SqoopOperator Demonstration

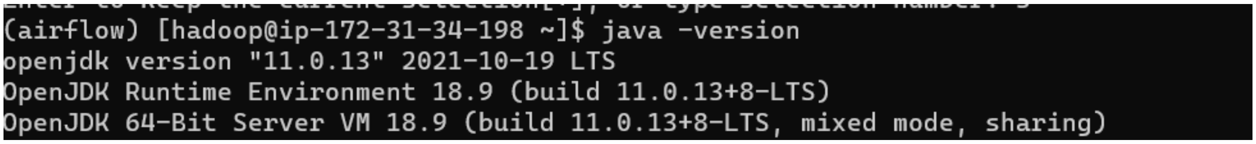
This document will guide you through the demonstration of the SqoopOperator in Session 2 of the Airflow module.

Prerequisites:

- transactions.dump (contains SQL queries that will create tables and dump the data into our local MySQL)
- sample_sqoop.py(the code explained in the video)
- Make sure that the current **Java version is Java 11**(11.x.x). **Note**: This is only needed to run the Sqoop operator and you will need to switch back to Java 8 after you are done with the Sqoop operator. The steps to switch back to Java 8 can be found at the end of this document.

You can check this by running the following command:

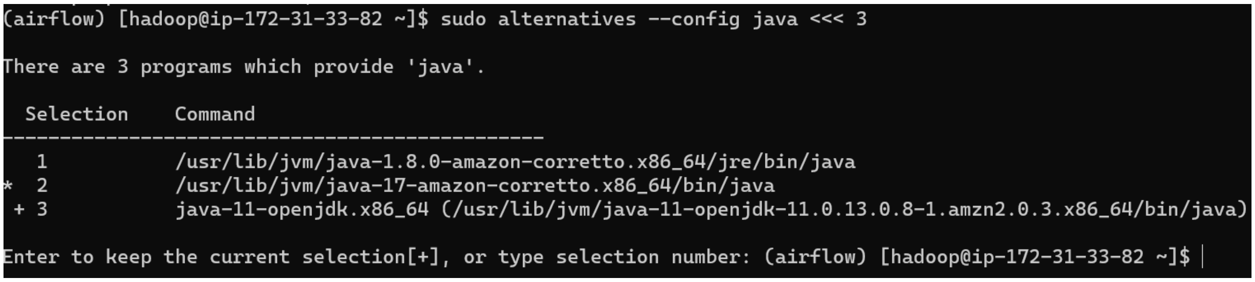
```
java -version
```



```
(airflow) [hadoop@ip-172-31-34-198 ~]$ java -version
openjdk version "11.0.13" 2021-10-19 LTS
OpenJDK Runtime Environment 18.9 (build 11.0.13+8-LTS)
OpenJDK 64-Bit Server VM 18.9 (build 11.0.13+8-LTS, mixed mode, sharing)
```

If you still have Java 8, then you need to switch to Java 11 by running the following command.

```
sudo alternatives --config java <<< 3
```



```
(airflow) [hadoop@ip-172-31-33-82 ~]$ sudo alternatives --config java <<< 3

There are 3 programs which provide 'java'.

  Selection    Command
-----------------------------------------------
   1           /usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre/bin/java
*  2           /usr/lib/jvm/java-17-amazon-corretto.x86_64/bin/java
 + 3           java-11-openjdk.x86_64 (/usr/lib/jvm/java-11-openjdk-11.0.13.0.8-1.amzn2.0.3.x86_64/bin/java)

Enter to keep the current selection[+], or type selection number: (airflow) [hadoop@ip-172-31-33-82 ~]$ |
```

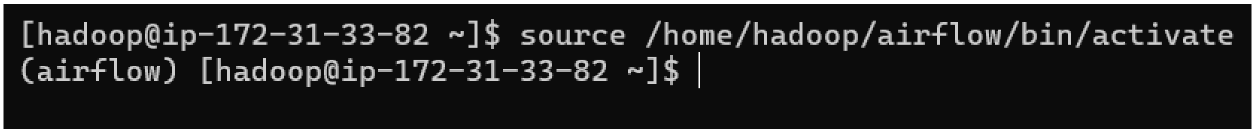
What are we doing?

In this demonstration, we need to create a DAG with one task. This will be a Sqoop task that will connect to the local MySQL and bring data to HDFS.

Please follow the instructions below:

1. Login to your EMR instance.
2. Activate the Python virtual environment using the following command:

   source /home/hadoop/airflow/bin/activate

   ```
   [hadoop@ip-172-31-33-82 ~]$ source /home/hadoop/airflow/bin/activate
   (airflow) [hadoop@ip-172-31-33-82 ~]$
   ```

3. We need to make sure that the output HDFS directory doesn't exist and if it does you can remove it using the following command:

   sudo hdfs dfs -rm -r -skipTrash hdfs:///data/credit_card/transactions

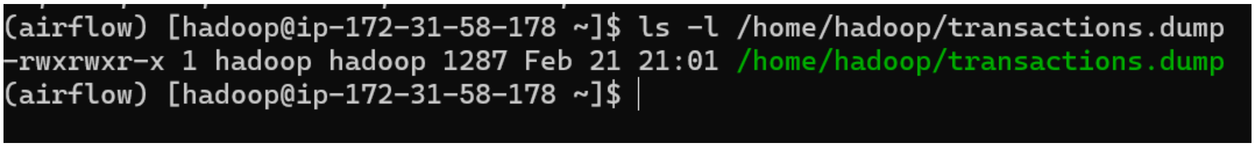
   Your sqoop task will fail if the target_dir is in use, which is why *you'll have to run this command if you choose re-run this task/DAG*

4. Now you need to load the data into your local MySQL, firstly you need to place the transactions.dump file in some location in your EMR machine. We will store in the /home/hadoop/ directory

   (You can use WinSCP or create a new file called transactions.dump in the /home/hadoop directory and paste the contents in that file)

5. Run the following command to make sure that the file is in the correct directory.

   ls -l /home/hadoop/transactions.dump

   ```
   (airflow) [hadoop@ip-172-31-58-178 ~]$ ls -l /home/hadoop/transactions.dump
   -rwxrwxr-x 1 hadoop hadoop 1287 Feb 21 21:01 /home/hadoop/transactions.dump
   (airflow) [hadoop@ip-172-31-58-178 ~]$
   ```

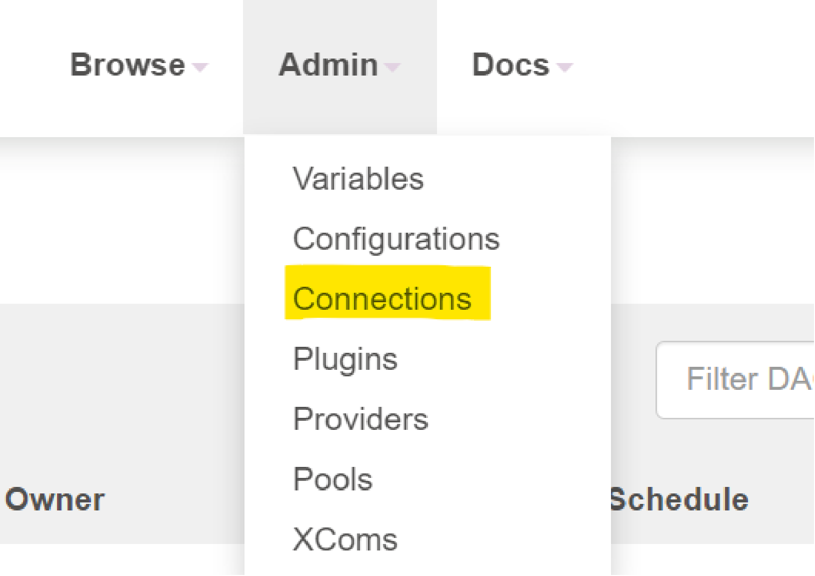
6. Now run the following command to execute the SQL commands in /tmp/transactions.dump and create our tables :

   mysql -u root -p123 < /home/hadoop/transactions.dump

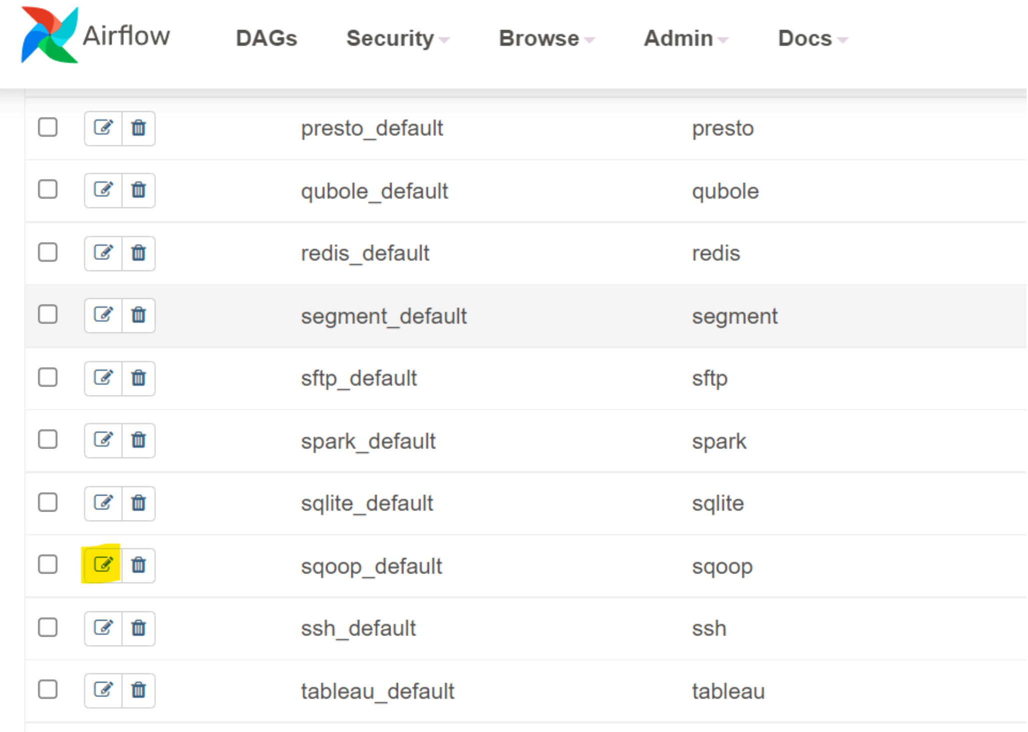
7. Next, you need to set up the Sqoop connection from the Airflow UI which is hosted in the URL : **your_public_dns:8082**

   Note: You can find you_publin_ip in your AWS EMR dashboard (IPv4 Public DNS))

Go to the Admin tab and click on Connections

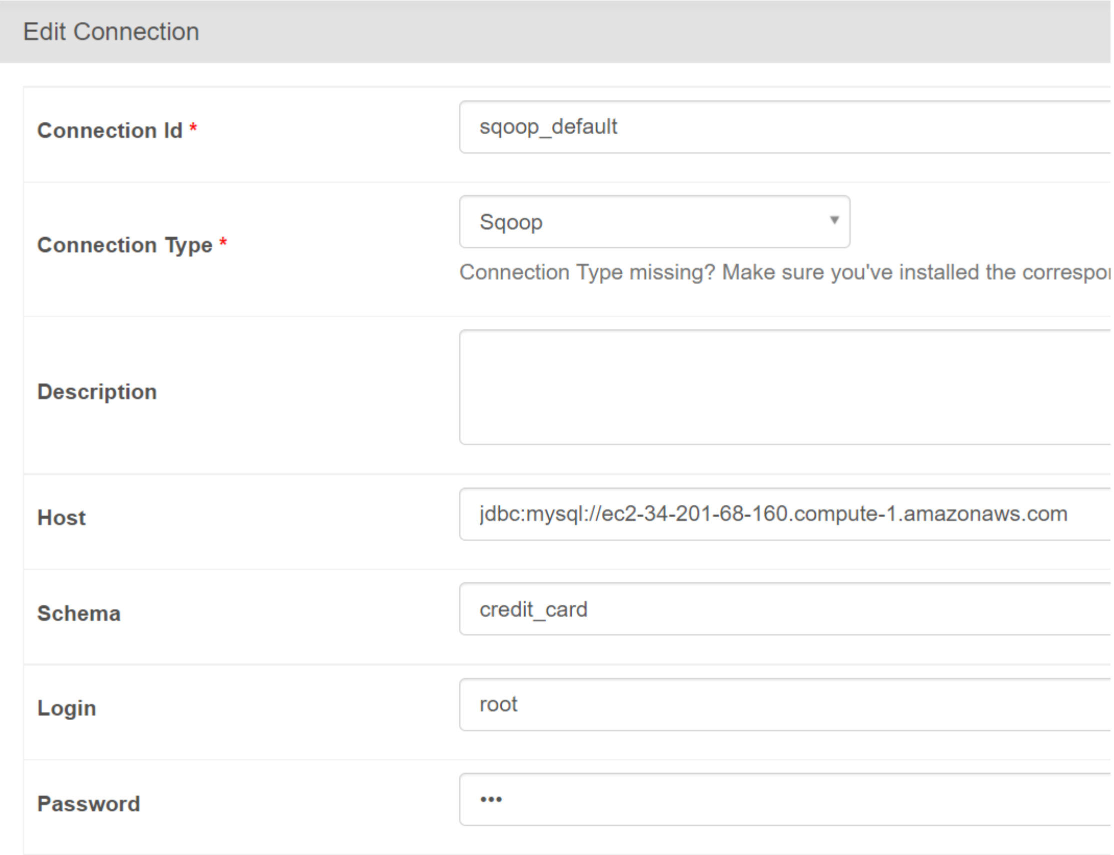


Now click on the edit button to the left of the **sqoop_default** connection.



□	presto_default	presto
□	qubole_default	qubole
□	redis_default	redis
□	segment_default	segment
□	sftp_default	sftp
□	spark_default	spark
□	sqlite_default	sqlite
□	sqoop_default	sqoop
□	ssh_default	ssh
□	tableau_default	tableau

Next, you need to make changes in the following details and click on Save



Edit Connection

Connection Id *	sqoop_default
Connection Type *	Sqoop Connection Type missing? Make sure you've installed the correspo
Description	
Host	jdbc:mysql://ec2-34-201-68-160.compute-1.amazonaws.com
Schema	credit_card
Login	root
Password	•••

Conn Id: sqoop_default

Conn Type: Sqoop (Select from the drop-down)

Connection URL: **jdbc:mysql://<public DNS>**

Schema: **credit_card**

Login: **root**

Password: **123**

Click on the Save button after you are done.

8. Now you need to place the **sample_sqoop.py** file in the **/home/hadoop/airflow/dags** directory. (You can use WinSCP or create a new file and paste the code in that file)

9. To ensure that the file there are no issues/errors with the file is it considered good practice to compile the program using the following command:

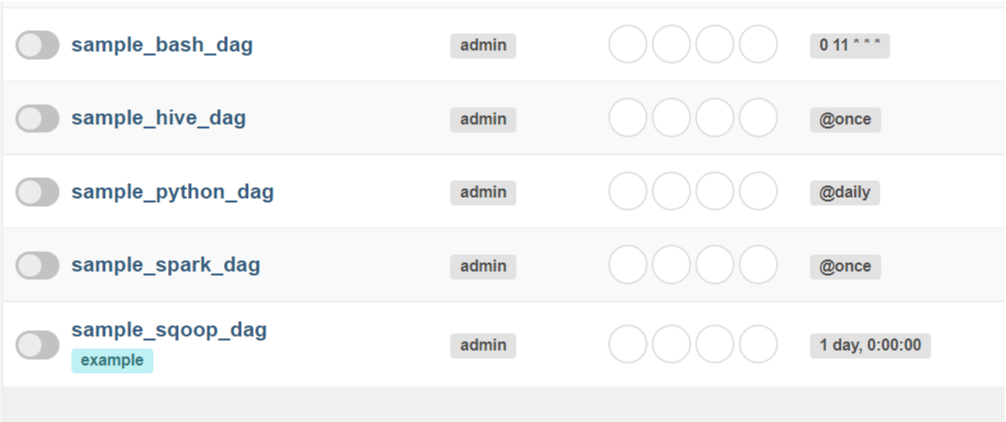
   python sample_sqoop.py

10. You can also use the following command to list the dags in your instance:

    airflow dags list

11. Once you have made sure that your dag file has no issues you can back go to the Airflow UI

12. Switch ON the DAG(sample_sqoop_dag)

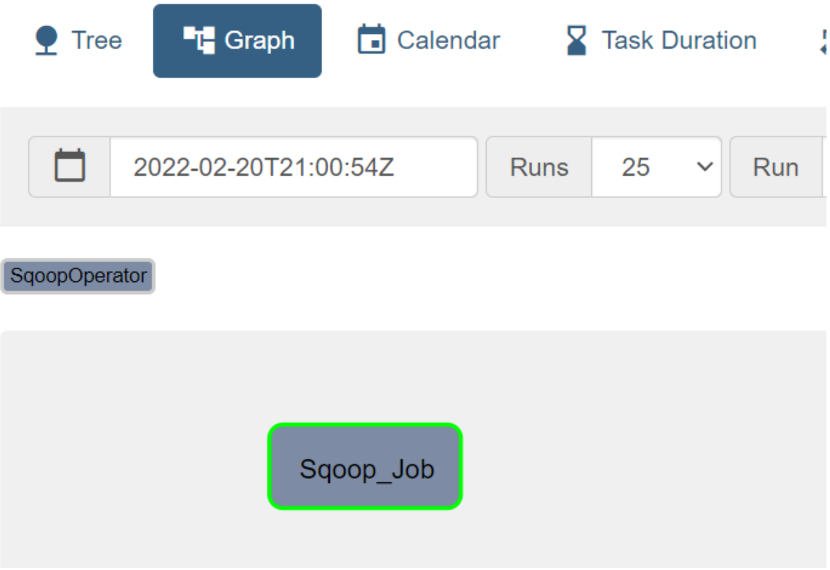


sample_bash_dag	admin	○○○○	0 11 * * *
sample_hive_dag	admin	○○○○	@once
sample_python_dag	admin	○○○○	@daily
sample_spark_dag	admin	○○○○	@once
sample_sqoop_dag example	admin	○○○○	1 day, 0:00:00

    (Note: The DAG might take a while to show up on the UI. Keep refreshing and wait patiently)

13. Click on the sample_sqoop_dag and go to the graph view

    You will see the task is running



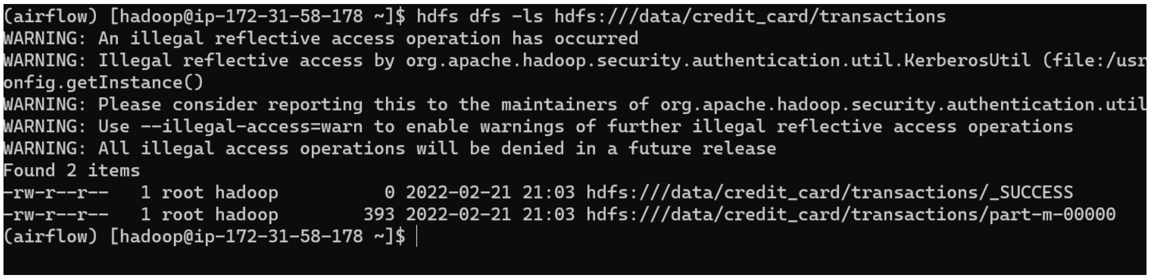
2022-02-20T21:00:54Z

Runs 25

Click on refresh and eventually, it will have successfully completed

14. Once the DAG has completed execution, the output will be generated in the location tardet_dir - **hdfs:///data/credit_card/transactions**

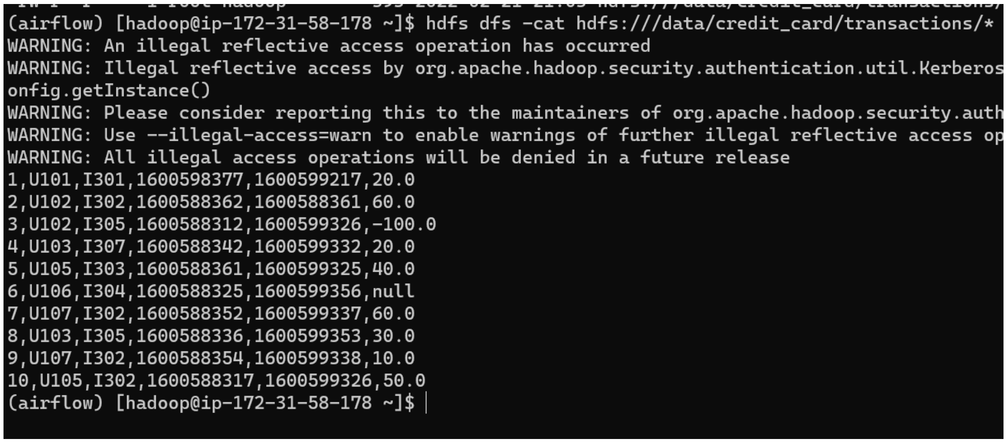
    You can use the **hdfs dfs -ls hdfs:///data/credit_card/transactions** command to see the same



```
(airflow) [hadoop@ip-172-31-58-178 ~]$ hdfs dfs -ls hdfs:///data/credit_card/transactions
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr
onfig.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-02-21 21:03 hdfs:///data/credit_card/transactions/_SUCCESS
-rw-r--r--   1 root hadoop        393 2022-02-21 21:03 hdfs:///data/credit_card/transactions/part-m-00000
(airflow) [hadoop@ip-172-31-58-178 ~]$
```

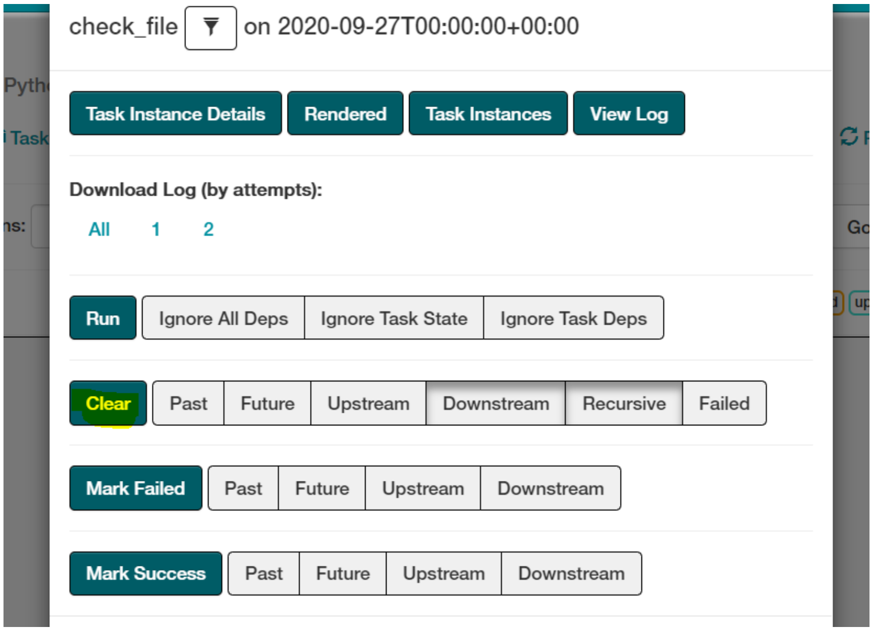
15. To view the result in the HDFS location use the following command:

    hdfs dfs -cat hdfs:///data/credit_card/transactions/*



```
(airflow) [hadoop@ip-172-31-58-178 ~]$ hdfs dfs -cat hdfs:///data/credit_card/transactions/*
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.Kerberos
onfig.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.auth
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access op
WARNING: All illegal access operations will be denied in a future release
1,U101,I301,1600598377,1600599217,20.0
2,U102,I302,1600588362,1600588361,60.0
3,U102,I305,1600588312,1600599326,-100.0
4,U103,I307,1600588342,1600599332,20.0
5,U105,I303,1600588361,1600599325,40.0
6,U106,I304,1600588325,1600599356,null
7,U107,I302,1600588352,1600599337,60.0
8,U103,I305,1600588336,1600599353,30.0
9,U107,I302,1600588354,1600599338,10.0
10,U105,I302,1600588317,1600599326,50.0
(airflow) [hadoop@ip-172-31-58-178 ~]$
```

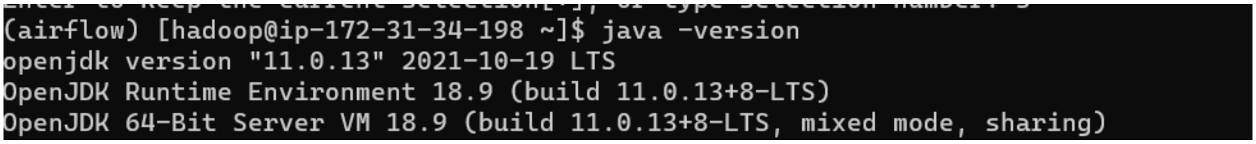
16. At any point, if a task fails you can click on it and click on the clear button to restart the task also remove the target_dir as mentioned in point 3 before you rerun the DAG



17. You can switch off your DAG if you don't want it to run anymore.

18. Make sure that you switch back to Java 8 by running the following command once you have run the Sqoop DAG. You can check this by running the following command:

```
java -version
```



```
Enter to keep the current selection[+], or type selection number: 3
(airflow) [hadoop@ip-172-31-34-198 ~]$ java -version
openjdk version "11.0.13" 2021-10-19 LTS
OpenJDK Runtime Environment 18.9 (build 11.0.13+8-LTS)
OpenJDK 64-Bit Server VM 18.9 (build 11.0.13+8-LTS, mixed mode, sharing)
```

If you still have Java 11, then you need to switch to Java 8 by running the following command.

```
sudo alternatives --config java <<< 1
```