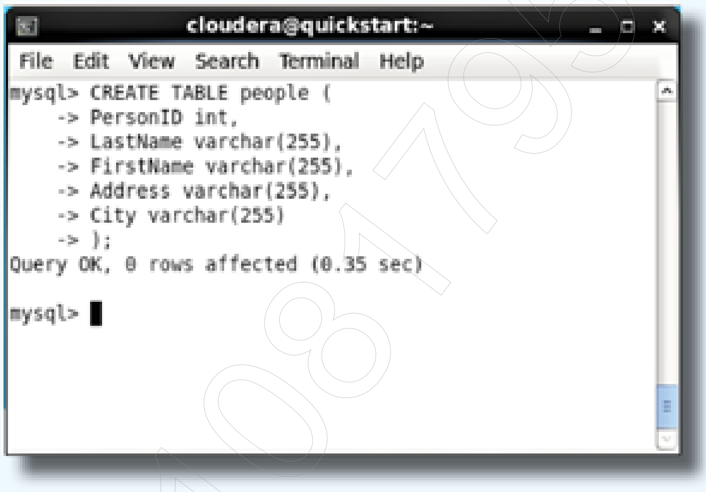
```
CREATE database trendytech;

USE trendytech;
```



```
mysql> create database trendytech;
Query OK, 1 row affected (0.18 sec)

mysql> use trendytech;
Database changed
mysql> 
```

Create a table in MySQL:

```
CREATE TABLE people
(
PersonID int,
LastName varchar(255),
FirstName varchar(255),
Address varchar(255),
City varchar(255)
);
```

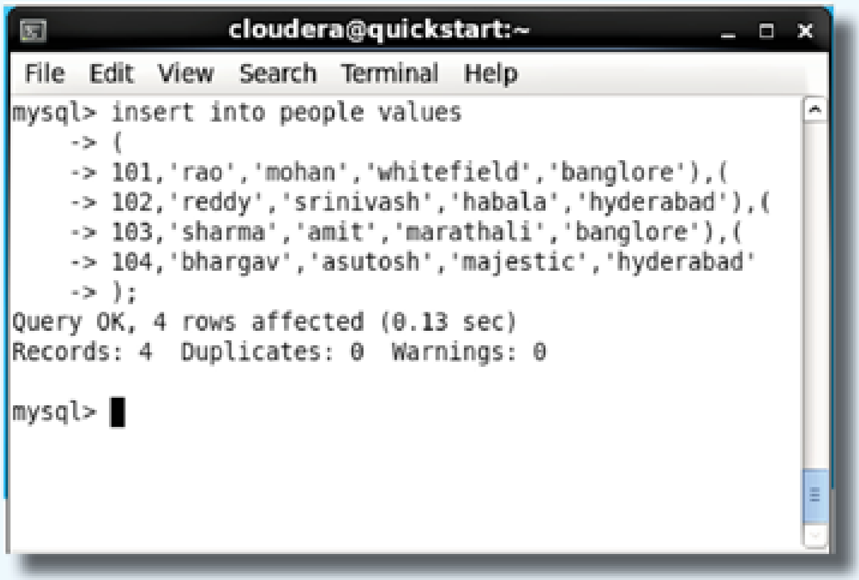


```
mysql> CREATE TABLE people (
    -> PersonID int,
    -> LastName varchar(255),
    -> FirstName varchar(255),
    -> Address varchar(255),
    -> City varchar(255)
    -> );
Query OK, 0 rows affected (0.35 sec)

mysql>
```

Insert records into the people table:

```
insert into people values
(
101,'rao','mohan','whitefield','banglore'),(
102,'reddy','srinivash','habala','hyderabad'),(
103,'sharma','amit','marathali','banglore'),(
104,'bhargav','asutosh','majestic','hyderabad'
);
```

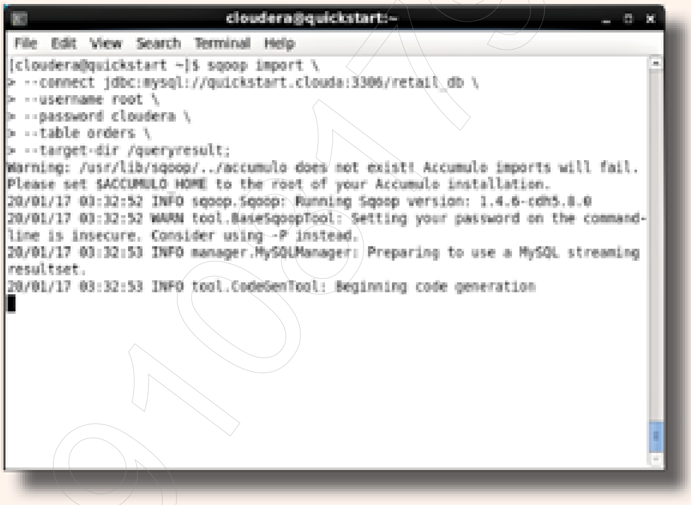
```
commit;
```



```
mysql> insert into people values
    -> (
    -> 101,'rao','mohan','whitefield','banglore'),(
    -> 102,'reddy','srinivash','habala','hyderabad'),(
    -> 103,'sharma','amit','marathali','banglore'),(
    -> 104,'bhargav','asutosh','majestic','hyderabad'
    -> );
Query OK, 4 rows affected (0.13 sec)
Records: 4  Duplicates: 0  Warnings: 0

mysql>
```

Import data from MySQL to Sqoop:

```
sqoop import \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username root \
--password cloudera \
--table orders \
--target-dir /queryresult
```

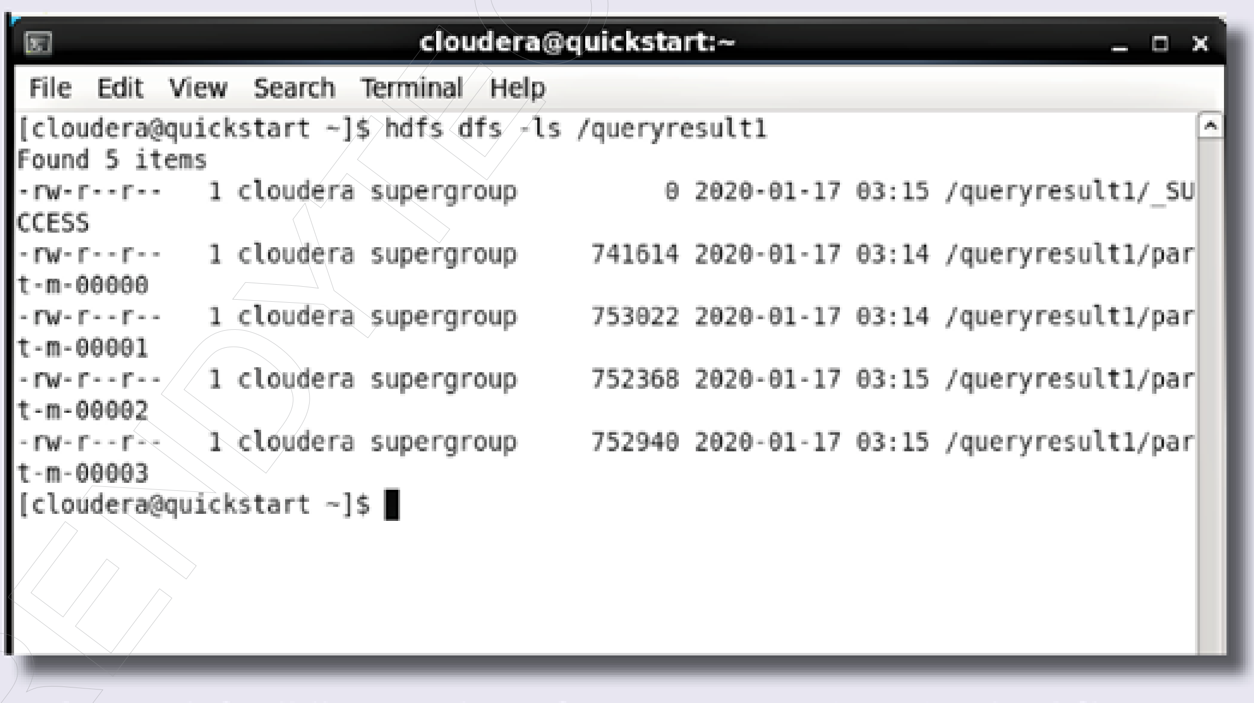


```
[cloudera@quickstart ~]$ sqoop import \
> --connect jdbc:mysql://quickstart.clouda:3306/retail_db \
> --username root \
> --password cloudera \
> --table orders \
> --target-dir /queryresult;
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/17 03:32:52 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
20/01/17 03:32:52 WARN tool.BaseSqoopTool: Setting your password on the command-
line is insecure. Consider using -P instead.
20/01/17 03:32:53 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
20/01/17 03:32:53 INFO tool.CodeGenTool: Beginning code generation
```

Note: If table don't have primary key than it will not import.

To display contents of queryresult directory in HDFS (use terminal):

```
hadoop fs -ls /queryresult
```



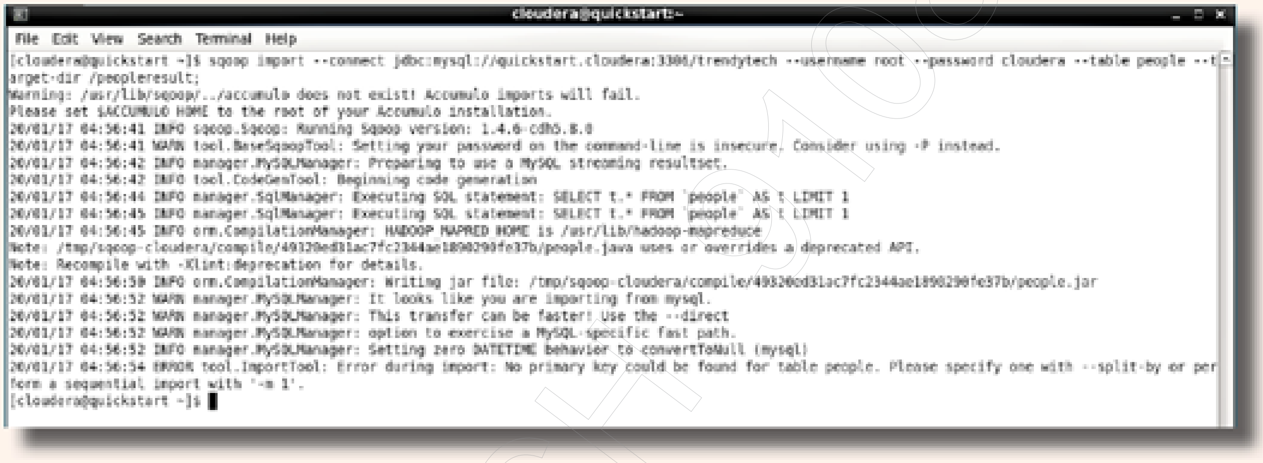
```
[cloudera@quickstart ~]$ hdfs dfs -ls /queryresult1
Found 5 items
-rw-r--r--   1 cloudera supergroup          0 2020-01-17 03:15 /queryresult1/_SU
CCESS
-rw-r--r--   1 cloudera supergroup     741614 2020-01-17 03:14 /queryresult1/par
t-m-00000
-rw-r--r--   1 cloudera supergroup     753022 2020-01-17 03:14 /queryresult1/par
t-m-00001
-rw-r--r--   1 cloudera supergroup     752368 2020-01-17 03:15 /queryresult1/par
t-m-00002
-rw-r--r--   1 cloudera supergroup     752940 2020-01-17 03:15 /queryresult1/par
t-m-00003
[cloudera@quickstart ~]$ 
```

Note: By default the number of mappers are 4, so 4 output files are created.

Instructions ► Import the **people** table (which we have created earlier in MySQL) with same command as we did above.

To import people table from MySQL to HDFS:

```
sqoop import \
--connect jdbc:mysql://quickstart.cloudera:3306/trendytech \
--username root --password cloudera --table people \
--target-dir /peopleresult
```



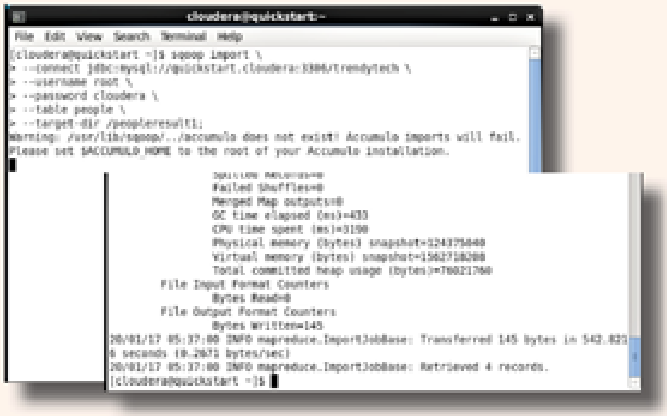
```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/trendytech --username root --password cloudera --table people --t
arget-dir /peopleresult;
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/17 04:56:41 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
20/01/17 04:56:41 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/01/17 04:56:42 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
20/01/17 04:56:42 INFO tool.CodeGenTool: Beginning code generation
20/01/17 04:56:44 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `people` AS t LIMIT 1
20/01/17 04:56:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `people` AS t LIMIT 1
20/01/17 04:56:45 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/49320ed31ac7fc2344ae1898290fe37b/people.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/01/17 04:56:50 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/49320ed31ac7fc2344ae1898290fe37b/people.jar
20/01/17 04:56:52 WARN manager.MySQLManager: It looks like you are importing from mysql.
20/01/17 04:56:52 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
20/01/17 04:56:52 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
20/01/17 04:56:52 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
20/01/17 04:56:54 ERROR tool.ImportTool: Error during import: No primary key could be found for table people. Please specify one with --split-by or per
form a sequential import with '-m 1'.
[cloudera@quickstart ~]$ 
```

NOTE: it will throw error. Becasue **people** table doesn't have primary key.

Instructions ► Now, run the above command with mapper (`-m 1`):

To import people table from MySQL to HDFS with one Mapper:

```
sqoop import \
--connect jdbc:mysql://quickstart.cloudera:3306/trendytech \
--username root \
--password cloudera \
--table people \
-m 1 \
--target-dir /peopleresult1
```

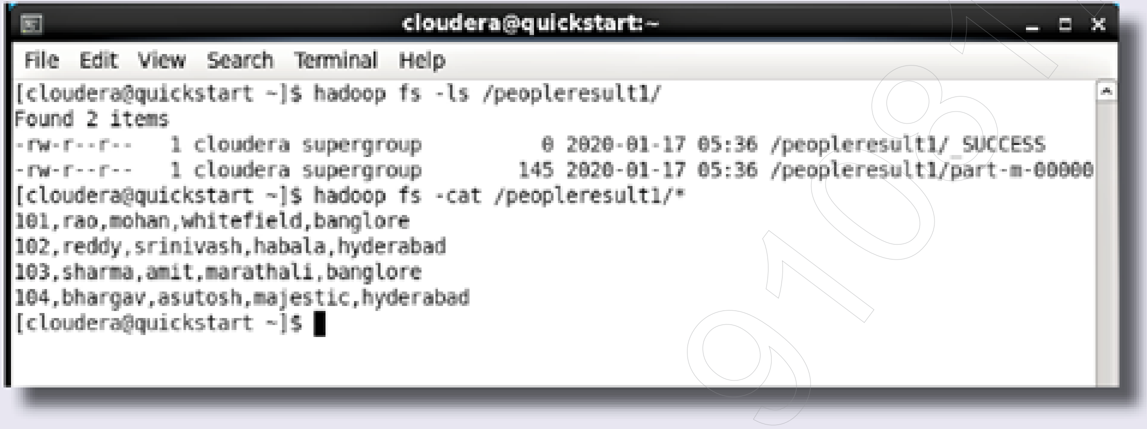


```
[cloudera@quickstart ~]$ sqoop import \
> --connect jdbc:mysql://quickstart.cloudera:3306/trendytech \
> --username root \
> --password cloudera \
> --table people \
> --target-dir /peopleresult1;
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
```

```
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=433
                CPU time spent (ms)=3190
                Physical memory (bytes) snapshot=124375040
                Virtual memory (bytes) snapshot=1562718208
                Total committed heap usage (bytes)=70021760
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=145
20/01/17 05:37:00 INFO mapreduce.ImportJobBase: Transferred 145 bytes in 542.821
6 seconds (0.2671 bytes/sec)
20/01/17 05:37:00 INFO mapreduce.ImportJobBase: Retrieved 4 records.
[cloudera@quickstart ~]$ 
```

To display people table from HDFS:

```
hadoop fs -ls /peopleresult1


hadoop fs -cat /peopleresult1/*
```



```
[cloudera@quickstart ~]$ hadoop fs -ls /peopleresult1/
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2020-01-17 05:36 /peopleresult1/_SUCCESS
-rw-r--r--   1 cloudera supergroup        145 2020-01-17 05:36 /peopleresult1/part-m-00000
[cloudera@quickstart ~]$ hadoop fs -cat /peopleresult1/*
101,rao,mohan,whitefield,banglore
102,reddy,srinivash,habala,hyderabad
103,sharma,amit,marathali,banglore
104,bhargav,asutosh,majestic,hyderabad
[cloudera@quickstart ~]$ 
```

Note: You will find one mapper file only (**part-m-00000**).

To import all tables from "MySQL" database:

```
sqoop-import-all-tables \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username retail_dba \
--password cloudera \
--as-sequencefile \   <------------------------ File Format
-m 4 \
--warehouse-dir /user/cloudera/sqoopdir
```

Note: Here no of mappers are 4 that means we will get 4 files

We can also mention **file format** while importing data as mentioned above.

Sqoop supports **4 types** of file formats:

- Text file format
- Sequence file format
- Avro file format
- Parquet file format

Note: If you do not mention any file format, by default it will be text file format.

By default Sqoop provides 4 mappers - so we can skip the above `-m 4` command and still get the same result.

Difference between Sqoop **target** directory & **warehouse** directory.

The difference is that:

–target-dir is a full directory path and the data files will be created directly inside the specified folder.

–warehouse-dir is used to specify a base directory within hdfs where SQOOP will create a sub folder inside with the name of the source table, and import the data files into that folder.

Directory structure for **retail_db** will be:

/user/cloudera/sqoopdir/employee
/user/cloudera/sqoopdir/customer
/user/cloudera/sqoopdir/table3
/user/cloudera/sqoopdir/tablw4

Now try to run following code to import the *orders* table with `--warehouse-dir` path:

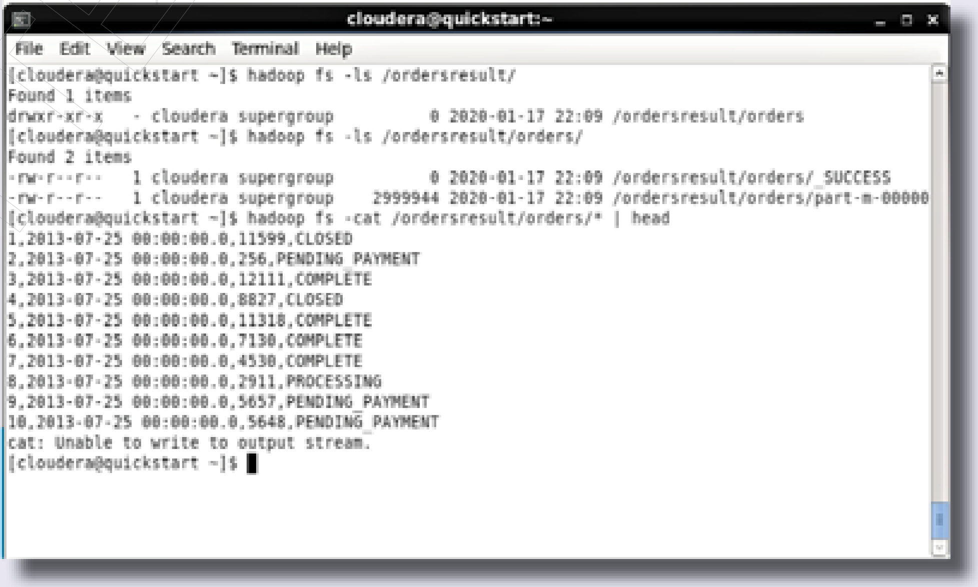
```
sqoop import \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username root \
--password cloudera \
--table orders \
--warehouse-dir /ordersresult
```

To check the file structure in HDFS:

```
hadoop fs -ls /ordersresult/
```

/user/cloudera/warehouse/ordersresult/orders

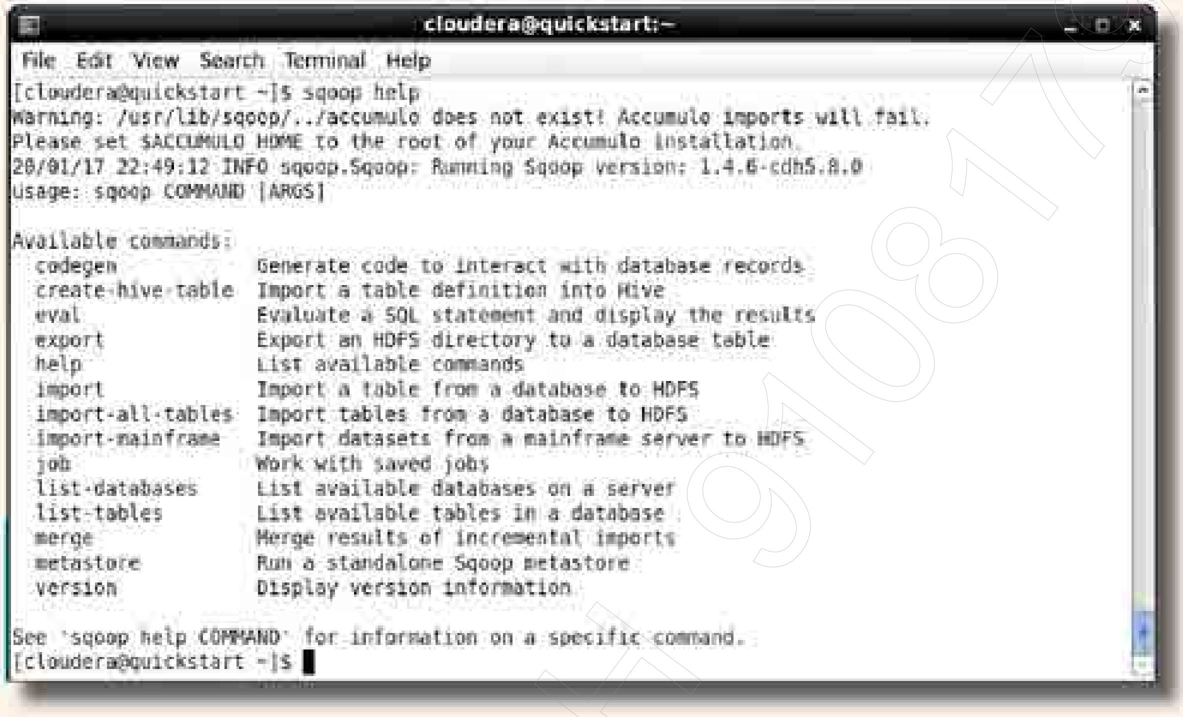
part-m-00000_0



```
[cloudera@quickstart ~]$ hadoop fs -ls /ordersresult/
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2020-01-17 22:09 /ordersresult/orders
[cloudera@quickstart ~]$ hadoop fs -ls /ordersresult/orders/
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2020-01-17 22:09 /ordersresult/orders/_SUCCESS
-rw-r--r--   1 cloudera supergroup    2999944 2020-01-17 22:09 /ordersresult/orders/part-m-00000
[cloudera@quickstart ~]$ hadoop fs -cat /ordersresult/orders/* | head
1,2013-07-25 00:00:00.0,11599,CLOSED
2,2013-07-25 00:00:00.0,256,PENDING_PAYMENT
3,2013-07-25 00:00:00.0,12111,COMPLETE
4,2013-07-25 00:00:00.0,8827,CLOSED
5,2013-07-25 00:00:00.0,11318,COMPLETE
6,2013-07-25 00:00:00.0,7130,COMPLETE
7,2013-07-25 00:00:00.0,4530,COMPLETE
8,2013-07-25 00:00:00.0,2911,PROCESSING
9,2013-07-25 00:00:00.0,5657,PENDING_PAYMENT
10,2013-07-25 00:00:00.0,5648,PENDING_PAYMENT
cat: Unable to write to output stream.
[cloudera@quickstart ~]$ 
```

To display a list of all available tools:

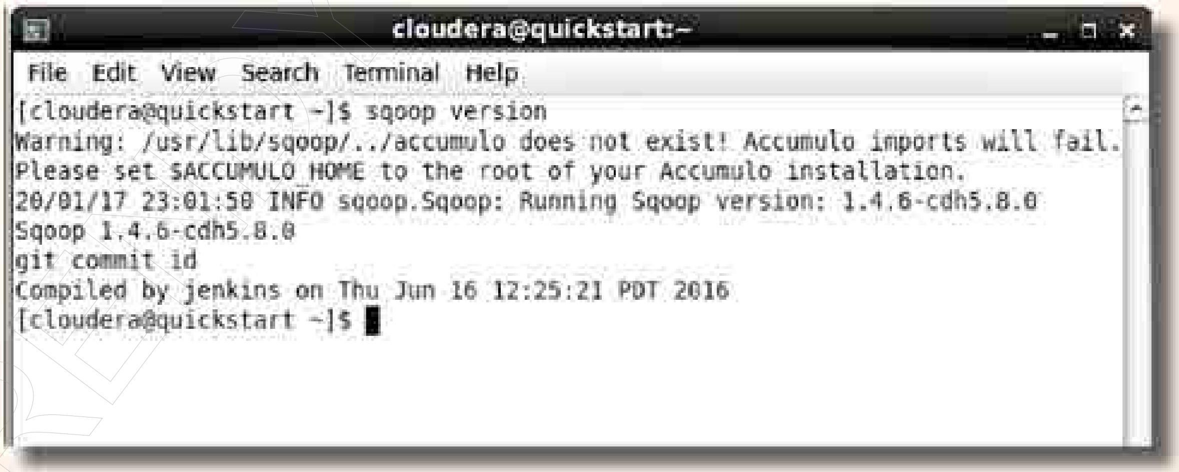
```
sqoop help
```



```
[cloudera@quickstart ~]$ sqoop help
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/17 22:49:12 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
usage: sqoop COMMAND [ARGS]

Available commands:
  codegen            Generate code to interact with database records
  create-hive-table  Import a table definition into Hive
  eval               Evaluate a SQL statement and display the results
  export             Export an HDFS directory to a database table
  help               List available commands
  import             Import a table from a database to HDFS
  import-all-tables  Import tables from a database to HDFS
  import-mainframe   Import datasets from a mainframe server to HDFS
  job                Work with saved jobs
  list-databases     List available databases on a server
  list-tables        List available tables in a database
  merge              Merge results of incremental imports
  metastore          Run a standalone Sqoop metastore
  version            Display version information

See 'sqoop help COMMAND' for information on a specific command.
[cloudera@quickstart ~]$ 
```

To know sqoop version:

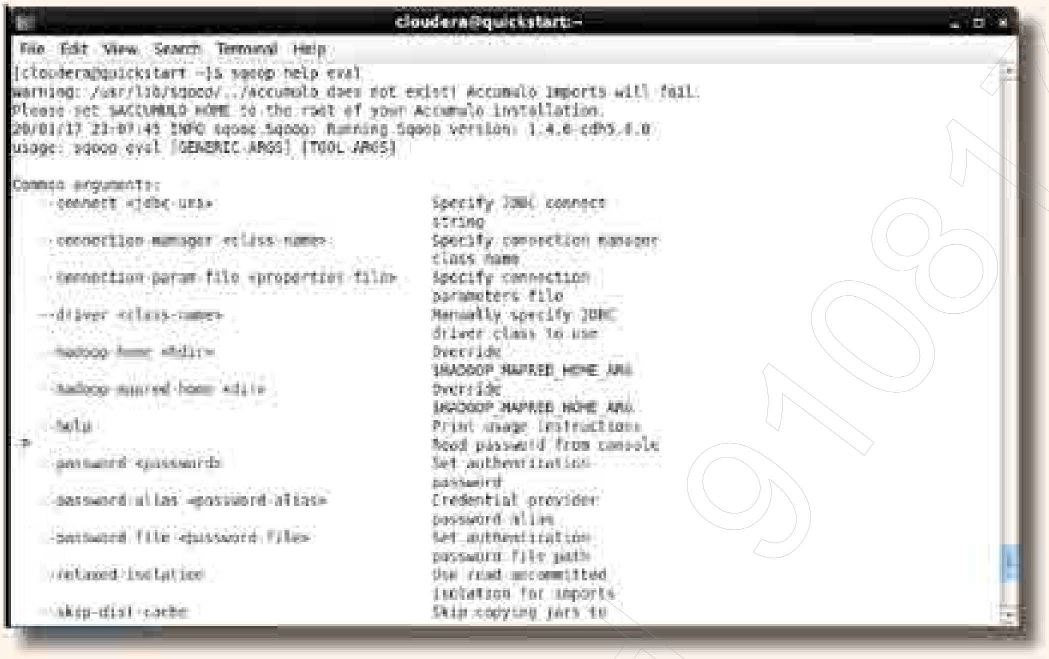
```
sqoop version
```



```
[cloudera@quickstart ~]$ sqoop version
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/17 23:01:50 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
Sqoop 1.4.6-cdh5.8.0
git commit id
Compiled by jenkins on Thu Jun 16 12:25:21 PDT 2016
[cloudera@quickstart ~]$ 
```

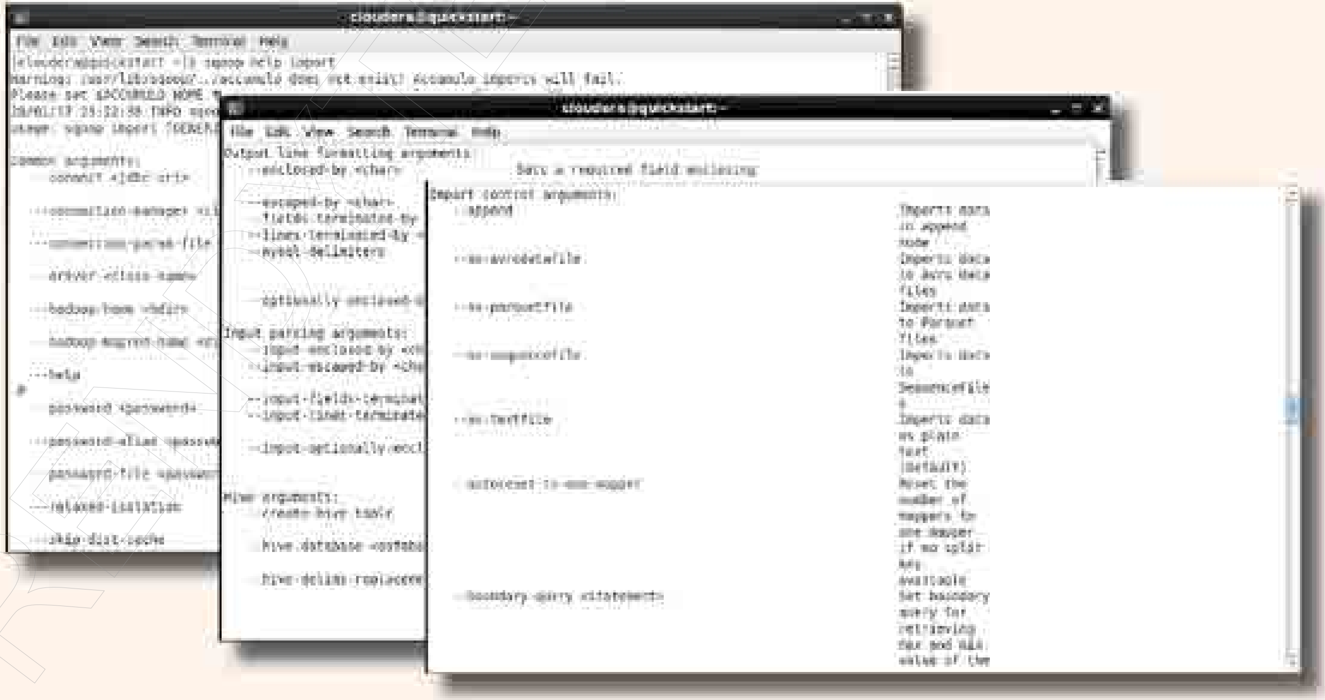
Sqoop help with command Aliases

```
sqoop help eval
```



```
[cloudera@quickstart ~]$ sqoop help eval
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/17 23:07:45 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.8.0
usage: sqoop eval [GENERIC-ARGS] [TOOL-ARGS]

Common arguments:
   --connect <jdbc-uri>                         Specify JDBC connect
                                                string
   --connection-manager <class-name>            Specify connection manager
                                                class name
   --connection-param-file <properties-file>    Specify connection
                                                parameters file
   --driver <class-name>                        Manually specify JDBC
                                                driver class to use
   --hadoop-home <hdir>                         Override
                                                $HADOOP_MAPRED_HOME_ARG
   --hadoop-mapred-home <dir>                   Override
                                                $HADOOP_MAPRED_HOME_ARG
   --help                                       Print usage instructions
-P                                              Read password from console
   --password <password>                        Set authentication
                                                password
   --password-alias <password-alias>            Credential provider
                                                password alias
   --password-file <password-file>              Set authentication
                                                password file path
   --relaxed-isolation                          Use read-uncommitted
                                                isolation for imports
   --skip-dist-cache                            Skip copying jars to
```

```
sqoop help import
```



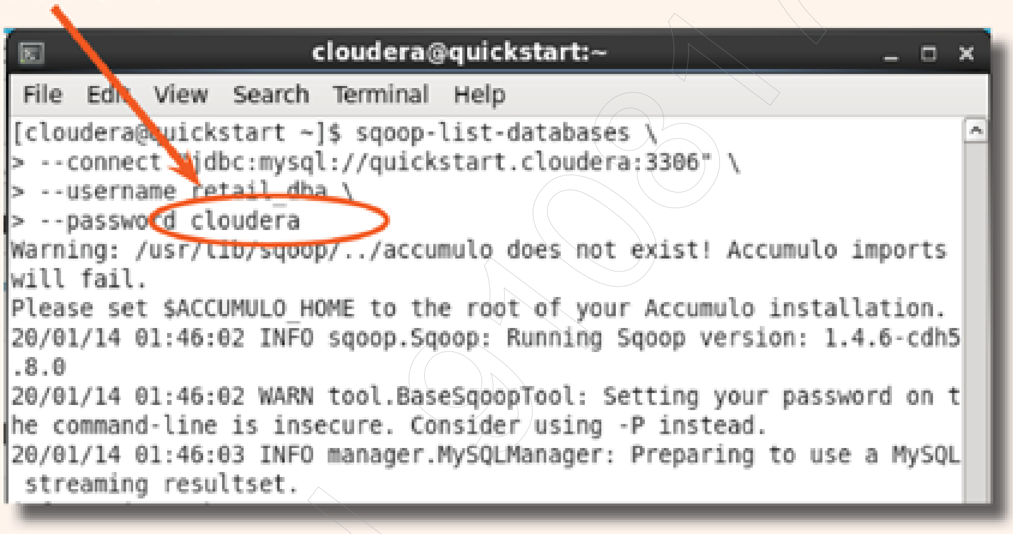
```
[cloudera@quickstart ~]$ sqoop help import
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME t
20/01/17 23:22:38 INFO sqoo
usage: sqoop import [GENERI

Common arguments:
   --connect <jdbc-uri>

   --connection-manager <cl

   --connection-param-file

   --driver <class-name>

   --hadoop-home <hdir>

   --hadoop-mapred-home <di

   --help
-P
   --password <password>

   --password-alias <passwo

   --password-file <passwor

   --relaxed-isolation

   --skip-dist-cache
```

```
Output line formatting arguments:
   --enclosed-by <char>              Sets a required field enclosing

   --escaped-by <char>
   --fields-terminated-by
   --lines-terminated-by <
   --mysql-delimiters


   --optionally-enclosed-b

Input parsing arguments:
   --input-enclosed-by <ch
   --input-escaped-by <cha

   --input-fields-terminat
   --input-lines-terminate

   --input-optionally-encl


Hive arguments:
   --create-hive-table

   --hive-database <databa

   --hive-delims-replaceme
```

```
Import control arguments:
   --append                                       Imports data
                                                  in append
                                                  mode
   --as-avrodatafile                              Imports data
                                                  to Avro data
                                                  files
   --as-parquetfile                               Imports data
                                                  to Parquet
                                                  files
   --as-sequencefile                              Imports data
                                                  to
                                                  SequenceFile
                                                  s
   --as-textfile                                  Imports data
                                                  as plain
                                                  text
                                                  (default)
   --autoreset-to-one-mapper                      Reset the
                                                  number of
                                                  mappers to
                                                  one mapper
                                                  if no split
                                                  key
                                                  available
   --boundary-query <statement>                   Set boundary
                                                  query for
                                                  retrieving
                                                  max and min
                                                  value of the
```

The argument **--password** takes authentication password in plain text.

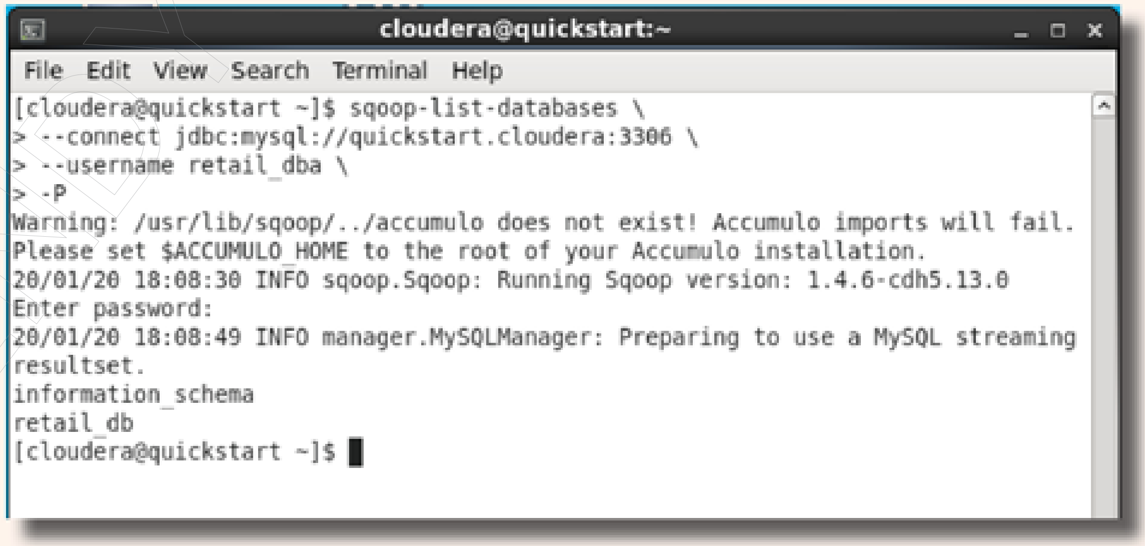
```
sqoop-list-databases \
--connect jdbc:mysql://quickstart.cloudera:3306 \
--username retail_dba \
--password cloudera
```



```
[cloudera@quickstart ~]$ sqoop-list-databases \
> --connect "jdbc:mysql://quickstart.cloudera:3306" \
> --username retail_dba \
> --password cloudera
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports
will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/14 01:46:02 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5
.8.0
20/01/14 01:46:02 WARN tool.BaseSqoopTool: Setting your password on t
he command-line is insecure. Consider using -P instead.
20/01/14 01:46:03 INFO manager.MySQLManager: Preparing to use a MySQL
 streaming resultset.
```

While the argument **-P** read password from console.

```
sqoop-list-databases \
--connect jdbc:mysql://quickstart.cloudera:3306 \
--username retail_dba \
-P
```



```
[cloudera@quickstart ~]$ sqoop-list-databases \
> --connect jdbc:mysql://quickstart.cloudera:3306 \
> --username retail_dba \
> -P
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/20 18:08:30 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
Enter password:
20/01/20 18:08:49 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
information_schema
retail_db
[cloudera@quickstart ~]$ █
```

The argument `--query` can be replaced with `-e`.

```
sqoop-eval \
--connect jdbc:mysql://quickstart.cloudera:3306 \
--username retail_dba \
--password cloudera \
--query "select * from retail_db.customers limit 10"
```

OR

```
sqoop-eval \
--connect "jdbc:mysql://quickstart.cloudera:3306" \
--username retail_dba \
--password cloudera \
-e "select * from retail_db.customers limit 10"
```

Similarly `-m` and `--num-mappers` are same.

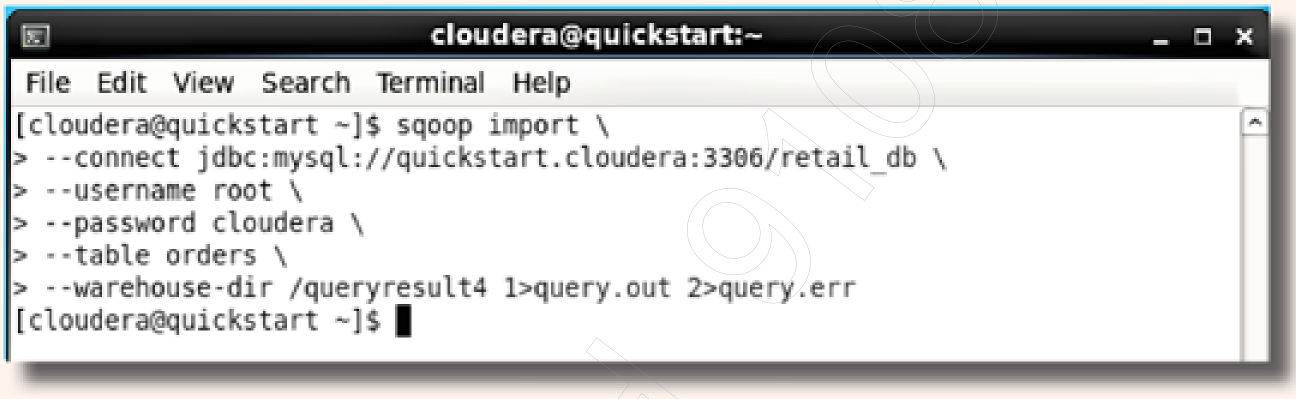
```
sqoop import \
--connect jdbc:mysql://quickstart.cloudera:3306/trendytech \
--username root \
--password cloudera \
--table people -m 1 \
--target-dir /peopleresult1
```

OR

```
sqoop import \
--connect jdbc:mysql://quickstart.cloudera:3306/trendytech \
--username root \
--password cloudera \
--table people --num-mappers 1 \
--target-dir /peopleresult1
```

Redirecting logs:

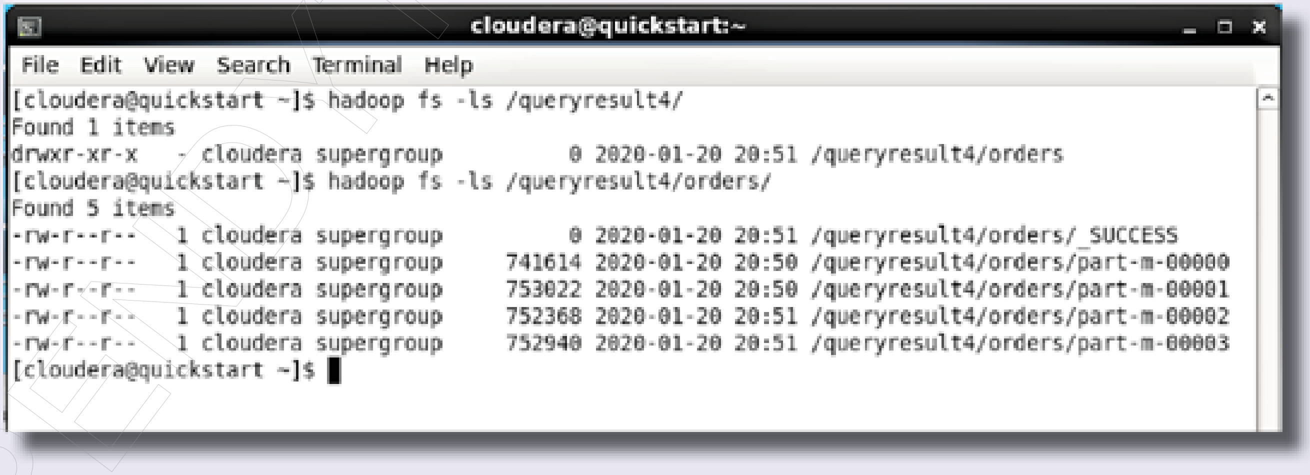
```
sqoop import \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username root \
--password cloudera \
--table orders \
--warehouse-dir /queryresult4 1>query.out 2>query.err
```



```
[cloudera@quickstart ~]$ sqoop import \
> --connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
> --username root \
> --password cloudera \
> --table orders \
> --warehouse-dir /queryresult4 1>query.out 2>query.err
[cloudera@quickstart ~]$ 
```

To check the content of the queryresult4:

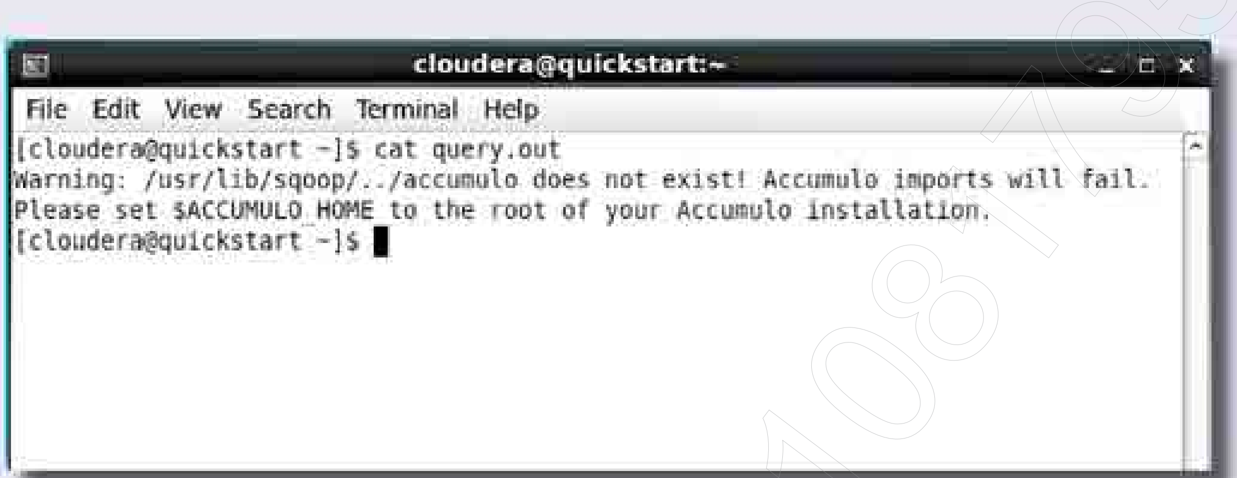
```
hadoop fs -ls /queryresult4/orders/
```



```
[cloudera@quickstart ~]$ hadoop fs -ls /queryresult4/
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2020-01-20 20:51 /queryresult4/orders
[cloudera@quickstart ~]$ hadoop fs -ls /queryresult4/orders/
Found 5 items
-rw-r--r--   1 cloudera supergroup          0 2020-01-20 20:51 /queryresult4/orders/_SUCCESS
-rw-r--r--   1 cloudera supergroup     741614 2020-01-20 20:50 /queryresult4/orders/part-m-00000
-rw-r--r--   1 cloudera supergroup     753022 2020-01-20 20:50 /queryresult4/orders/part-m-00001
-rw-r--r--   1 cloudera supergroup     752368 2020-01-20 20:51 /queryresult4/orders/part-m-00002
-rw-r--r--   1 cloudera supergroup     752940 2020-01-20 20:51 /queryresult4/orders/part-m-00003
[cloudera@quickstart ~]$ 
```

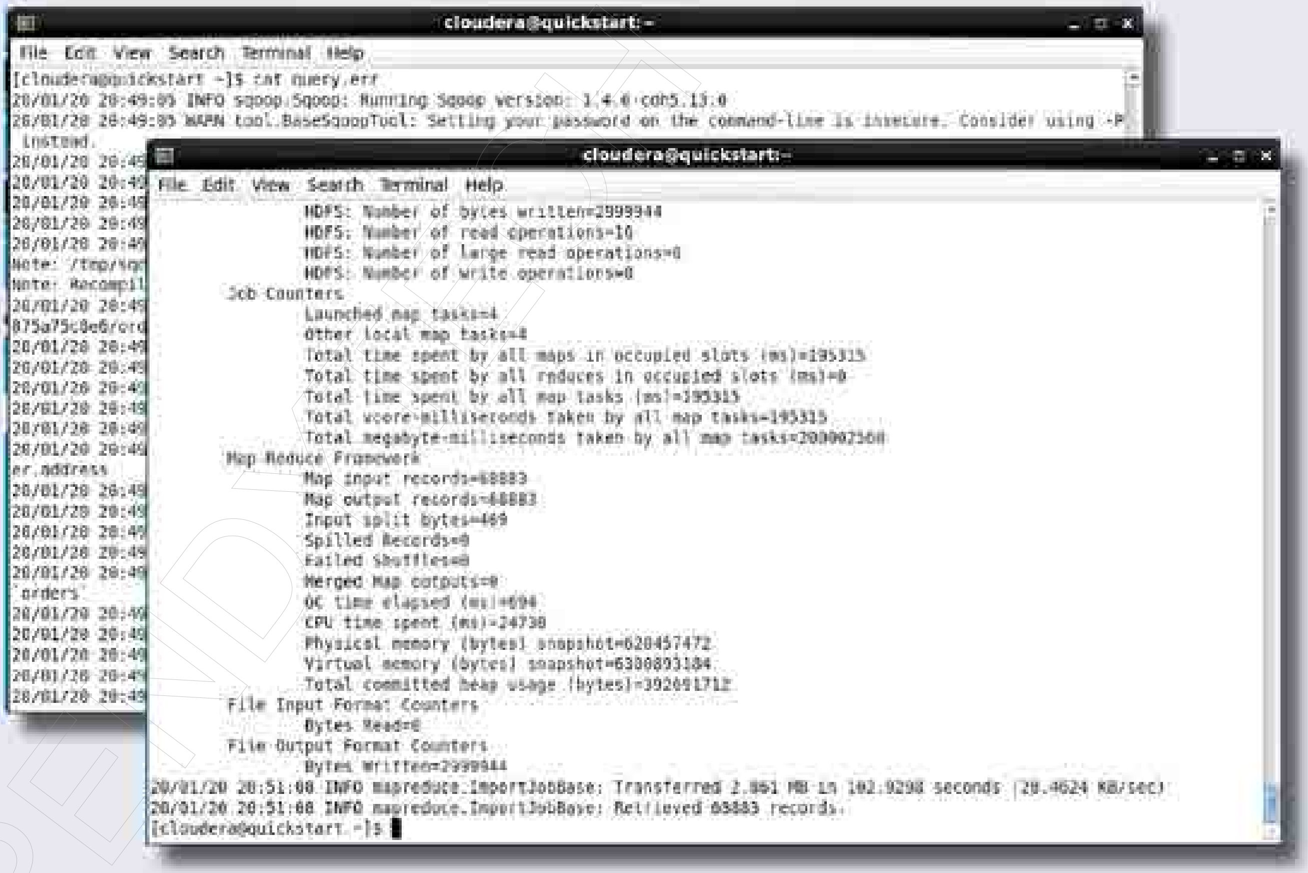
To check the contents of log files:

```
cat query.out
```



```
[cloudera@quickstart ~]$ cat query.out
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
[cloudera@quickstart ~]$ 
```

```
cat query.err
```



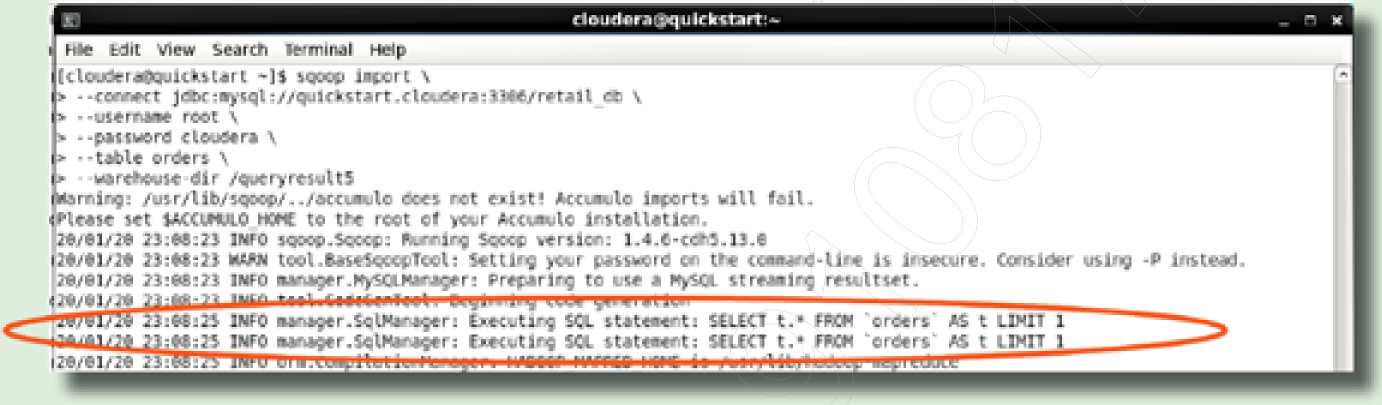
```
[cloudera@quickstart ~]$ cat query.err
20/01/20 20:49:05 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
20/01/20 20:49:05 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P
 instead.
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
Note: /tmp/sqo
Note: Recompil
20/01/20 20:49
875a75c0e6/ord
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
er.address
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
`orders`
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
20/01/20 20:49
```

```
                HDFS: Number of bytes written=2999944
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Job Counters
                Launched map tasks=4
                Other local map tasks=4
                Total time spent by all maps in occupied slots (ms)=195315
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=195315
                Total vcore-milliseconds taken by all map tasks=195315
                Total megabyte-milliseconds taken by all map tasks=200002560
        Map-Reduce Framework
                Map input records=68883
                Map output records=68883
                Input split bytes=469
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=694
                CPU time spent (ms)=24730
                Physical memory (bytes) snapshot=620457472
                Virtual memory (bytes) snapshot=6300893184
                Total committed heap usage (bytes)=392691712
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=2999944
20/01/20 20:51:08 INFO mapreduce.ImportJobBase: Transferred 2.861 MB in 102.9298 seconds (28.4624 KB/sec)
20/01/20 20:51:08 INFO mapreduce.ImportJobBase: Retrieved 68883 records.
[cloudera@quickstart ~]$ 
```

Sqoop import execution flow

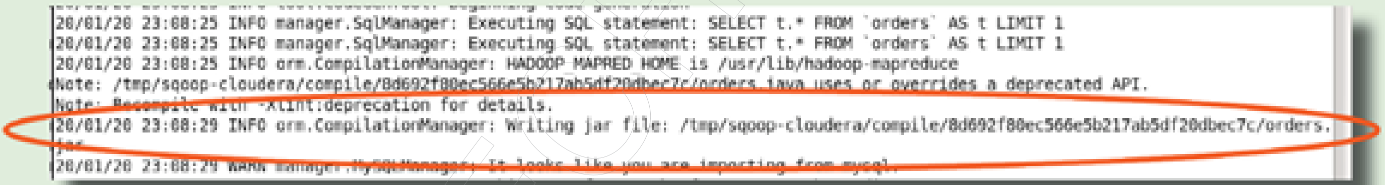
How Mappers devide their work when a query fired:

- Selects 1 record and by using that it gets the metadata and builds the java file



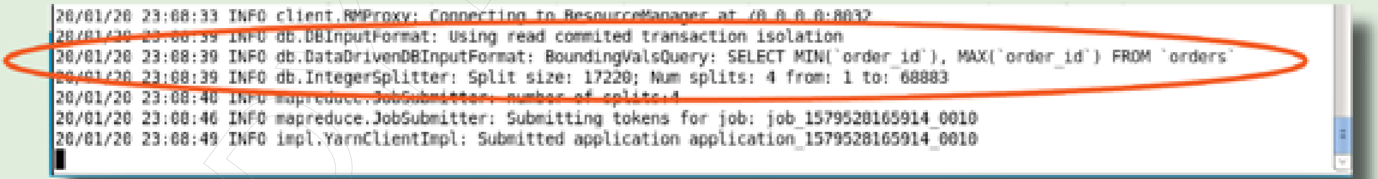
```
[cloudera@quickstart ~]$ sqoop import \
> --connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
> --username root \
> --password cloudera \
> --table orders \
> --warehouse-dir /queryresult5
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/01/20 23:08:23 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
20/01/20 23:08:23 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/01/20 23:08:23 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
20/01/20 23:08:23 INFO tool.CodeGenTool: Beginning code generation
20/01/20 23:08:25 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
20/01/20 23:08:25 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
20/01/20 23:08:25 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
```

- Using above java file it builds the jar file



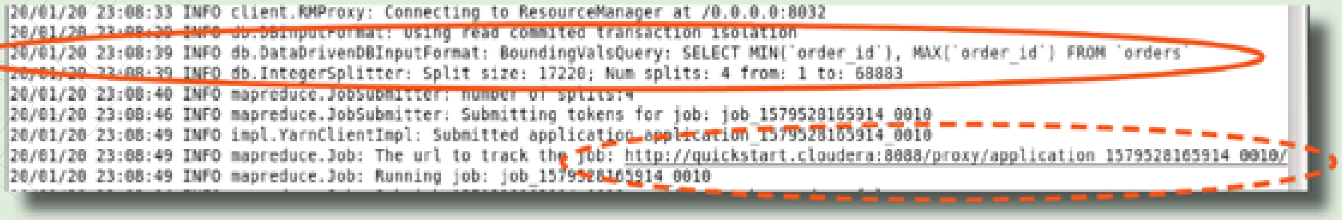
```
20/01/20 23:08:25 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
20/01/20 23:08:25 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `orders` AS t LIMIT 1
20/01/20 23:08:25 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/8d692f80ec566e5b217ab5df20dbec7c/orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/01/20 23:08:29 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/8d692f80ec566e5b217ab5df20dbec7c/orders.
jar
20/01/20 23:08:29 WARN manager.MySQLManager: It looks like you are importing from mysql.
```

- **BoundingValsQuery** based on min and max on primary key



```
20/01/20 23:08:33 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/01/20 23:08:39 INFO db.DBInputFormat: Using read commited transaction isolation
20/01/20 23:08:39 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`order_id`), MAX(`order_id`) FROM `orders`
20/01/20 23:08:39 INFO db.IntegerSplitter: Split size: 17220; Num splits: 4 from: 1 to: 68883
20/01/20 23:08:40 INFO mapreduce.JobSubmitter: number of splits:4
20/01/20 23:08:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1579528165914_0010
20/01/20 23:08:49 INFO impl.YarnClientImpl: Submitted application application_1579528165914_0010
```

- Calculates (max - min)/4 and it gets the split size.



```
20/01/20 23:08:33 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/01/20 23:08:39 INFO db.DBInputFormat: Using read commited transaction isolation
20/01/20 23:08:39 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`order_id`), MAX(`order_id`) FROM `orders`
20/01/20 23:08:39 INFO db.IntegerSplitter: Split size: 17220; Num splits: 4 from: 1 to: 68883
20/01/20 23:08:40 INFO mapreduce.JobSubmitter: number of splits:4
20/01/20 23:08:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1579528165914_0010
20/01/20 23:08:49 INFO impl.YarnClientImpl: Submitted application application_1579528165914_0010
20/01/20 23:08:49 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1579528165914_0010/
20/01/20 23:08:49 INFO mapreduce.Job: Running job: job_1579528165914_0010
```

File formats:

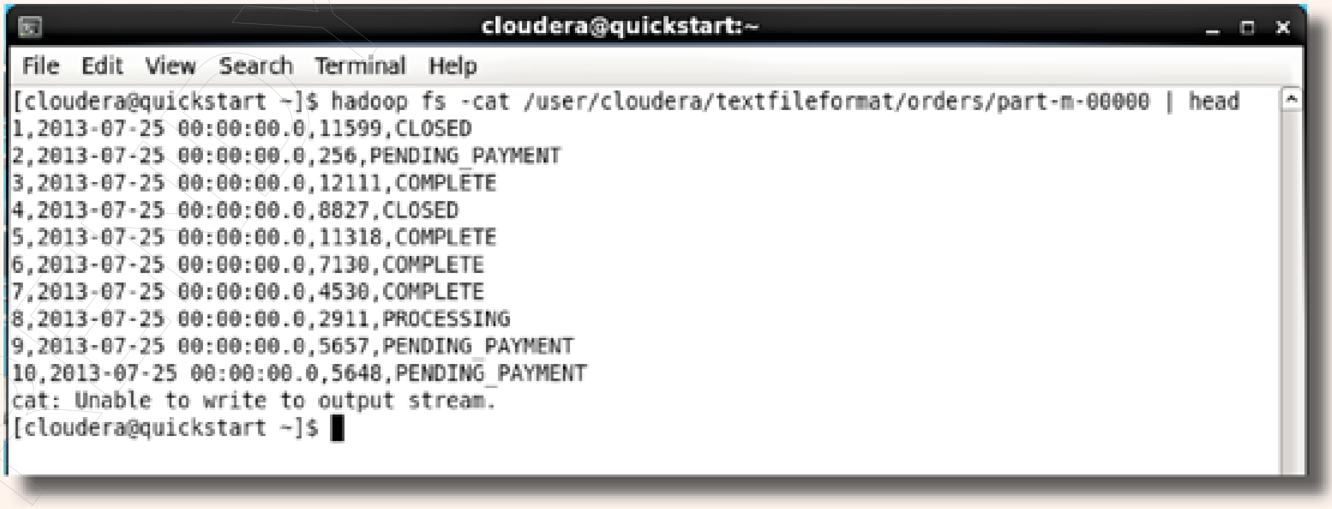
Sqoop import supports following file formats:

1. Text file format - command argument **--as-textfile**
2. Sequence file format - command argument **--as-sequencefile**
3. Avro file format - command argument **--as-avrodatafile**
4. Parquet file format - command argument **--as-parquetfile**

Note: If you are not mentioning any file format, by default sqoop uses **--as-textfile**

Text file format:

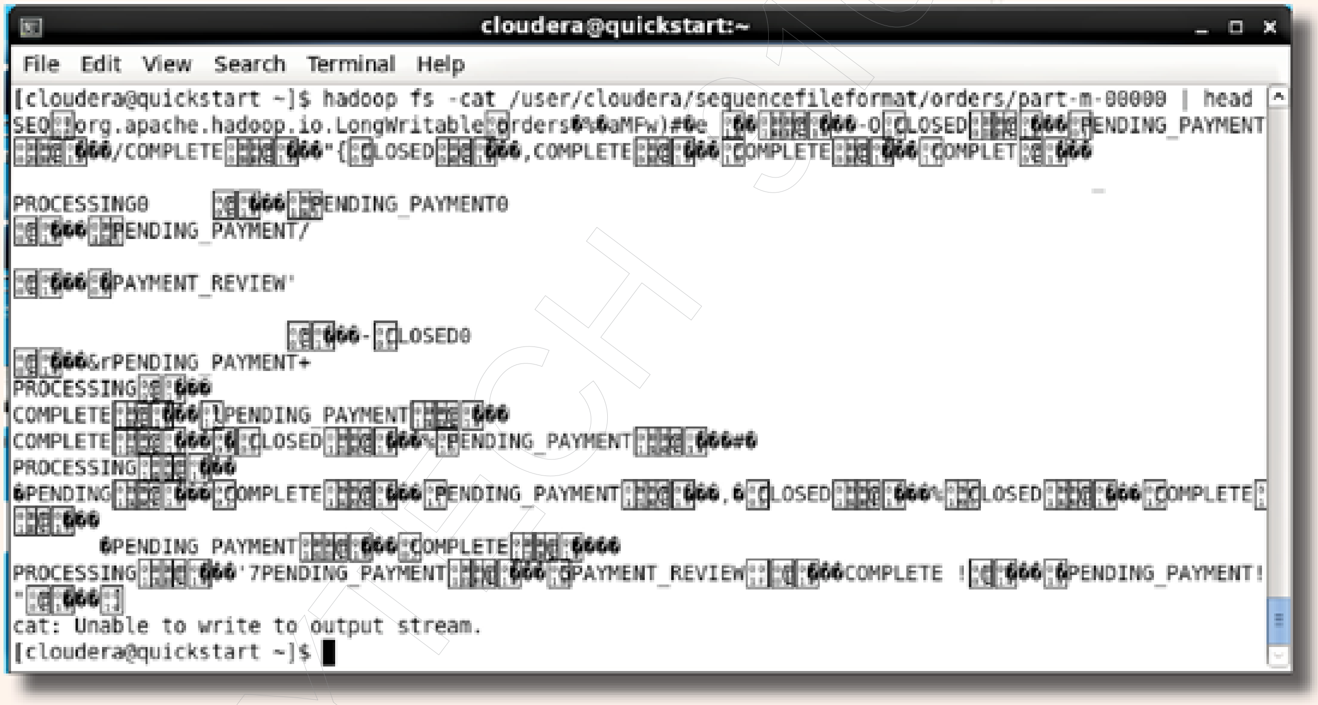
```
sqoop-import \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username retail_dba \
--password cloudera \
--table orders \
--as-textfile \
-m 4 \
--warehouse-dir /user/cloudera/textfileformat
```



```
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/textfileformat/orders/part-m-00000 | head
1,2013-07-25 00:00:00.0,11599,CLOSED
2,2013-07-25 00:00:00.0,256,PENDING_PAYMENT
3,2013-07-25 00:00:00.0,12111,COMPLETE
4,2013-07-25 00:00:00.0,8827,CLOSED
5,2013-07-25 00:00:00.0,11318,COMPLETE
6,2013-07-25 00:00:00.0,7130,COMPLETE
7,2013-07-25 00:00:00.0,4530,COMPLETE
8,2013-07-25 00:00:00.0,2911,PROCESSING
9,2013-07-25 00:00:00.0,5657,PENDING_PAYMENT
10,2013-07-25 00:00:00.0,5648,PENDING_PAYMENT
cat: Unable to write to output stream.
[cloudera@quickstart ~]$
```

Sequence file format:

```
sqoop-import \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username retail_dba \
--password cloudera \
--table orders \
--as-sequencefile \
-m 4 \
--warehouse-dir /user/cloudera/sequencefileformat
```

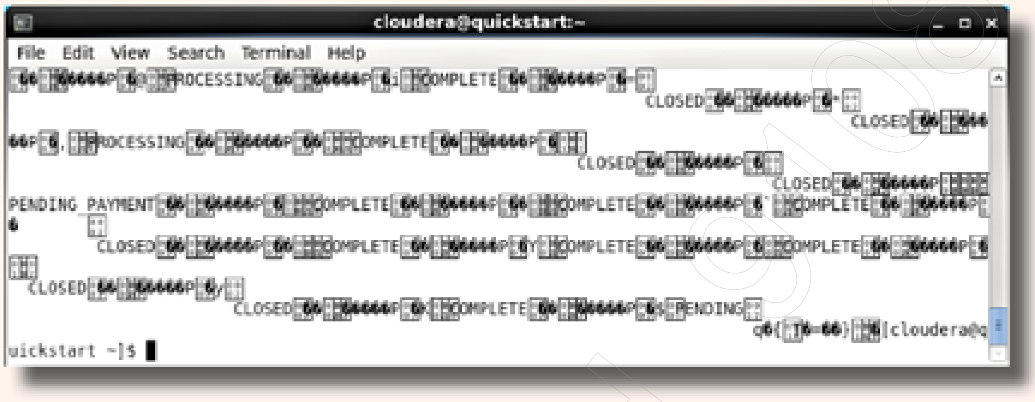


```
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/sequencefileformat/orders/part-m-00000 | head
SEQ□□org.apache.hadoop.io.LongWritable□orders��aMFw)#�e □��□□□□���-O□CLOSED□□□□���□PENDING_PAYMENT
□□□□���/COMPLETE□□□□���"{□CLOSED□□□□���,COMPLETE□□□□���□COMPLETE□□□□���□COMPLET□□□���

PROCESSING0     □□□���□□PENDING_PAYMENT0
□□□���□□PENDING_PAYMENT/

□□□���□PAYMENT_REVIEW'

                        □□□���-□CLOSED0
□□□���&rPENDING_PAYMENT+
PROCESSING□□□���
COMPLETE□□□□���□lPENDING_PAYMENT□□□□���
COMPLETE□□□□���□�□CLOSED□□□□���%□PENDING_PAYMENT□□□□����#�
PROCESSING□□□□���
�PENDING□□□□���□COMPLETE□□□□���□PENDING_PAYMENT□□□□���,�□CLOSED□□□□���%□□CLOSED□□□□���□COMPLETE□
□□□□���
       �PENDING_PAYMENT□□□□���□COMPLETE□□□□����
PROCESSING□□□□���'7PENDING_PAYMENT□□□□���□□PAYMENT_REVIEW□□□□���COMPLETE !□□□���□PENDING_PAYMENT!
"□□□���□
cat: Unable to write to output stream.
[cloudera@quickstart ~]$ █
```

Note: SequenceFiles are a binary format that store individual records in custom record-specific data types. These data types are manifested as Java classes.

Avro file format:

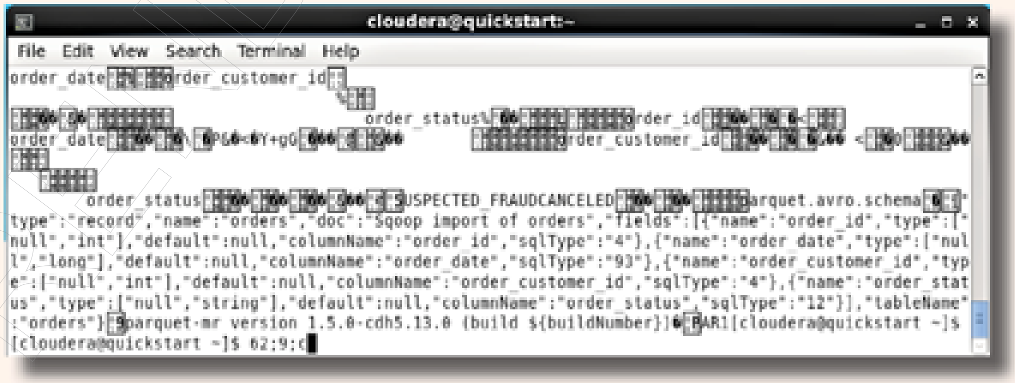
```
sqoop-import \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username retail_dba --password cloudera \
--table orders \
--as-avrodatafile -m 4 \
--warehouse-dir /user/cloudera/avrofileformat
```



```
��������P�����PROCESSING�������������P����COMPLETE�����������P��-��
                                                           CLOSED��������P��*��
                                                                          CLOSED������
��P�,���PROCESSING������������P��������COMPLETE�����������P�����
                                                       CLOSED�����������P���
                                                                      CLOSED���������P����
PENDING_PAYMENT�������������P������COMPLETE������������P��������COMPLETE������������P���COMPLETE������������P��
�      ��
      CLOSED�����������P�������COMPLETE������������P��Y���COMPLETE������������P�����COMPLETE������������P��
���
  CLOSED���������P��/��
                              CLOSED���������P���COMPLETE������������P���PENDING��
                                                                    q�{��T�=��}����[cloudera@q
uickstart ~]$ 
```

Parquet file format:

```
sqoop-import \
--connect jdbc:mysql://quickstart.cloudera:3306/retail_db \
--username retail_dba --password cloudera \
--table orders \
--as-parquetfile -m 4 \
--warehouse-dir /user/cloudera/parquetfileformat
```



```
order_date�����order_customer_id��
                                     ����

��������������                        order_status������������order_id����������
order_date������\��P&�<�Y+gG��������           ��������order_customer_id�������� <����������
���
   �����
       order_status�����������������SUSPECTED_FRAUDCANCELED�������������parquet.avro.schema���"
type":"record","name":"orders","doc":"Sqoop import of orders","fields":[{"name":"order_id","type":["
null","int"],"default":null,"columnName":"order_id","sqlType":"4"},{"name":"order_date","type":["nul
l","long"],"default":null,"columnName":"order_date","sqlType":"93"},{"name":"order_customer_id","typ
e":["null","int"],"default":null,"columnName":"order_customer_id","sqlType":"4"},{"name":"order_stat
us","type":["null","string"],"default":null,"columnName":"order_status","sqlType":"12"}],"tableName"
:"orders"}��parquet-mr version 1.5.0-cdh5.13.0 (build ${buildNumber})���PAR1[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ 62;9;c
```

Google REVIEWS

5 Star Google Rated Big Data Course

LEARN FROM THE EXPERT

9108179578

Call for more details

Follow US

Trainer	Mr. Sumit Mittal
LinkedIn	https://www.linkedin.com/in/bigdatabysumit/
Website	https://trendytech.in/courses/big-data-online-training/
Phone	9108179578
Email	trendytech.sumit@gmail.com
Youtube	TrendyTech
Twitter	@BigdataBySumit
Instagram	bigdatabysumit
Facebook	https://www.facebook.com/trendytech.in/

TRENDY TECH

UPLIFT YOUR CAREER!