Big Data And Hadoop

Session 20 – Assignment 1

Problem Statement:

Perform UPSERT in Sqoop export.

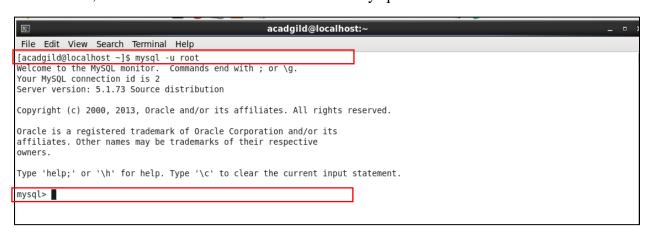
Read a file from HDFS and based on the id field, perform UPSERT in MySQL table. In UPSERT, if the field exists, then it is updated else it is inserted.

Solution:

1. We first start mysql service using the command **sudo service mysqld start** as follows:



2. Next, we start the command line interface for mysql as follows:



3. As we have to place the data from HDFS to MySQL, we should first have the table created already in MySQL. We will create the table in a database named **db1**. So we will first create the database using the command **create database db1**; as follows:



4. We specify which database to work in using the command use db1;



5. We now create the table named **employee** using the command

Here, we have made **id** to be the primary key of the table.



6. We can see the schema of the table created using **describe** command as follows:



7. We have the input file present locally at /home/acadgild/Abhilasha. Name of the file is employee.txt. It can be list using ls command as follows:



8. We put this file on HDFS at /abhilasha/sqoop using the put command as follows:



9. Listing of files at HDFS shows **employee.txt** placed successfully.



10. The content of the file is shown using **cat** command as follows:

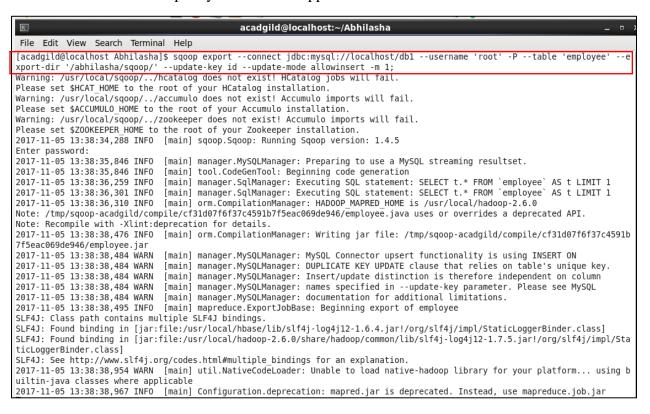


The file has three records.

- 11. Now we perform the sqoop operation to put data from HDFS into MySQL as follows: The input parameters are:
 - a. Connect: This specifies the mysql system to be connected to. Here its value is jdbc:mysql//localhost/db1

Here, **db1** is name of the database to be used.

- b. Username: This is to specify the username to be used to access MySQL. Here its value is **root.**
- c. P: This is for password. Using P doesn't need to type the password in the sqoop command, instead it takes it from standard input stream.
- d. Table: This is to specify the table name, our input is **employee**.
- e. Export-dir: This is the path of export directory in HDFS. Its value is /abhilasha/sqoop/
- f. Update-key: This is the column name that will be used for update. Our column for update is **id**.
- g. Update-mode: We need the command to perform UPSERT that is if a new record comes, if should update the previous one else it should insert the record. Hence, we have used the mode to be **allowinsert**.
- h. M: This is to specify number of mappers to use. We have used 1 in our case.



12. The output of this command execution results in the **employee** table getting populated in MySQL as follows:



13. Now to verify that the UPSERT works, we will modify the input file locally and place it on HDFS and re-run the sqoop command. The updated input file on the local file system is as follows:

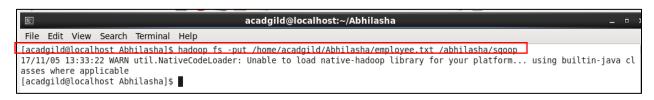


Please note that here, we have modified the salary of employee with **id 2.** Earlier, its salary was 20000, now we have made it to 50000. Also we have added a new entry for employee id **4.**

14. Before we place this file on HDFS, we remove the old file using **rm** command as follows:



15. Now, we place the updated file on HDFS as follows:



16. The content of file on HDFS is seen using **cat** command as follows:

```
acadgild@localhost:~/Abhilasha

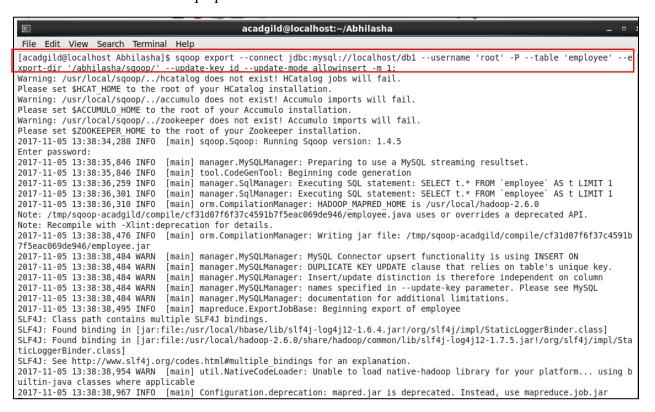
File Edit View Search Terminal Help

[acadgild@localhost Abhilasha]$ hadoop fs -cat /abhilasha/sqoop/employee.txt

17/11/05 13:33:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
1,ABCD,10000

2,EFGH,50000
3,IGKL,30000
4,MNOP,40000
[acadgild@localhost Abhilasha]$
```

17. We now re-run the sqoop command as follows:



18. On running the query **select * from employee** in MySQL, we see the data below.



Here, we observe that the records with ids ${\bf 1}$ and ${\bf 3}$ have not been re-inserted.

The record with id **2** has its salary updated, hence, update operation is performed.

Record with id **4** is a new record, was not present earlier, and now has been inserted.

This shows that UPSERT is performed.