ETL CASESTUDY

Problem Statement

Extract data of atm transactions form RDS and perform ETL operations to **load** data into Amazon Redshift warehouse, using Sqoop pipeline, HDFS and S3 storages and PySpark for processing and **transformation**. And perform analytical queries.

Step 1: Check source data count

We extract the data from the MySql RDS using the **Apache Sqoop pipeline**. We first check the number of records in MySql SRC_ATM_TRANS table (source table).

```
mysql> select count(*) from SRC_ATM_TRANS;
+----------+
| count(*) |
+----------+
|  2468572 |
+----------+
1 row in set (11.23 sec)
```

The query results show us there are **2468572 records**.

Step 2 : Import data into HDFS

Command to **import data into 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_trans_data \
-m 1 \
--as-parquetfile
```

Data is extracted into HDFS path /user/root/atm_trans_data as a parquet file, and the null values will be treated as null and not a string. 1 mapper is used.

```
[root@ip-10-0-0-211 ~]# sqoop import \
> --connect jdbc:mysql://upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com/t
estdatabase \
> --table SRC_ATM_TRANS \
> --username student --password STUDENT123 \
> --null-string '\\N' --null-non-string '\\N' \
> --target-dir /user/root/atm_trans_data \
> -m 1 \
> --as-parquetfile
```

Result: 2468572 records are imported

```
21/03/27 10:17:26 INFO mapreduce.ImportJobBase: Transferred 42.6079 MB in 87.2749 seconds
(499.92 KB/sec)
21/03/27 10:17:26 INFO mapreduce.ImportJobBase: Retrieved 2468572 records.
```

Step 3 : List imported files in HDFS

Command to list the file imported.

Hadoop fs -ls */user/root/atm_trans_data*

```
[root@ip-10-0-0-211 ~]# hadoop fs -ls /user/root/atm_trans_data
Found 3 items
drwxr-xr-x   - root supergroup          0 2021-03-27 10:15 /user/root/atm_trans_data/.meta
data
drwxr-xr-x   - root supergroup          0 2021-03-27 10:17 /user/root/atm_trans_data/.sign
als
-rw-r--r--   3 root supergroup   44667278 2021-03-27 10:17 /user/root/atm_trans_data/99cab
c80-c250-4b0e-97f2-1c6766ab412e.parquet
```

File listed HDFS web UI

/user/root/atm_trans_data Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	Sat Mar 27 15:45:59 +0530 2021	0	0 B	.metadata
drwxr-xr-x	root	supergroup	0 B	Sat Mar 27 15:47:24 +0530 2021	0	0 B	.signals
-rw-r--r--	root	supergroup	42.6 MB	Sat Mar 27 15:47:24 +0530 2021	3	128 MB	99cabc80-c250-4b0e-97f2-1c6766ab412e.parquet