

# **Data Ingestion from the RDS to HDFS using Sqoop**

**Sqoop and Spark Cluster setup and configuration:**

aws

Services

Search for services, features, blogs, docs, and more

[Option+S]

N. Virginia

upgradadityaanand2 @ 0839-6347-6825

Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Cluster: Sqoop\_Spark\_Cluster

Waiting

Cluster ready after last step completed.

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

Configuration details

Application user interfaces

Network and hardware

Security and access

Feedback

English (US)

© 2022, Amazon Internet Services Private Ltd. or its affiliates.

Privacy

Terms

Cookie preferences

aws

Services

Search for services, features, blogs, docs, and more

[Option+S]

N. Virginia

upgradadityaanand2 @ 0839-6347-6825

Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Cluster: Sqoop\_Spark\_Cluster

Waiting

Cluster ready after last step completed.

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Persistent application user interfaces

On-cluster application user interfaces

Feedback

English (US)

© 2022, Amazon Internet Services Private Ltd. or its affiliates.

Privacy

Terms

Cookie preferences

```
> --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.cyaie1c9bmnf.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 ~]# █
```