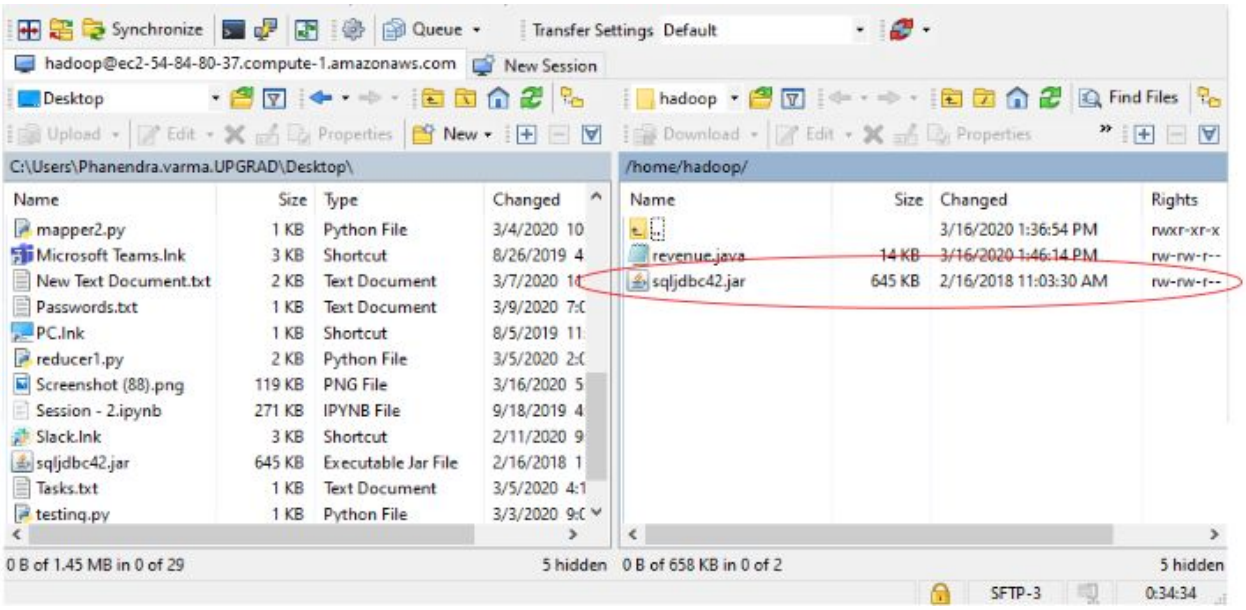
Things to ensure before running the sqoop job

1. Make sure the EMR version you are running is 5.24.0. The latest version of EMR isn't stable and creates problems while creating a workflow.
2. An RDS instance with necessary data loaded on the tables
3. Sample sqoop job running fine through CLI
4. Download the jdbc driver using the wget command and the link given below

   OR

   Download it locally and copy it to your EMR cluster using the WINSCP tool - https://www.microsoft.com/en-us/download/details.aspx?id=54671



5. Sample sqoop job running fine on your cluster

```
sqoop import --connect
jdbc:mysql://sample-database.cqsesz6h9yjg.us-east-1.rds.amazonaws.com:3306/
telco --table revenue --target-dir /user/hadoop/telco/revenue/ --username
admin -P -m 1
```

sample-database.cqsesz6h9yjg.us-east-1.rds.amazonaws.com - Replace this with the appropriate RDS instance address

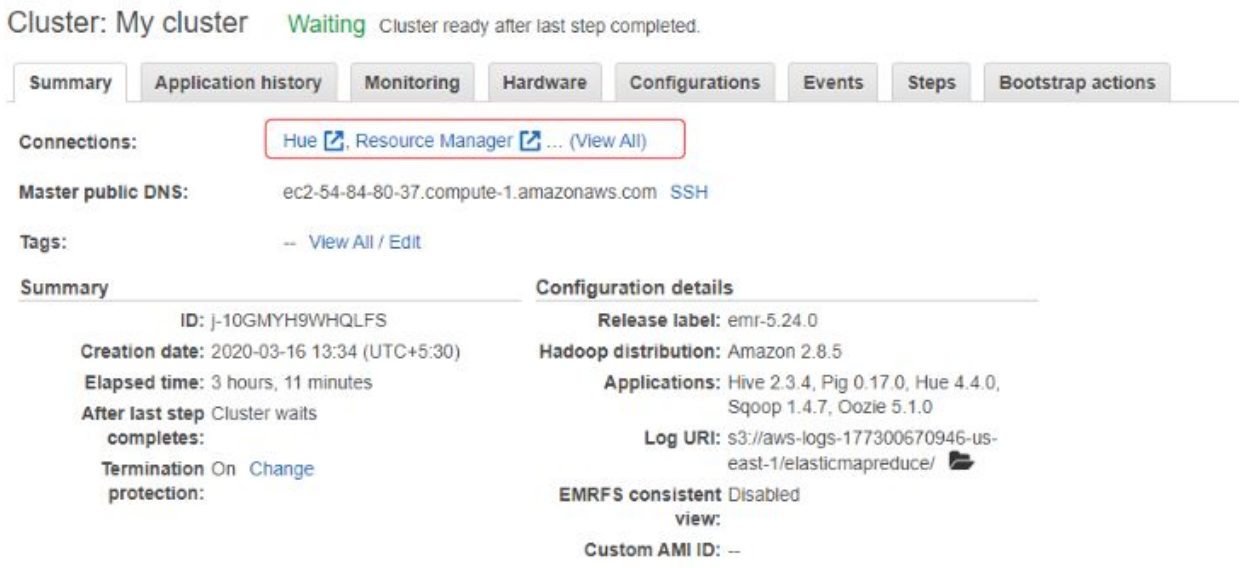
telco - Name of the database

revenue - Name of the table

/user/hadoop/telco/revenue/ - Target directory

Now, steps to add a sqoop job in a oozie workflow

1. Launch the hue service by clicking on hue



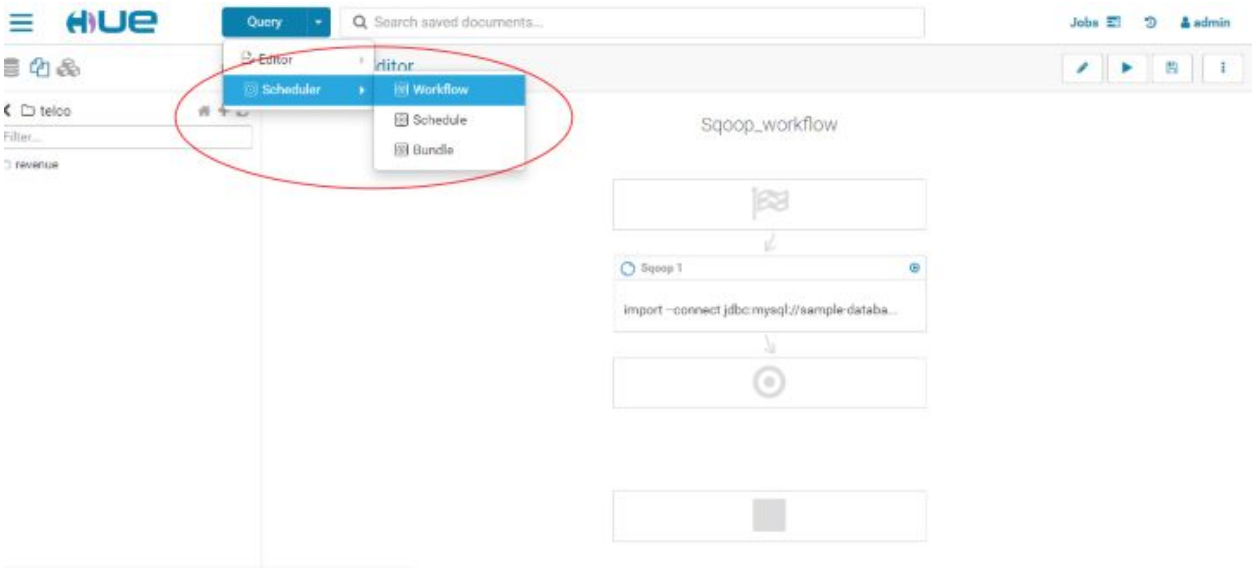
2. This will redirect to a new landing page where you are supposed to create a hue user.

3. On creating the hue user, the next step is to copy the driver to this user. This is to make sure that the jdbc driver is available with the hue user to run the sqoop job.

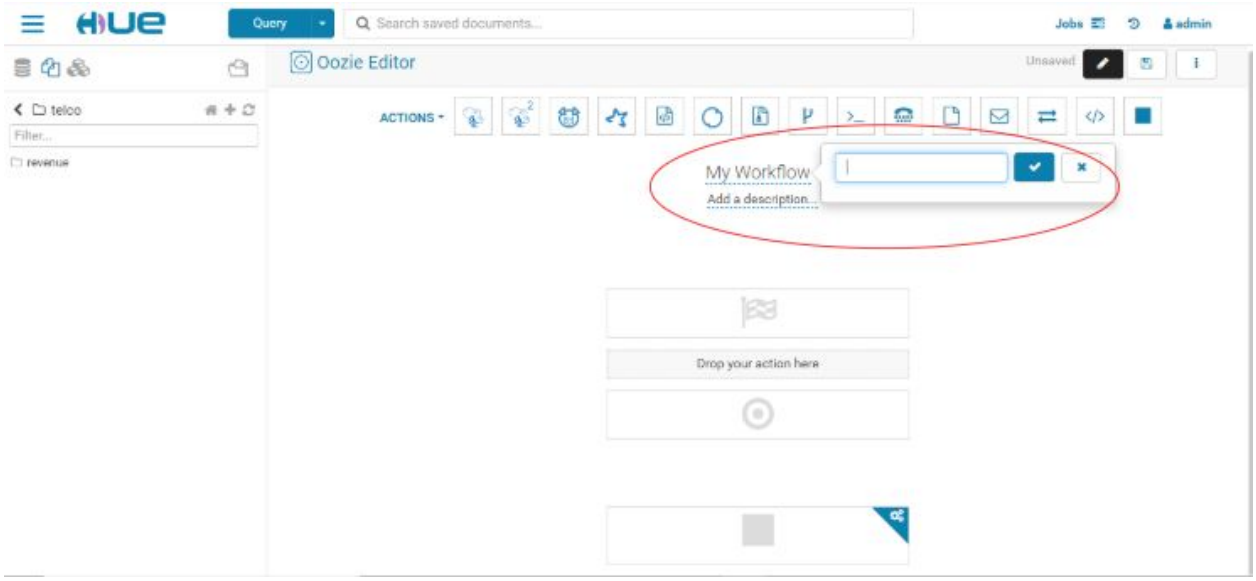
On the CLI run the following commands to copy the file to the hue user.

hadoop fs -put sqljdbc42.jar /user/admin/

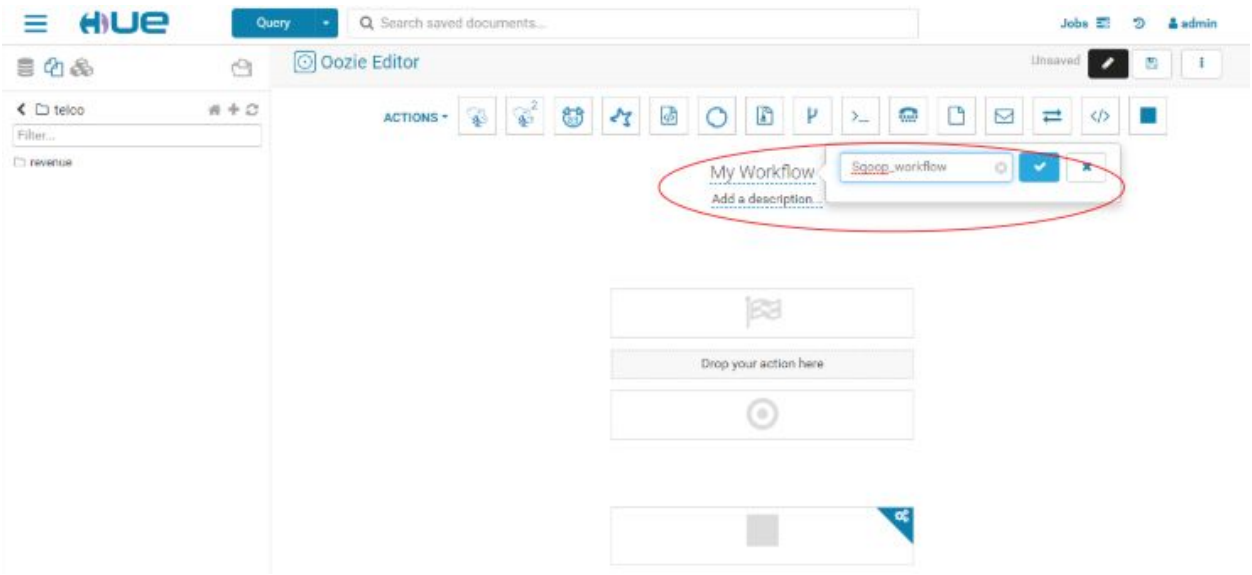
4. Now that you have the driver with the hue user the next step is to create a workflow. For this click on **schedule** and select **workflow**.



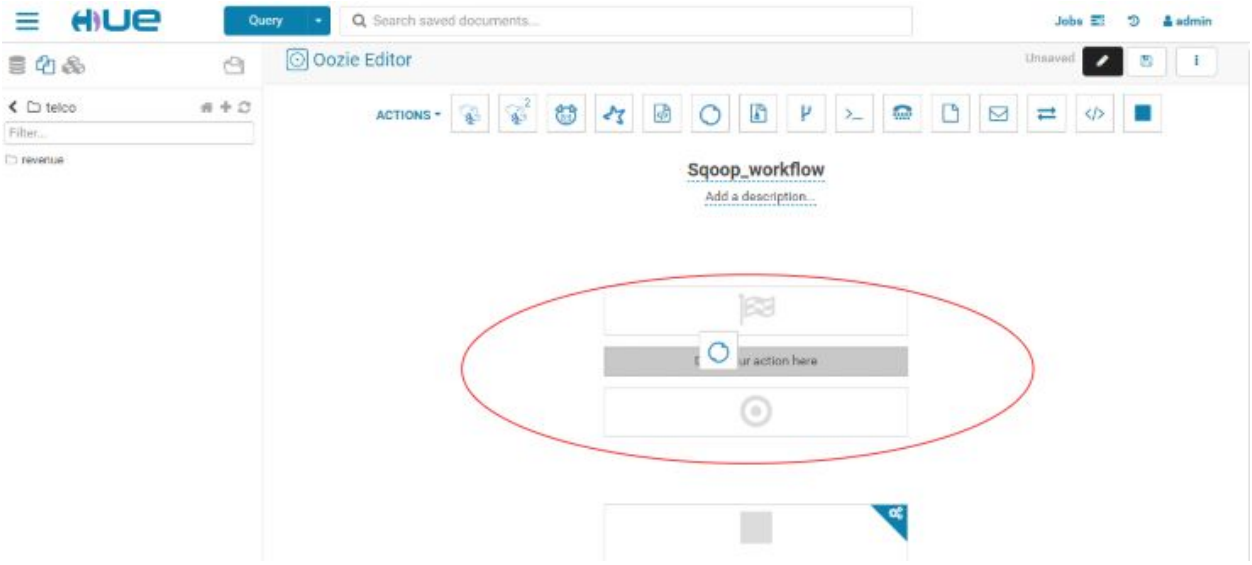
5. Give the name to the workflow as shown in the image below



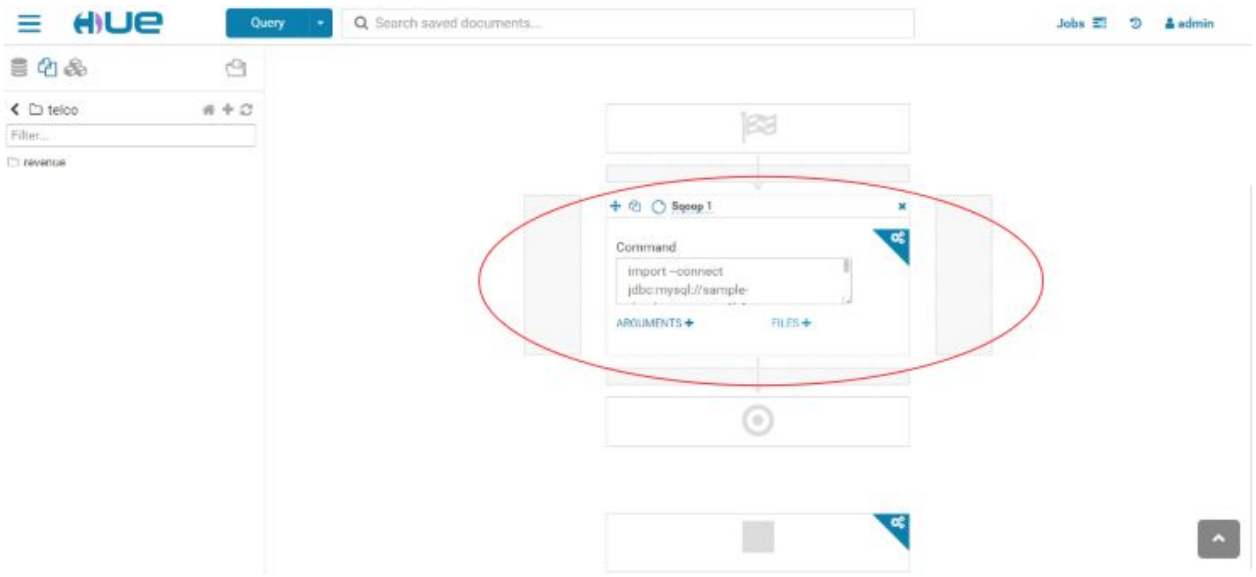
For example you can name it as **sqoop_workflow**



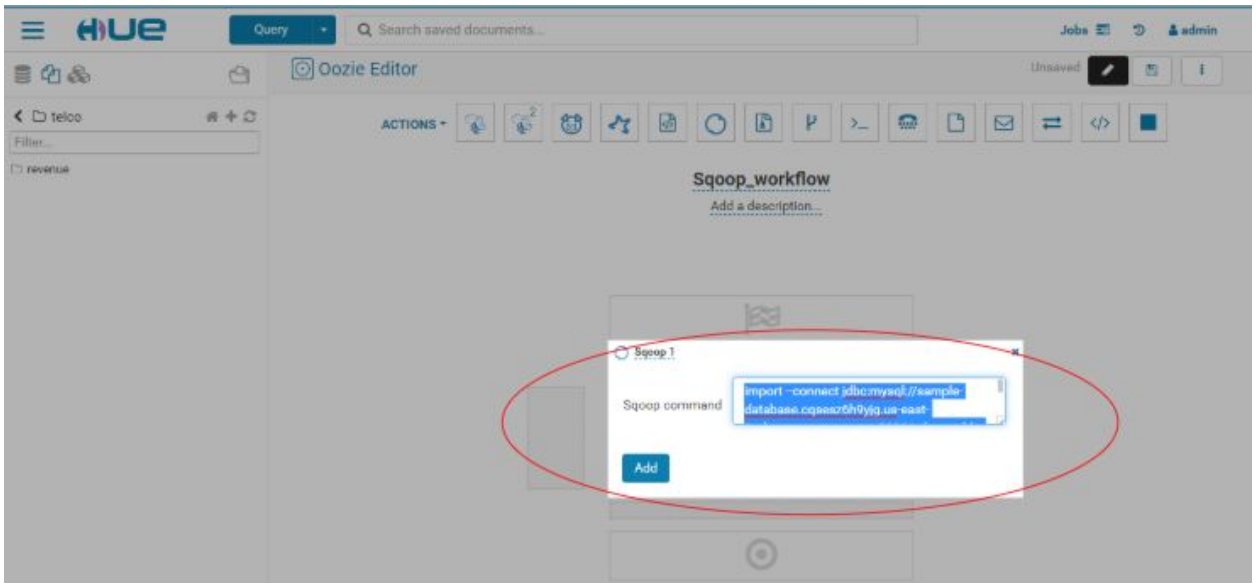
6. Next based on the desired action drag and drop the desired service. In this case drop the sqoop service as shown in the figure



Dropping the sqoop service enables you to enter the necessary command



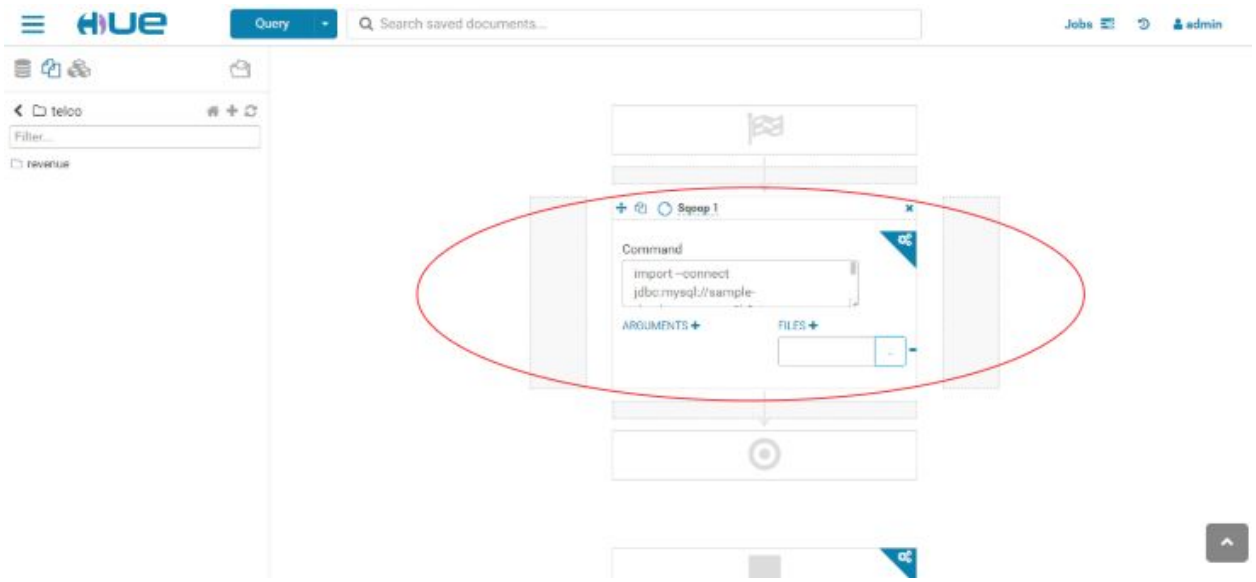
7. Under the command field give the following command

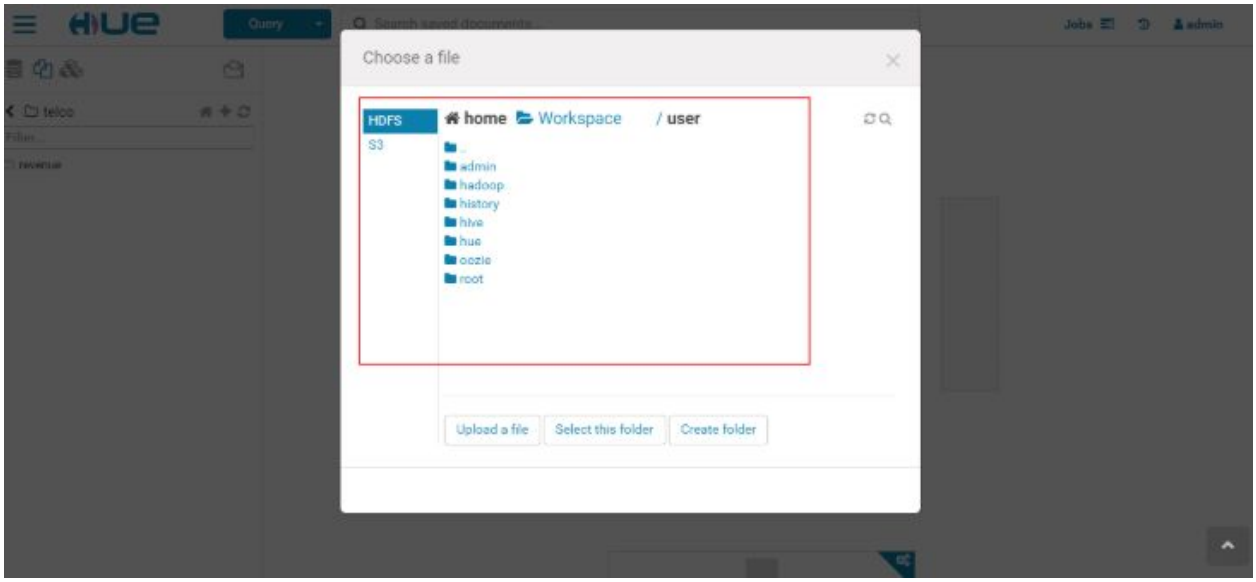


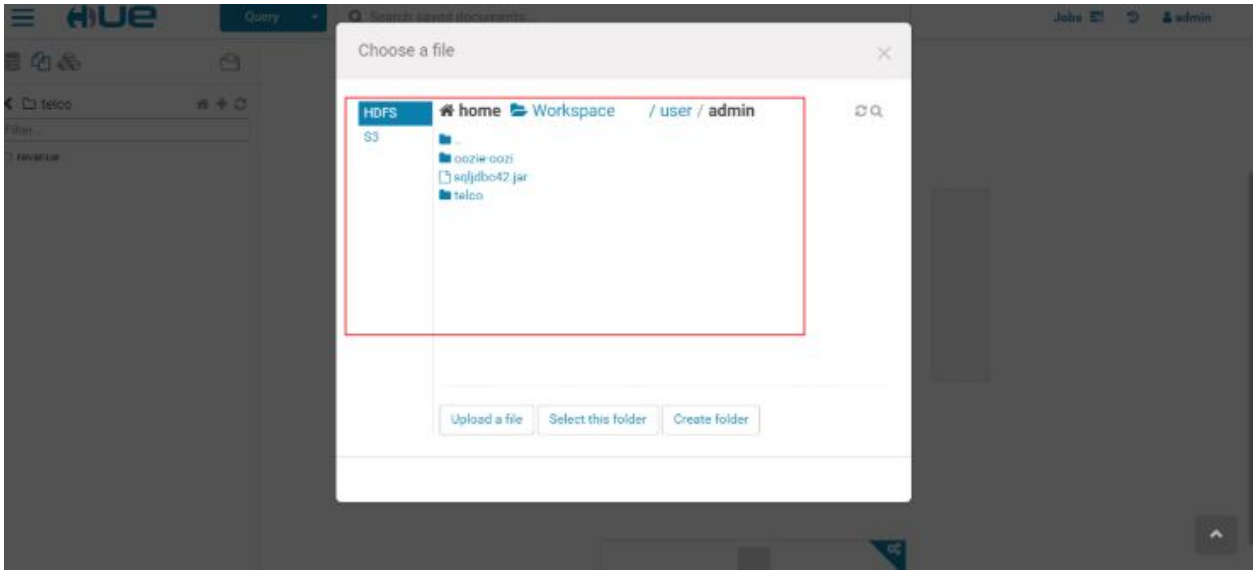
import --connect jdbc:mysql://**sample-database.cqsesz6h9yjg.us-east-1.rds.amazonaws.com**:3306/telco --table **revenue** --target-dir **/user/admin/telco/revenue/** --username **admin** --password **admin123** --m 1 --driver **com.microsoft.sqlserver.jdbc.SQLServerDrive**r

The username and password here are that of the RDS instance and not that of the hue service

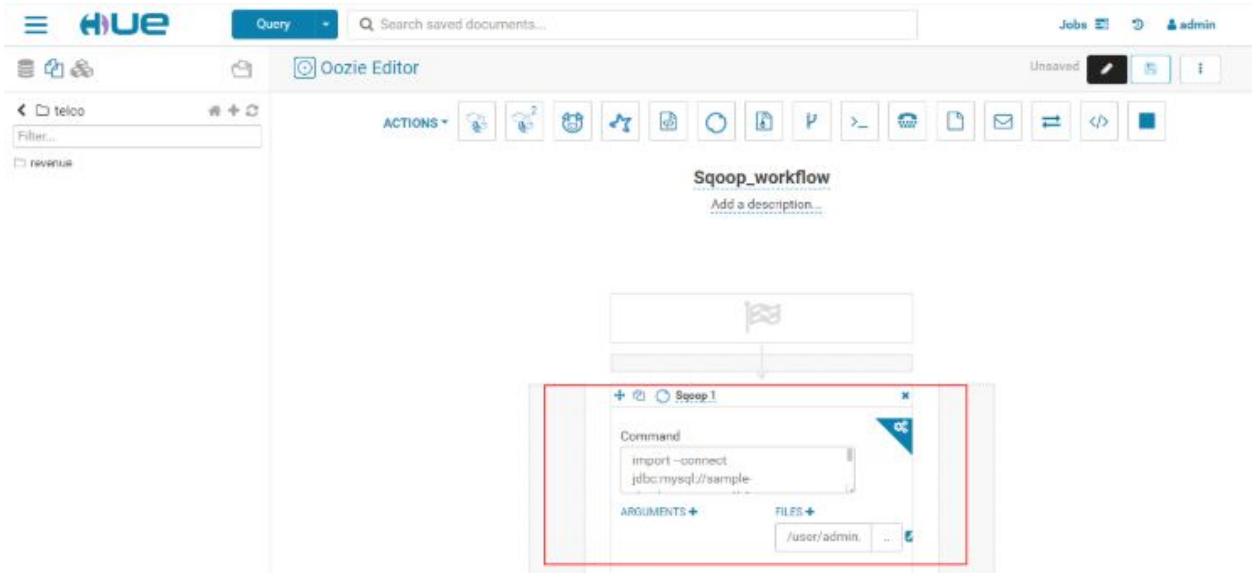
8. In the command section you can find the option to choose the necessary files while running the sqoop job.



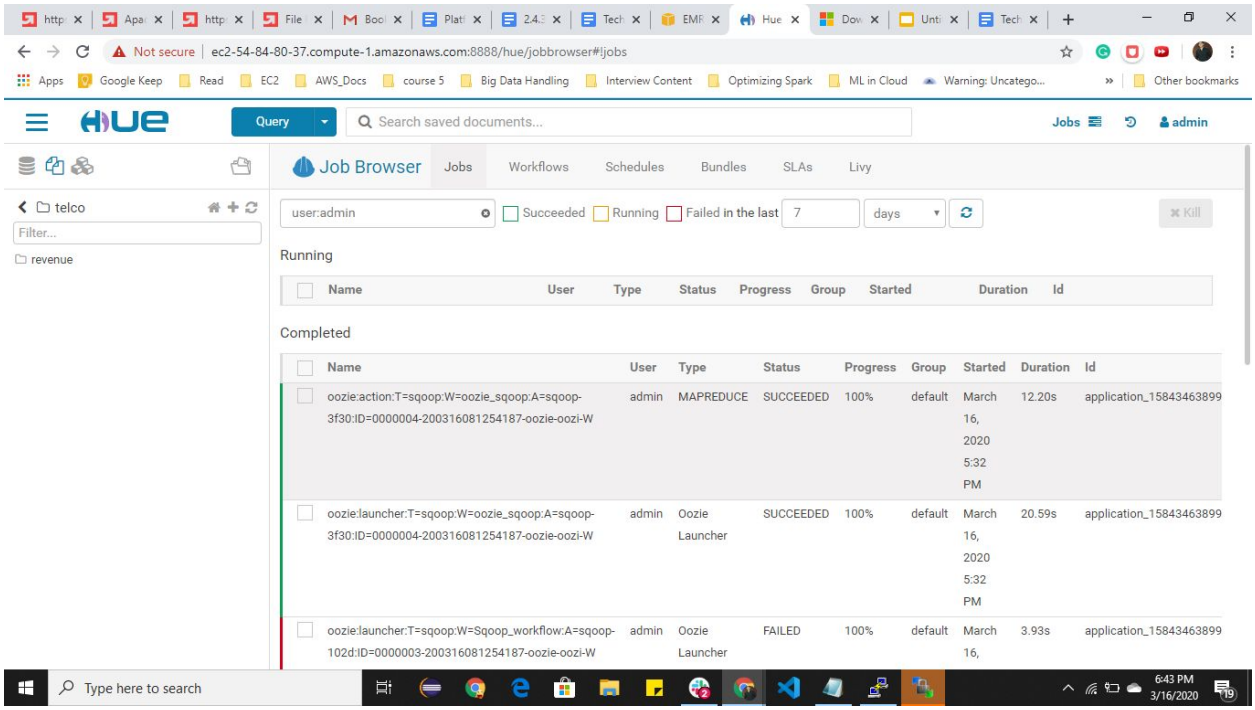


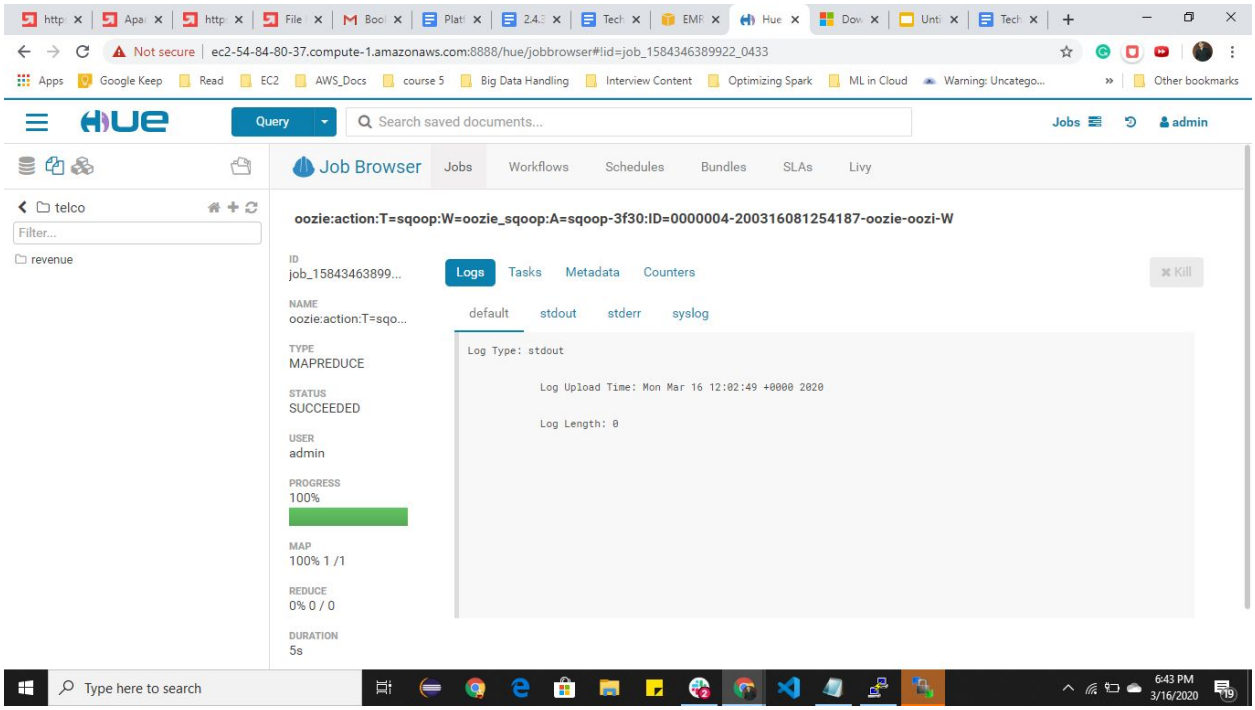


9. Finally after selecting the file save and run the job



10. Once you are job is submitted you can see the status under the **job browser section** and by clicking on the job you can see the details of the different tasks involved in the job.





11. After successful execution of the sqoop job, you can see the files created at the destination specified