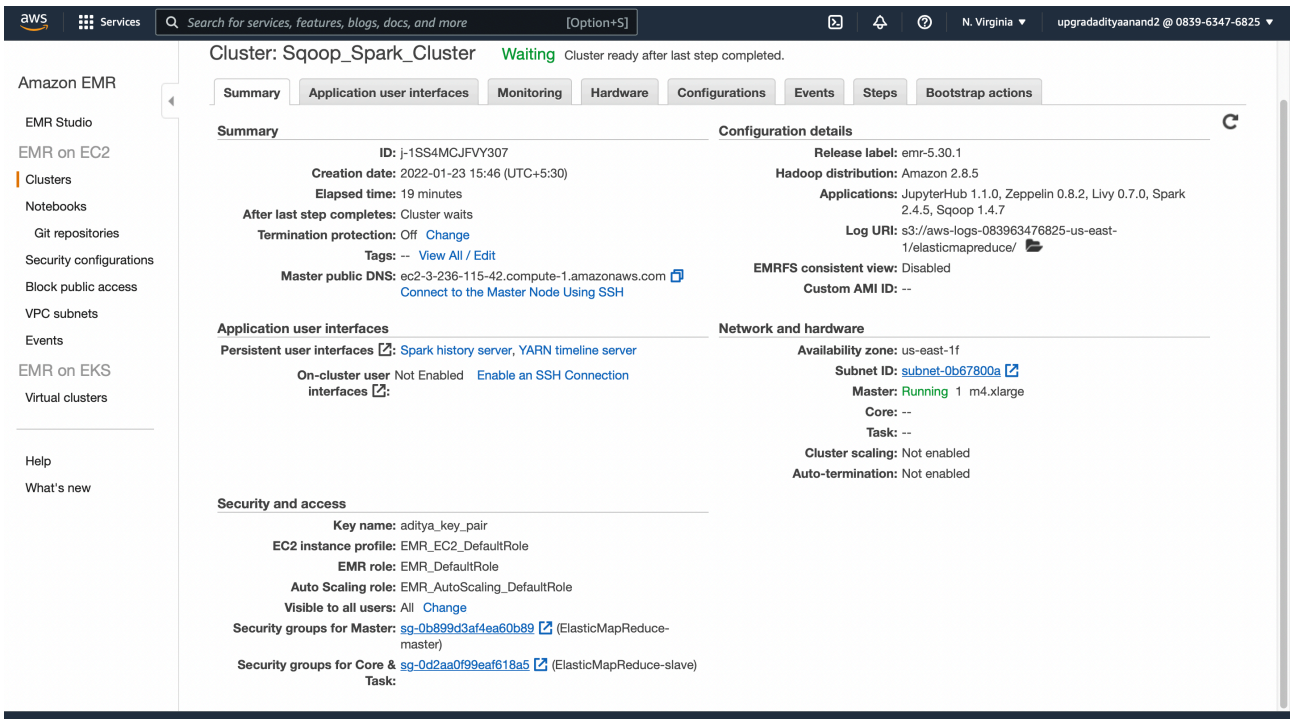
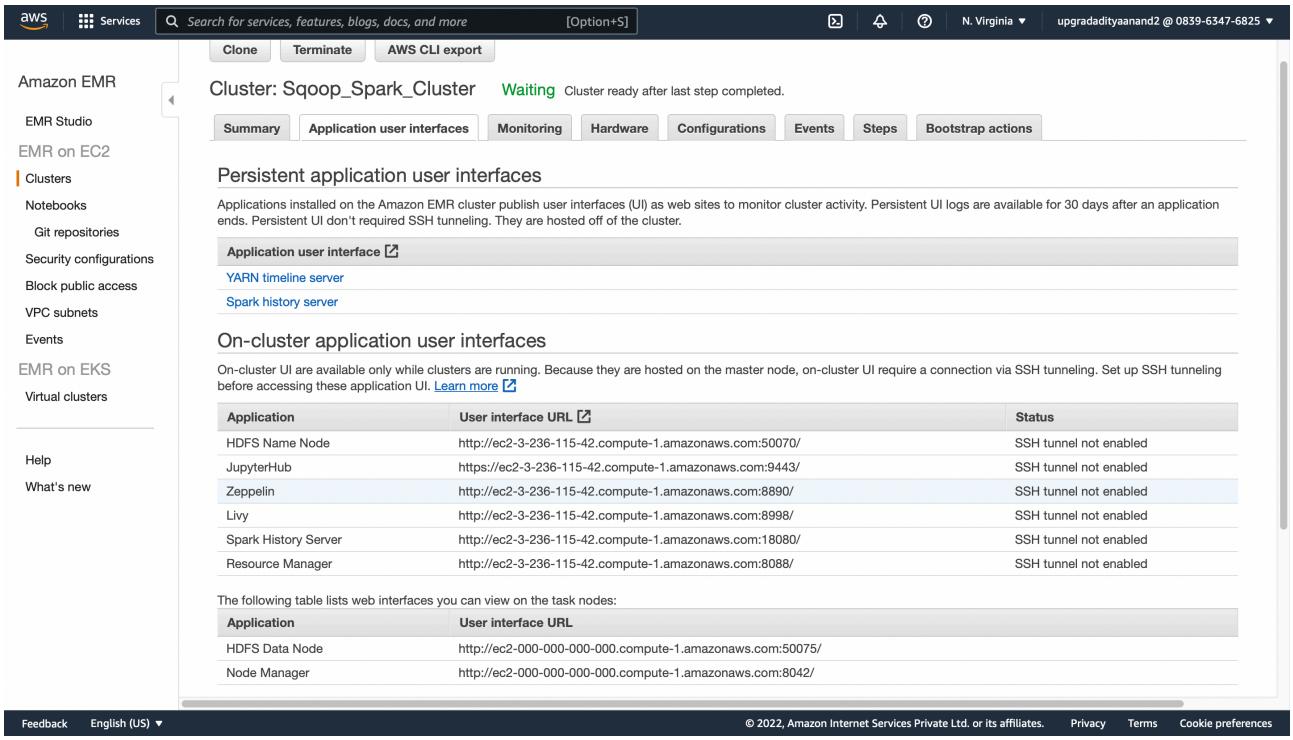
Data Ingestion from the RDS to HDFS using Sqoop

Sqoop and Spark Cluster setup and configuration:





Sqoop Import command used for importing table from RDS to HDFS:

```
sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/ testdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
```

> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/atm_txn_data \
> -m 1 --as-textfile

```
[root@ip-172-31-66-204 ~]# sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/testdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/atm_txn_data \
> -m 1 --as-textfile
```

Command used to see the list of imported data in HDFS:

hadoop fs -ls /user/root/atm_txn_data

Screenshot of the imported data:

Command used to copy the imported data to Livy so that Pyspark can access the data:

hadoop fs -cp /user/root/atm_txn_data/part-m-00000 /user/livy

```
22/01/23 10:46:25 INFO mapreduce.ImportJobBase: Transferred 506.6059 MB in 46.5815 seconds (10.8757 MB/sec)
22/01/23 10:46:25 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
[root@ip-172-31-66-204 ~]# hadoop fs -ls /user/root/atm_txn_data
Found 2 items
-rw-r--r--   1 root hadoop          0 2022-01-23 10:46 /user/root/atm_txn_data/_SUCCESS
-rw-r--r--   1 root hadoop  531214815 2022-01-23 10:46 /user/root/atm_txn_data/part-m-00000
[root@ip-172-31-66-204 ~]#
[root@ip-172-31-66-204 ~]# hadoop fs -cp /user/root/atm_txn_data/part-m-00000 /user/livy
[root@ip-172-31-66-204 ~]# hadoop fs -ls /user/livy
Found 1 items
-rw-r--r--   1 root livy  531214815 2022-01-23 10:49 /user/livy/part-m-00000
[root@ip-172-31-66-204 ~]#
```