

Session 6 - DATA INGESTION TOOL SQOOP

Assignment 1

In the demo of Sqoop import tool, the table 'Person' was imported into HDFS folder of 'sqoopout'.

To export the imported data back to mysql, 2 tables namely Person1 and Person2 are created with structure similar to Person table.

```
mysql> select * from Person1;
Empty set (0.08 sec)

mysql> select * from Person2;
Empty set (0.01 sec)

mysql> exit;
Bye
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hdfs dfs -ls /user/acadgild/sqoopout
18/04/21 01:54:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r-- 1 acadgild supergroup 0 2018-04-21 01:52 /user/acadgild/sqoopout/.SUCCESS
-rw-r--r-- 1 acadgild supergroup 150 2018-04-21 01:52 /user/acadgild/sqoopout/part-m-00000
```

Both 'Person1' and 'Person2' are empty tables.

Sqoopout folder in HDFS contains a part file that was imported from 'Person' table.

Task1 : Use Sqoop tool to export data present in SQOOPOUT folder made while demo of Import table.

Command :

```
sqoop export --connect jdbc:mysql://localhost/simplidb --table Person1 --export-dir /user/acadgild/sqoopout --username root -P
```

Parameters :

--table -> target table

--export-dir -> hdfs location of the directory from which data is to be exported.

```
acadgild@localhost:~$ sqoop export --connect jdbc:mysql://localhost/simpladb --table Person1 --export-dir /user/acadgild/sqoopout --username root -p
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/04/21 01:56:45 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
18/04/21 01:56:51 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/04/21 01:56:51 INFO Tool.CodeGenTool: Beginning code generation
Sat Apr 21 01:56:52 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
18/04/21 01:56:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Person1` AS t LIMIT 1
18/04/21 01:56:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Person1` AS t LIMIT 1
```

Since -p was specified in the command, the password is asked for explicitly.

```
18/04/21 01:57:54 INFO mapreduce.ExportJobBase: Transferred 1.0146 KB in 54.1348 seconds (19.1928 bytes/sec)
18/04/21 01:57:54 INFO mapreduce.ExportJobBase: Exported 5 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ mysql -u root -p
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 45
Server version: 8.0.3-rc-log MySQL Community Server (GPL)

Copyright (c) 2000, 2017, 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> use simpladb;
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> select * from Person1;
+-----+-----+-----+-----+-----+
| person_id | lname | fname | area | city |
+-----+-----+-----+-----+-----+
| 1 | Shyam | Ram | Patna | Bihar |
| 2 | Tanya | Priya | Whitefield | Bangalore |
| 3 | James | Brown | New York | United States |
| 4 | Jhon | Miller | Los Angeles | United States |
| 789 | a | b | c | d |
+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)

mysql>
```

The mapreduce job is successfully completed. Once logged into mysql, we can see that the table Person1 is populated with the data from sqoopout folder.

Task 2 : Use Sqoop tool to export data present in SQOOPOUT folder made while demo of Import table with parameter person_id =3.

- For selective/conditional export of data from HDFS to mysql, --query parameter is not supported with Sqoop Export tool.
- So I created a mysql Stored procedure which will be called from the sqoop command. The procedure will perform the condition checking. Below is the procedure.

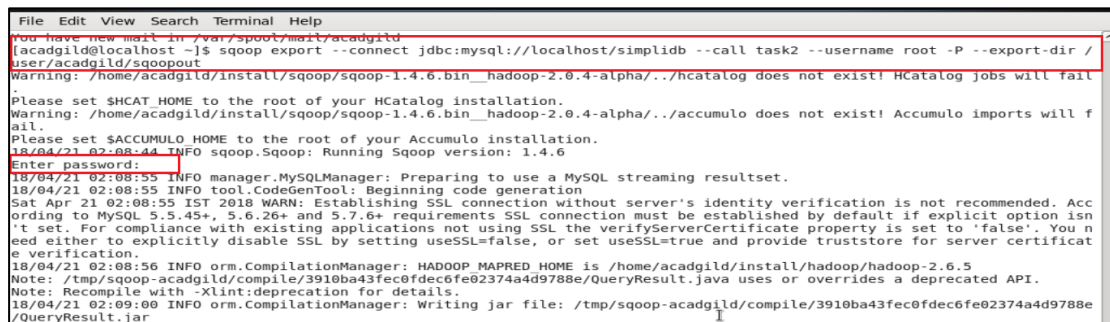
```
mysql>
delimiter //
mysql> create procedure task2(IN in_person_id int(11),in_lname varchar(20),in_fname varchar(20),in_area varchar(20),in_city
varchar(20))
-> begin
-> if in_person_id = 3 then
-> insert into Person2(person_id,lname,fname,area,city) values (in_person_id,in_lname,in_fname,in_area,in_city);
-> end if;
-> end //
Query OK, 0 rows affected (0.01 sec)
```

Command :

```
sqoop export --connect jdbc:mysql://localhost/simplidb --call task2 --username root
-P --export-dir /user/acadgild/sqoopout
```

Parameters :

--call -> instead of specifying the target table, a mysql stored procedure can be called.



```
File Edit View Search Terminal Help
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ sqoop export --connect jdbc:mysql://localhost/simplidb --call task2 --username root -P --export-dir /
user/acadgild/sqoopout
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./hcatalog does not exist! HCatalog jobs will fail
.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./accumulo does not exist! Accumulo imports will f
ail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/04/21 02:08:44 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
18/04/21 02:08:55 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/04/21 02:08:55 INFO tool.CodeGenTool: Beginning code generation
Sat Apr 21 02:08:55 IST 2018 WARN: Establishing SSL connection without server's identity verification is not recommended. Acc
ording to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn
't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You n
eed either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificat
e verification.
18/04/21 02:08:56 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/install/hadoop/hadoop-2.6.5
Note: /tmp/sqoop-acadgild/compile/3910ba43fec0fdec6fe02374a4d9788e/QueryResult.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/04/21 02:09:00 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/3910ba43fec0fdec6fe02374a4d9788e
/QueryResult.jar
```

Since -p was specified in the command, the password is asked for explicitly.

```
Applications Places System acadgild@localhost:~
File Edit View Search Terminal Help
Virtual memory (bytes) snapshot=8250687488
Total committed heap usage (bytes)=195035136
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
18/04/21 02:09:55 INFO mapreduce.ExportJobBase: Transferred 1.0146 KB in 52.1512 seconds (19.9228 bytes/sec)
18/04/21 02:09:55 INFO mapreduce.ExportJobBase: Exported 5 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 67
Server version: 8.0.3-rc-log MySQL Community Server (GPL)

Copyright (c) 2000, 2017, 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> select * from Person2;
ERROR 1046 (3D000): No database selected
mysql> use simplldb;
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> select * from Person2;
+-----+-----+-----+-----+-----+
| person_id | lname | fname | area | city |
+-----+-----+-----+-----+-----+
| 3 | James | Brown | New York | United States |
+-----+-----+-----+-----+-----+
1 row in set (0.00 sec)

mysql>
```

The mapreduce job is successfully completed. Once logged into mysql, we can see that the table Person2 is populated with only one record matching to the condition that is person_id=3.