

Big Data Management and Analytics Lab

Lab Exercise 11: Sqoop

Objective of this exercise is to import MySQL into HDFS using Sqoop

1. Under Hue, you have to go on to Query, then choose Editor and finally, choose Sqoop. The screen looks like, as shown below:
2. Write the statement for Get MySQL database into MySQL server
`mysql -u root -p`

```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 14
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.

3. Show Databases available in the MySQL
`show databases;`

Output

```
mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| cm |
| firehose |
| hue |
| metastore |
| mysql |
| nav |
| navms |
| oozie |
| retail_db |
| rman |
| sentry |
+-----+
12 rows in set (0.01 sec)
```

Big Data Management and Analytics Lab

4. Use retail_db database and check the tables in it:

```
use retail_db

// command for list available tables in retail_db
show tables;
```

Output

```
mysql> use retail_db
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

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

5. List available tables in local host using Sqoop

```
// command for get local server name
Hostname -f

//command for display local host tables
sqoop list-databases --connect jdbc:mysql://localhost/ --password cloudera --
username root; //Lists the databases using Sqoop
```

Output

Big Data Management and Analytics Lab

```
[cloudera@quickstart ~]$ sqoop list-databases --connect jdbc:mysql://localhost/ --password cloudera --username root;
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/10/10 21:44:57 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
22/10/10 21:44:57 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/10/10 21:44:57 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
information_schema
cm
firehose
hue
metastore
mysql
nav
navms
oozie
retail_db
rman
sentry
[cloudera@quickstart ~]$
```

6. To connect to the MySQL database and import the customer's database to HDFS. Let's go ahead and run our first import command:

```
sqoop import --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table departments;
```

Output

```
Launched map tasks=4
Other local map tasks=4
Total time spent by all maps in occupied slots (ms)=59827
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=59827
Total vcore-milliseconds taken by all map tasks=59827
Total megabyte-milliseconds taken by all map tasks=61262848
Map-Reduce Framework
  Map input records=6
  Map output records=6
  Input split bytes=481
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=1237
  CPU time spent (ms)=2110
  Physical memory (bytes) snapshot=468836352
  Virtual memory (bytes) snapshot=6040522752
  Total committed heap usage (bytes)=243007488
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=60
22/10/10 21:49:52 INFO mapreduce.ImportJobBase: Transferred 60 bytes in 31.2524 seconds (1.9199 bytes/sec)
22/10/10 21:49:52 INFO mapreduce.ImportJobBase: Retrieved 6 records.
[cloudera@quickstart ~]$
```

7. Check whether data is imported into HDFS or not

```
hadoop fs -ls departments
```

Output

```
[cloudera@quickstart ~]$ hadoop fs -ls departments
Found 5 items
-rw-r--r-- 1 cloudera cloudera 0 2022-10-10 21:49 departments/_SUCCESS
-rw-r--r-- 1 cloudera cloudera 21 2022-10-10 21:49 departments/part-m-00000
-rw-r--r-- 1 cloudera cloudera 10 2022-10-10 21:49 departments/part-m-00001
-rw-r--r-- 1 cloudera cloudera 7 2022-10-10 21:49 departments/part-m-00002
-rw-r--r-- 1 cloudera cloudera 22 2022-10-10 21:49 departments/part-m-00003
[cloudera@quickstart ~]$
```

8. Imports the departments data to a target directory:

```
scoop import --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --
username root --table departments --target-dir /user/cloudera/dept1;
```

9. Print imported data

```
hadoop fs -cat /user/cloudera/dept1/part*
```

Output

```
2,Fitness
3,Footwear
4,Apparel
5,Golf
6,Outdoors
7,Fan Shop
```

10. Export Data from HDFS to MySQL

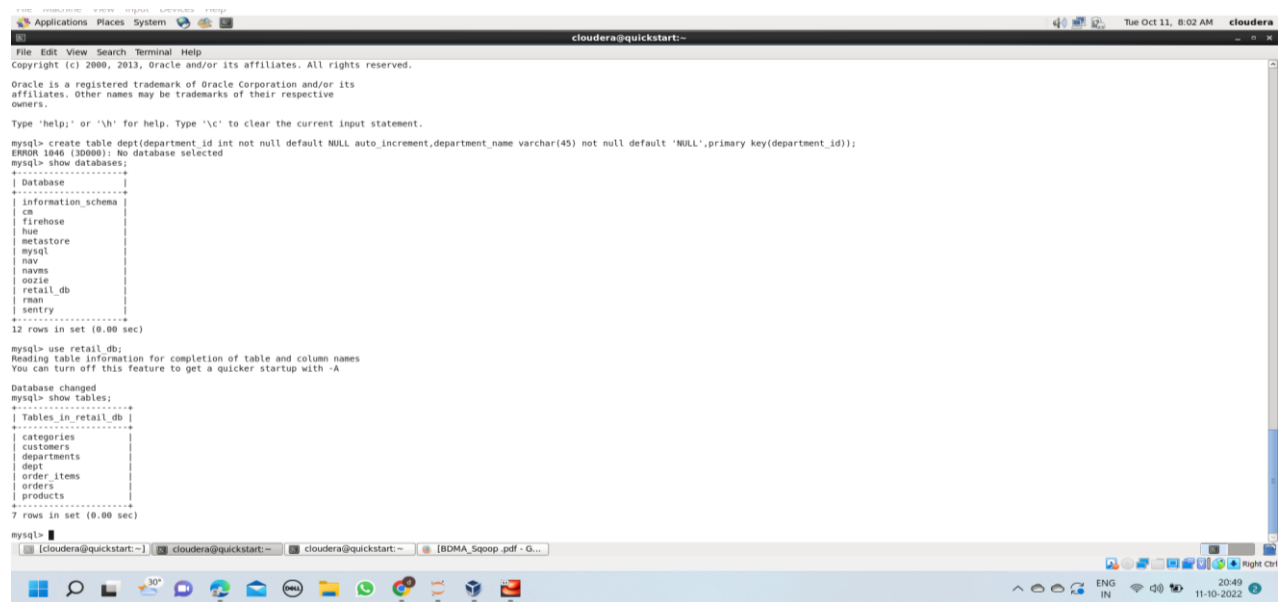
- a. Create following table in MySQL using following scheme

```
department_id    int
department_name  varchar(45)
```

create table dept(department_id int not null default NULL auto_increment, department_name varchar(45) not null default 'NULL', primary key(department_id));

```
mysql>
mysql>
mysql>
mysql>
mysql> create table dept(department_id int not null default NULL auto_increment,department_name varchar(45) not null default 'NULL',primary key(department_id)
Query OK, 0 rows affected (0.04 sec)
```

Big Data Management and Analytics Lab



```
mysql> create table dept(department_id int not null default NULL auto_increment,department_name varchar(45) not null default 'NULL',primary key(department_id));
ERROR 1046 (30000): No database selected
mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| cs |
| firehose |
| hue |
| metastore |
| mysql |
| nav |
| names |
| oozie |
| retail_db |
| rman |
| sentry |
+-----+
12 rows in set (0.00 sec)

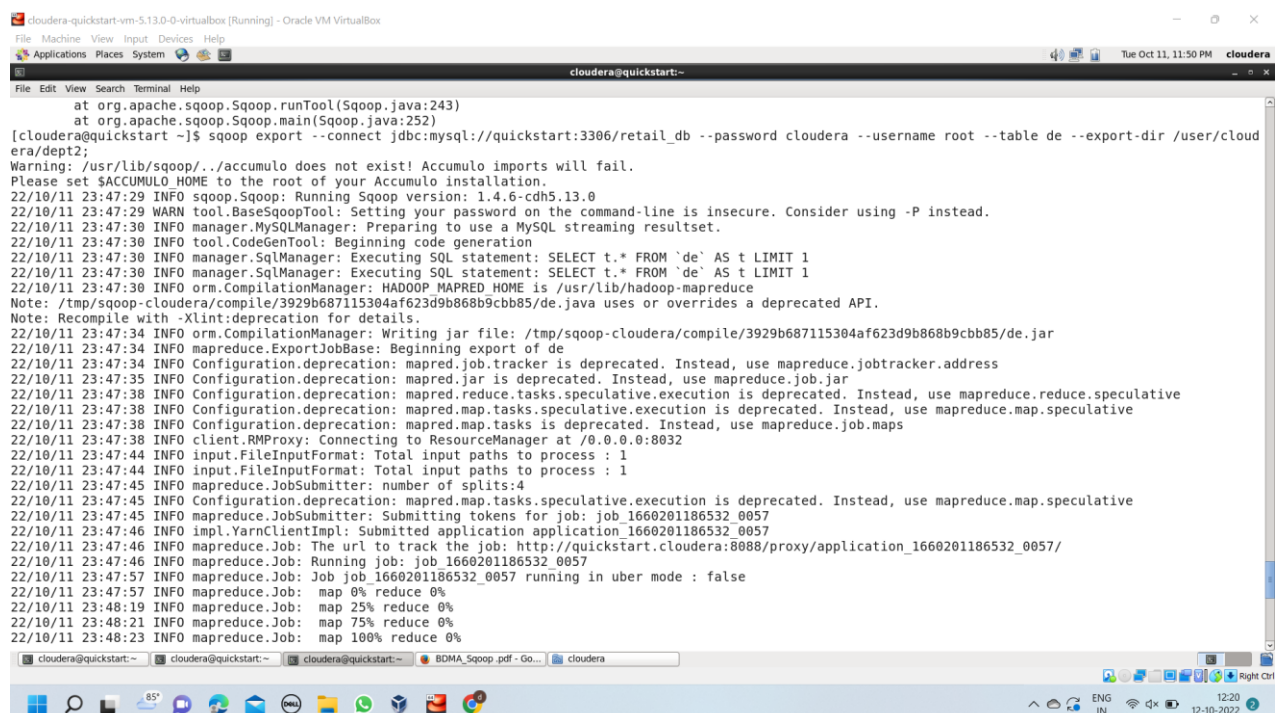
mysql> use retail_db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

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

mysql>
```

b. Export table from HDFS to MySQL

```
sqoop export --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table dept --export-dir /user/cloudera/dept2;
```



```
at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)
at org.apache.sqoop.Sqoop.main(Sqoop.java:252)
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table dept --export-dir /user/cloudera/dept2;
Warning: /usr/lib/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/10/11 23:47:29 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/10/11 23:47:29 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/10/11 23:47:30 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/10/11 23:47:30 INFO tool.CodeGenTool: Beginning code generation
22/10/11 23:47:30 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `dept` AS t LIMIT 1
22/10/11 23:47:30 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `dept` AS t LIMIT 1
22/10/11 23:47:30 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/3929b687115304af623d9b868b9cbb85/de.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/10/11 23:47:34 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/3929b687115304af623d9b868b9cbb85/de.jar
22/10/11 23:47:34 INFO mapreduce.ExportJobBase: Beginning export of dept
22/10/11 23:47:34 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
22/10/11 23:47:35 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/10/11 23:47:38 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
22/10/11 23:47:38 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
22/10/11 23:47:38 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/10/11 23:47:38 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/10/11 23:47:44 INFO input.FileInputFormat: Total input paths to process : 1
22/10/11 23:47:44 INFO input.FileInputFormat: Total input paths to process : 1
22/10/11 23:47:45 INFO mapreduce.JobSubmitter: number of splits:4
22/10/11 23:47:45 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
22/10/11 23:47:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1660201186532_0057
22/10/11 23:47:46 INFO impl.YarnClientImpl: Submitted application application_1660201186532_0057
22/10/11 23:47:46 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1660201186532_0057/
22/10/11 23:47:46 INFO mapreduce.Job: Running job: job_1660201186532_0057
22/10/11 23:47:57 INFO mapreduce.Job: Job job_1660201186532_0057 running in uber mode : false
22/10/11 23:47:57 INFO mapreduce.Job: map 0% reduce 0%
22/10/11 23:48:19 INFO mapreduce.Job: map 25% reduce 0%
22/10/11 23:48:21 INFO mapreduce.Job: map 75% reduce 0%
22/10/11 23:48:23 INFO mapreduce.Job: map 100% reduce 0%
```

Big Data Management and Analytics Lab

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
FILE: Number of bytes written=683596
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=622
HDFS: Number of bytes written=0
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters
  Launched map tasks=4
  Data-local map tasks=4
  Total time spent by all maps in occupied slots (ms)=79849
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=79849
  Total vcore-milliseconds taken by all map tasks=79849
  Total megabyte-milliseconds taken by all map tasks=81765376
Map-Reduce Framework
  Map input records=1
  Map output records=1
  Input split bytes=580
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=1069
  CPU time spent (ms)=5880
  Physical memory (bytes) snapshot=697503744
  Virtual memory (bytes) snapshot=6274482176
  Total committed heap usage (bytes)=695730176
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
22/10/11 23:48:24 INFO mapreduce.ExportJobBase: Transferred 622 bytes in 46.7273 seconds (13.3113 bytes/sec)
22/10/11 23:48:24 INFO mapreduce.ExportJobBase: Exported 1 records.
[cloudera@quickstart ~]$
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
mysql> insert into dept(department_id,department_name) values(9125,'Nambi');
Query OK, 1 row affected (0.00 sec)

mysql> insert into dept(department_id,department_name) values(9130,'rajesh');
Query OK, 1 row affected (0.01 sec)

mysql> select * from dept;
+-----+
| department_id | department_name |
+-----+
| 9104 | arunkumar |
| 9125 | Nambi |
| 9130 | rajesh |
+-----+
3 rows in set (0.00 sec)

mysql> create table de(department_id int not null default NULL auto_increment,department_name varchar(45) not null default 'NULL',primary key(department_id));
Query OK, 0 rows affected (0.06 sec)

mysql> show tables;
+-----+
| Tables_in_retail_db |
+-----+
| categories |
| customers |
| de |
| departments |
| dept |
| order_items |
| orders |
| products |
+-----+
8 rows in set (0.00 sec)

mysql> select * from de;
+-----+
| department_id | department_name |
+-----+
| 9130 | rajesh |
+-----+
1 row in set (0.02 sec)

mysql>
```

Notes