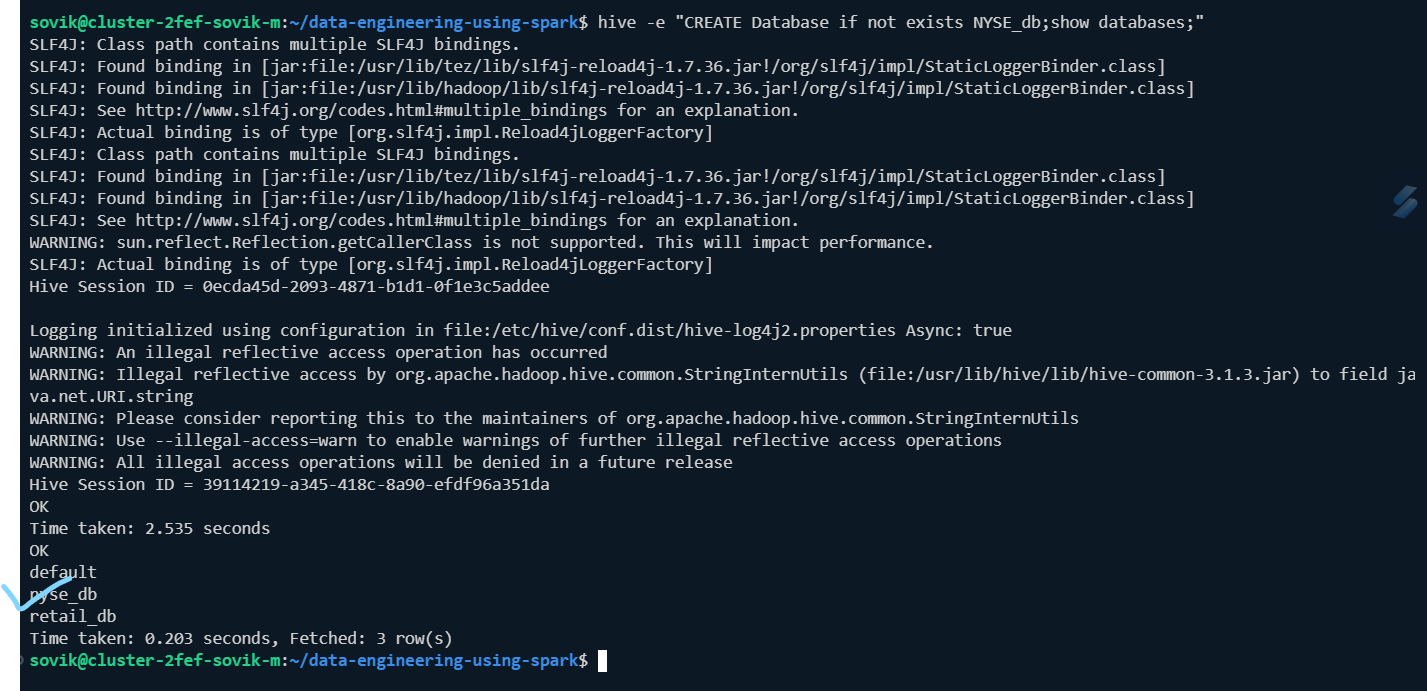
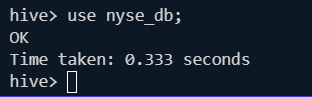
Create new database:

**sovik@cluster-2fef-sovik-m:~/data-engineering-using-spark$ hive -e "CREATE Database if not exists NYSE\_db;show databases;"**

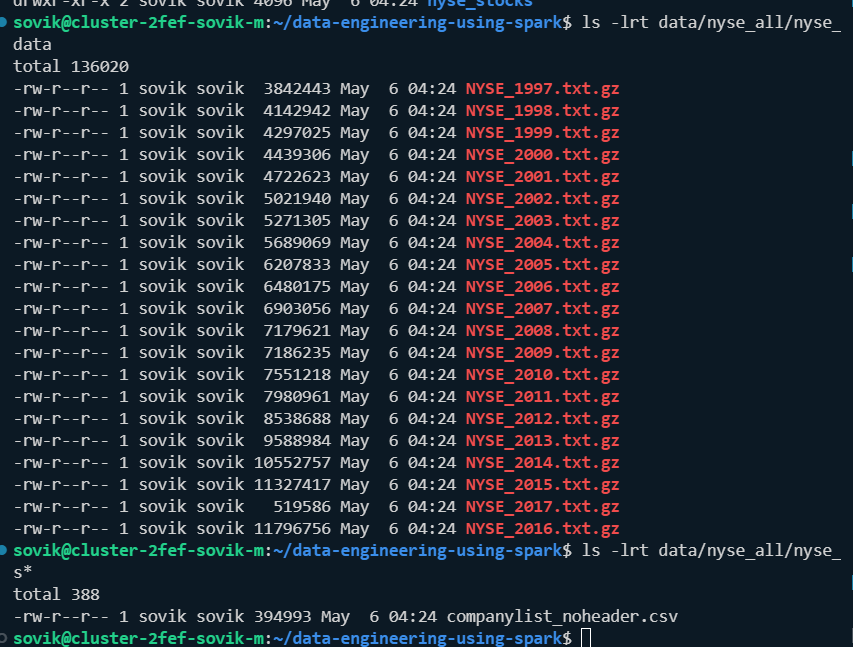


**hive> use nyse\_db;**

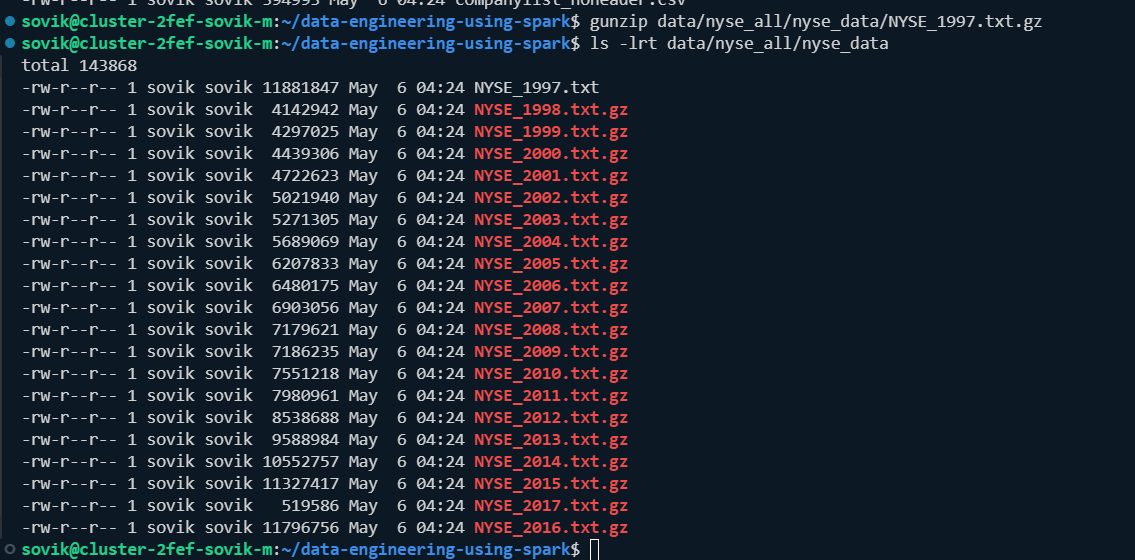
****

**Now we are going to create 2 tables 1 with *Parquet* format another using *Text* file format:**

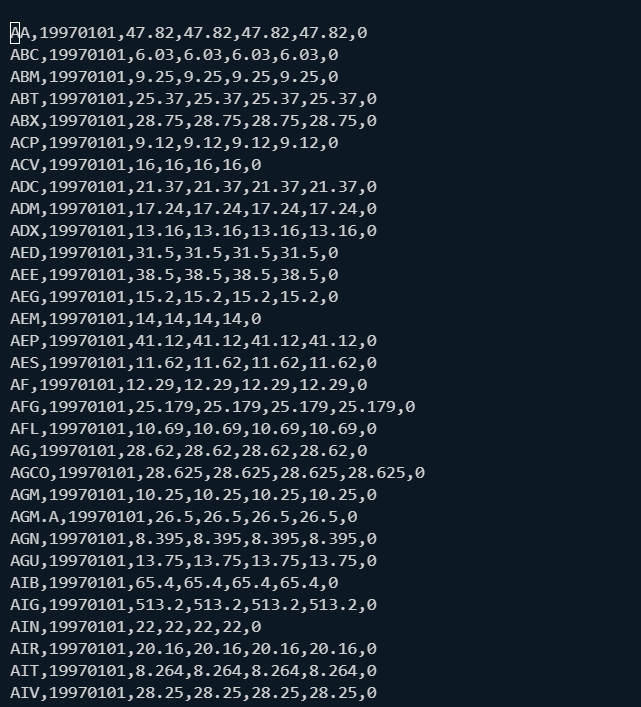
**We will be working on the below tables:**

****

**We can unzip 1 file and check:**

****

**Command:view path filename**

****

**The date can be Stockname,date(yyyymmdd),opening\_price,high\_price,low\_price,closing\_price,change\_pct**

**We need to create a table from the txt files,we then load it into a table and then convert the file from text format to parquet format and then partition the table based on months**

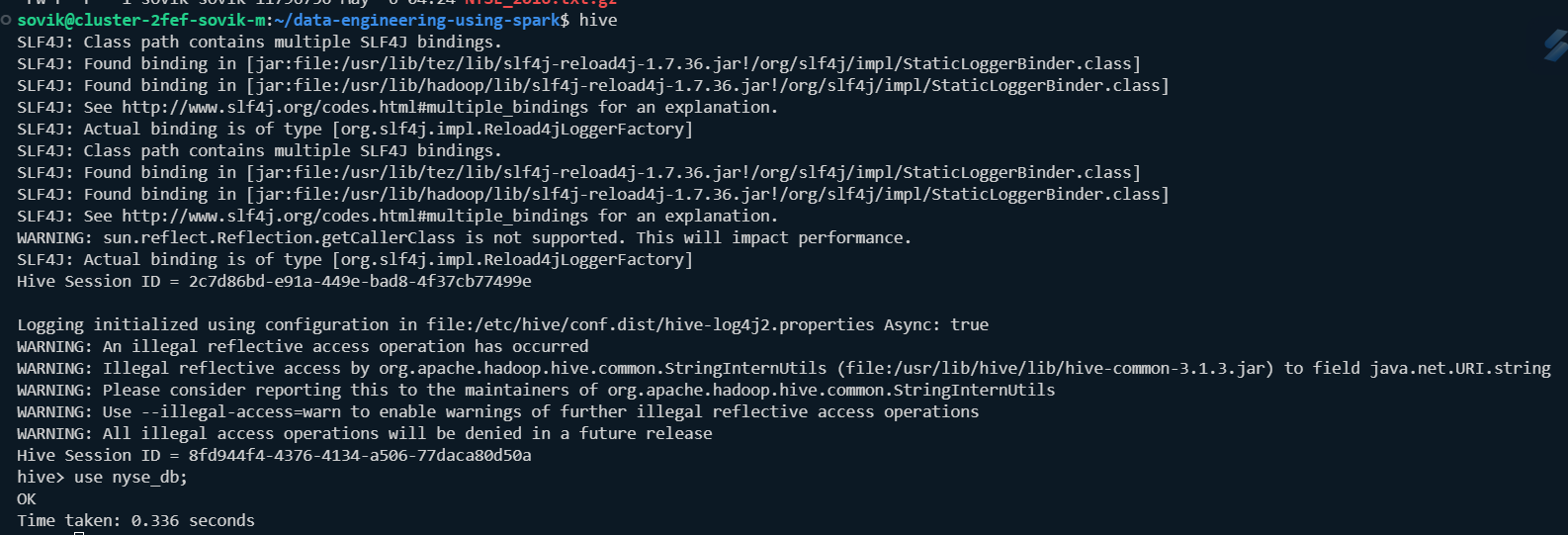
**PIPELINE:**

****

**For us local path is:**

**/home/Sovik/data-engineering-using-spark/data/nyse\_all/nyse\_data/**

**We use nyse\_db:**

****

1. **First lets create the Parquet final table : nyse\_daily**

**Format:Parquet**

**Partitioned by : trademonth**

**SQL command:** **create table if not exists nyse\_daily(**

**ticker STRING,**

**trade\_date INT,**

**open\_price FLOAT,**

**high\_price FLOAT,**

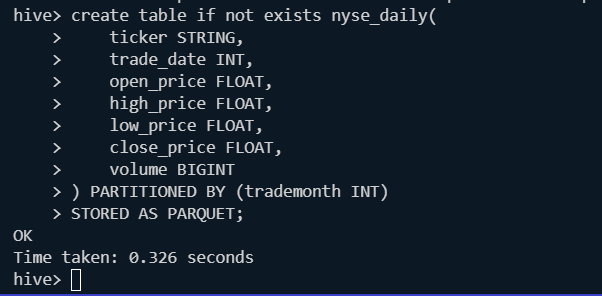
**low\_price FLOAT,**

**close\_price FLOAT,**

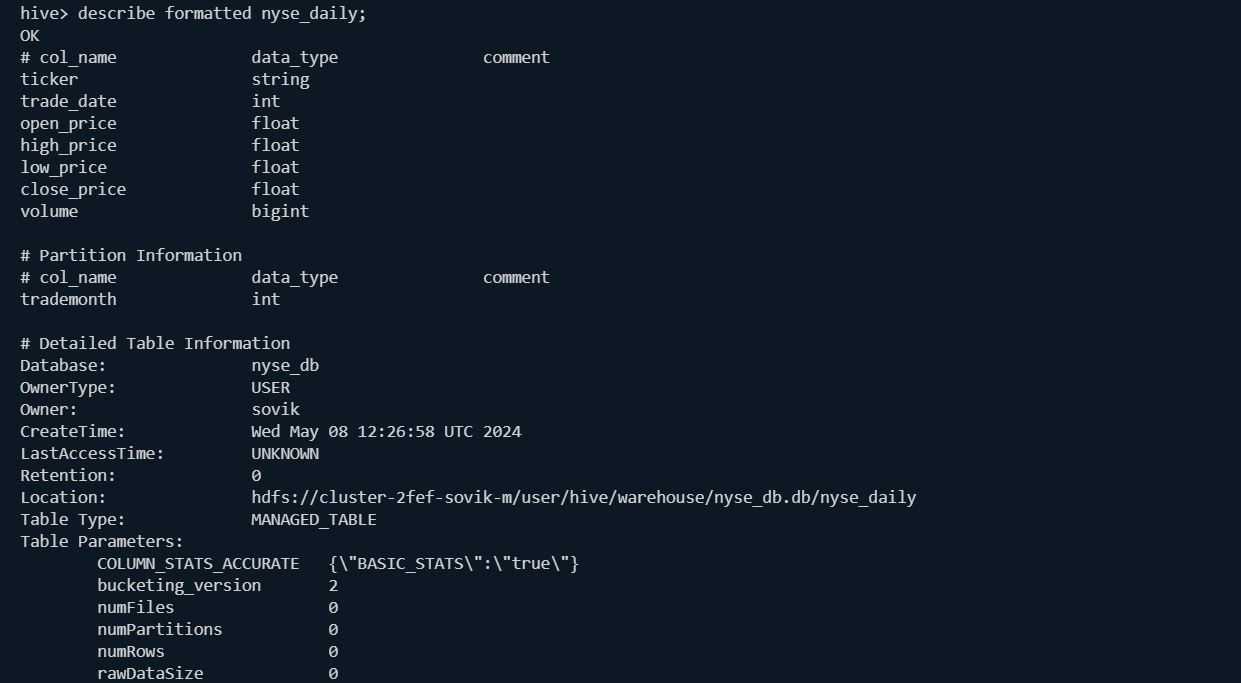
**volume BIGINT**

**) PARTITIONED BY (trademonth INT)**

**STORED AS PARQUET;**

****

**Check:**

****

**Lets create the staging table: nyse\_stg**

**Format:text**

**SQL: SQL command:** **create table if not exists nyse\_stg(**

**ticker STRING,**

**trade\_date INT,**

**open\_price FLOAT,**

**high\_price FLOAT,**

**low\_price FLOAT,**

**close\_price FLOAT,**

**volume BIGINT**

**)**

**Path : /home/sovik/data-engineering-using-spark/data/nyse\_all/nyse\_data/NYSE\_1997.txt.gz**

**Loading data in staging table:**

**Time taken: 0.538 seconds**

**hive> LOAD DATA LOCAL INPATH**

**> '/home/sovik/data-engineering-using-spark/data/nyse\_all/nyse\_data/NYSE\_1997.txt.gz'**

**> OVERWRITE INTO TABLE nyse\_stg;**

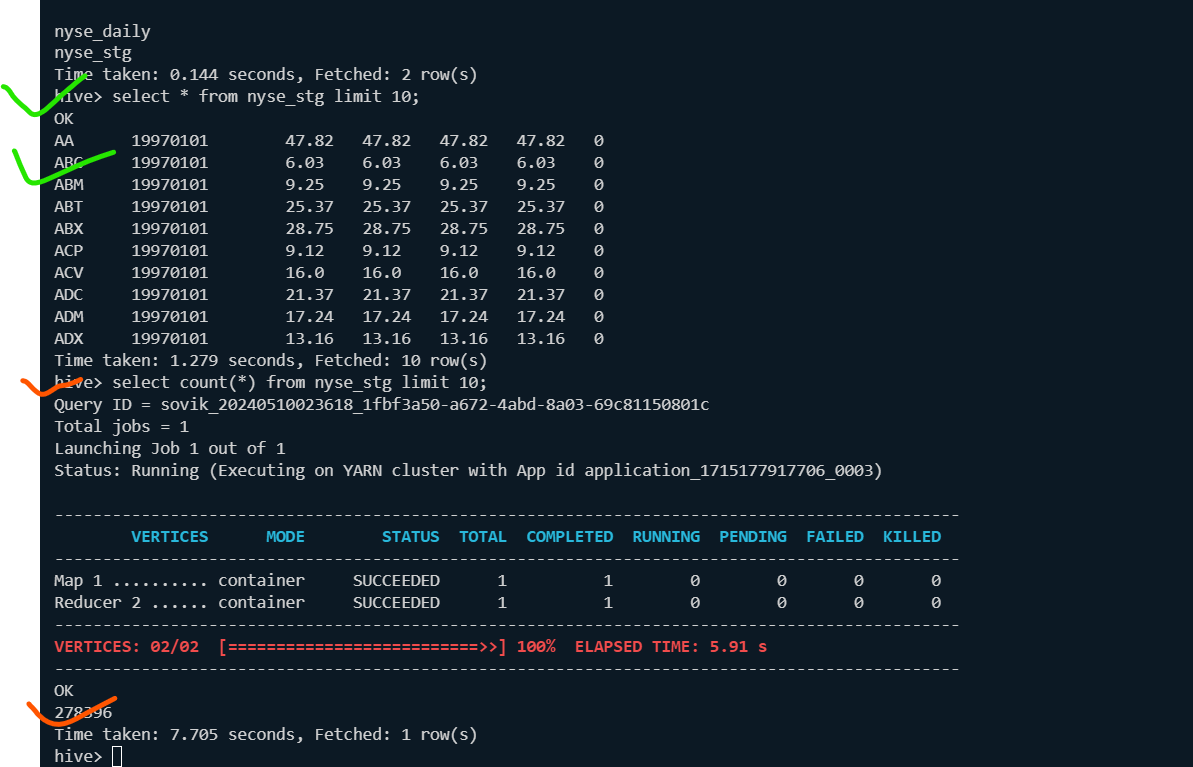
**Loading data to table nyse\_db.nyse\_stg**

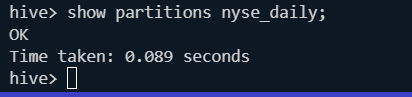
**OK**

**Time taken: 0.596 seconds**

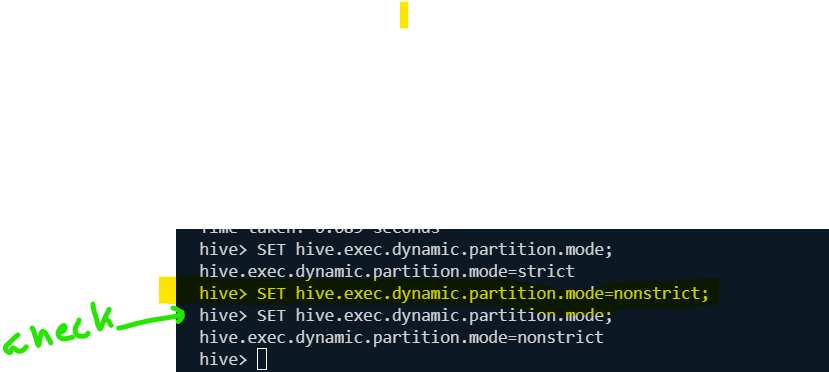
**hive>**

**Validate:**

**check existing partitions in nyse\_daily table;**

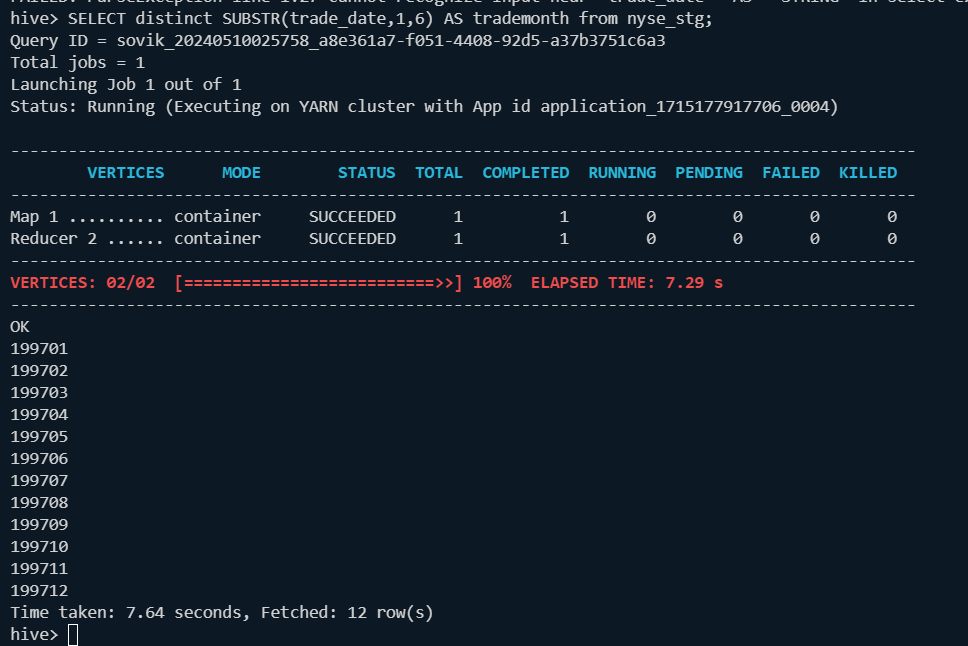
****

**To set dynamic partitioning, we do as follows:**

****

**Now we can write relevant insert statement to populate the table with partitions.**

**Computing the *trademonth*:**

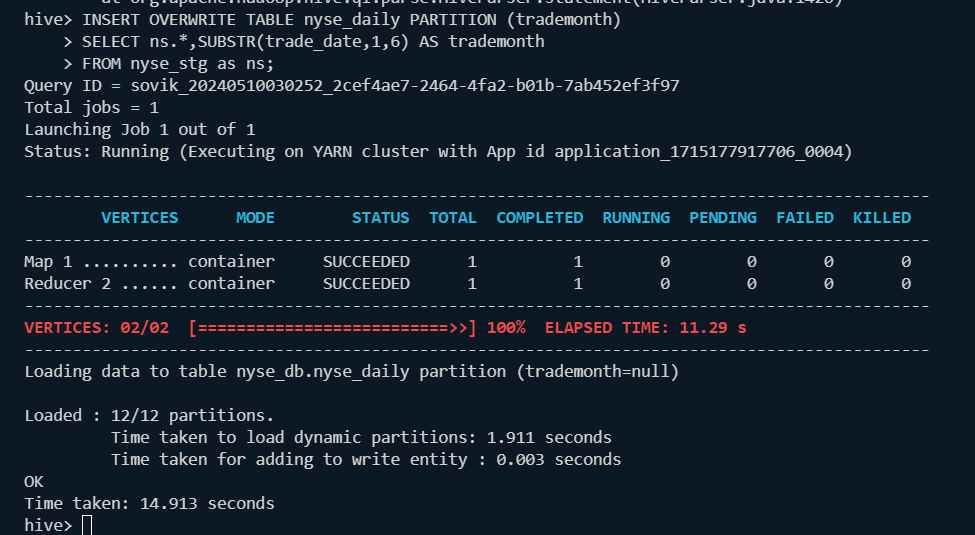
****

**Command to insert:**

**INSERT OVERWRITE TABLE nyse\_daily PARTITION (trademonth)**

**SELECT ns.\*,SUBSTR(trade\_date,1,6) AS trademonth**

**FROM nyse\_stg as ns;**

****

**Validate:**

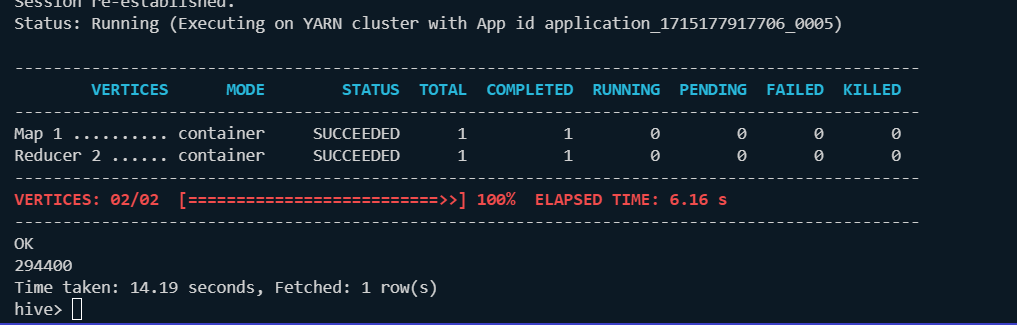
****

**Lets populate the 1998:**

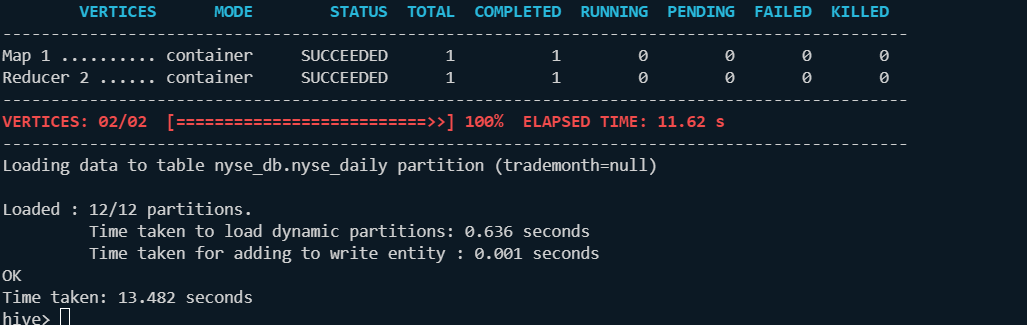
**Overwrite with 1998 data:**

**LOAD DATA LOCAL INPATH '/home/sovik/data-engineering-using-spark/data/nyse\_all/nyse\_data/NYSE\_1998.txt.gz' OVERWRITE INTO TABLE nyse\_stg;**

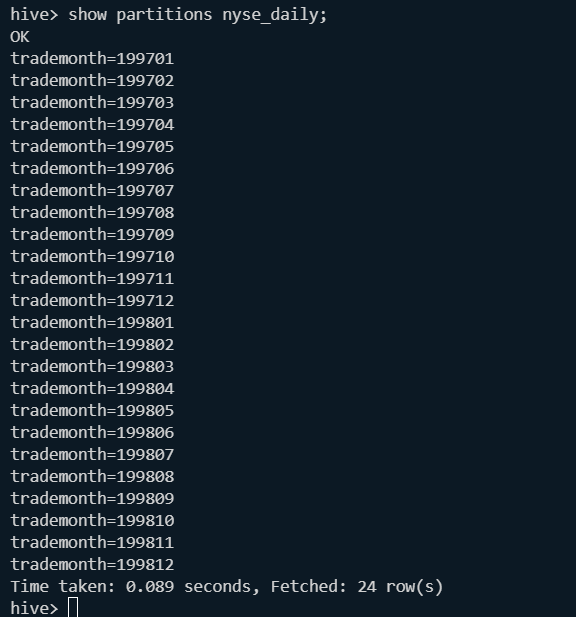
**Validate:**

****

**Now insert the staging data to nyse\_daily table:**

****

**Validate:**

****

**Validate with a query:**

hive> select trademonth,count(\*) as trades\_permonth from nyse\_daily

> group by trademonth

> order by trademonth;

Query ID = sovik\_20240510031824\_18f05312-43e8-4bb1-9ae1-a745a7dd0154

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1715177917706\_0005)

----------------------------------------------------------------------------------------------

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED

----------------------------------------------------------------------------------------------

Map 1 .......... container SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... container SUCCEEDED 1 1 0 0 0 0

Reducer 3 ...... container SUCCEEDED 1 1 0 0 0 0

----------------------------------------------------------------------------------------------

VERTICES: 03/03 [==========================>>] 100% ELAPSED TIME: 7.76 s

----------------------------------------------------------------------------------------------

OK

199701 23989

199702 20957

199703 22011

199704 23121

199705 23221

199706 22289

199707 24578

199708 22540

199709 23656

199710 24908

199711 21861

199712 25265

199801 24224

199802 22060

199803 24412

199804 24573

199805 23632

199806 24880

199807 26109

199808 23895

199809 25069

199810 25124

199811 24041

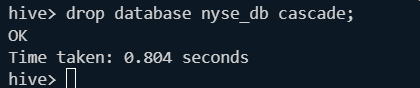
199812 26381

Time taken: 8.117 seconds, Fetched: 24 row(s)

hive>

**Now we create the application:**

**First we drop existing database:**

****

**To create a wrapper and hql script we can check the history;**

**Cat ./hivehistory**

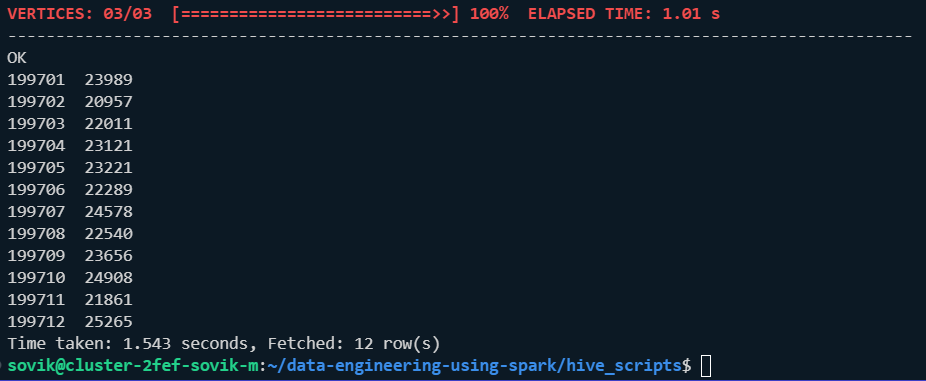
**Now after creating the script we go to path of script:**

****

**Command to execute it:**

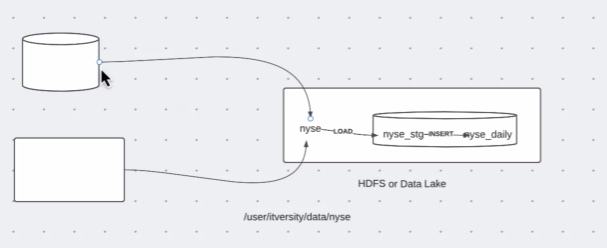
**hive -f nyse\_dbloader\_script.hql --hivevar tradeyear=1997**

**Validated:**

****

**If we use command:** **hive -f nyse\_dbloader\_script.hql --verbose --hivevar tradeyear=1998 it will show the command being executed along with the steps.**

**Lets redesign the script for ease of creation of wrappers:**

****

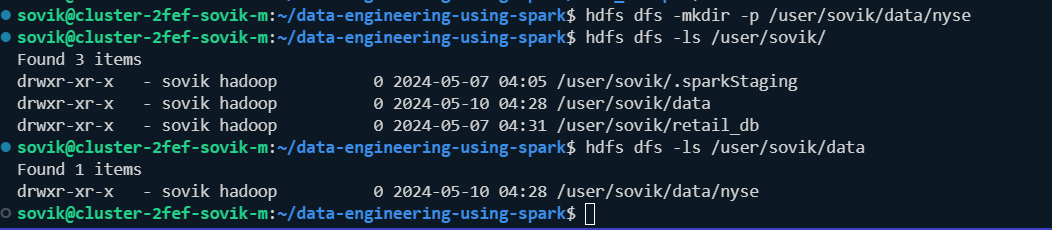
*Data can come from external db now the data is kept in hdfs****(/user/Sovik/data/nyse****) from there periodically data will move to nyse\_db.nyse\_stg -🡪nyse\_db.nyse\_daily*

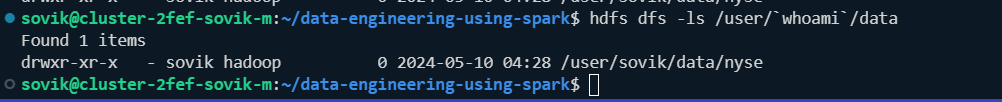
**We create the new script:**

**sovik@cluster-2fef-sovik-m:~/data-engineering-using-spark/hive\_scripts$ cp -pR nyse\_dbloader\_script.hql nyse\_dbloader\_script\_v2.hql**

**sovik@cluster-2fef-sovik-m:~/data-engineering-using-spark/hive\_scripts$**

**create the directory:/user/Sovik/data/nyse**

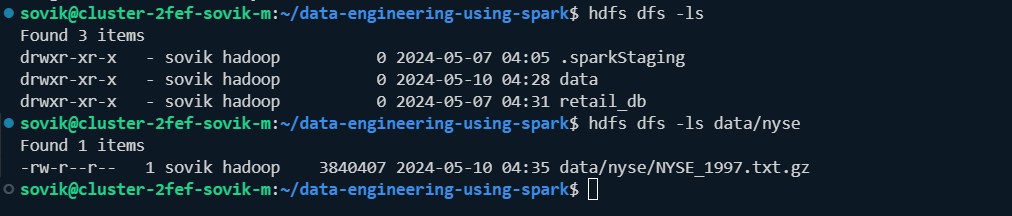
****

****

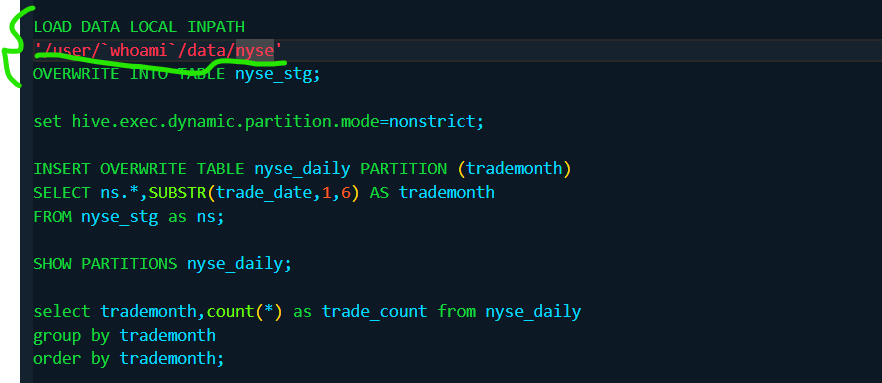
**Lets validate to load the data from nyse\_data in local to hdfs directory**

**Command:**

**sovik@cluster-2fef-sovik-m:~/data-engineering-using-spark$ hdfs dfs -put /home/sovik/data-engineering-using-spark/data/nyse\_all/nyse\_data/NYSE\_1997.txt.gz /user/sovik/data/nyse**

****

**Only change in script:**

****

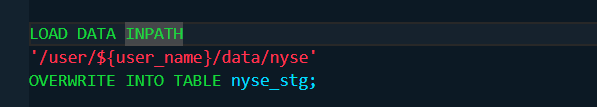
**Now we see as it is performed:**

**Hive -e “DROP database nyse\_db CASCADE;”**

**hive -f nyse\_dbloader\_script\_v2.hql –verbose**

**this is not working as we cannot put shell command in hive**

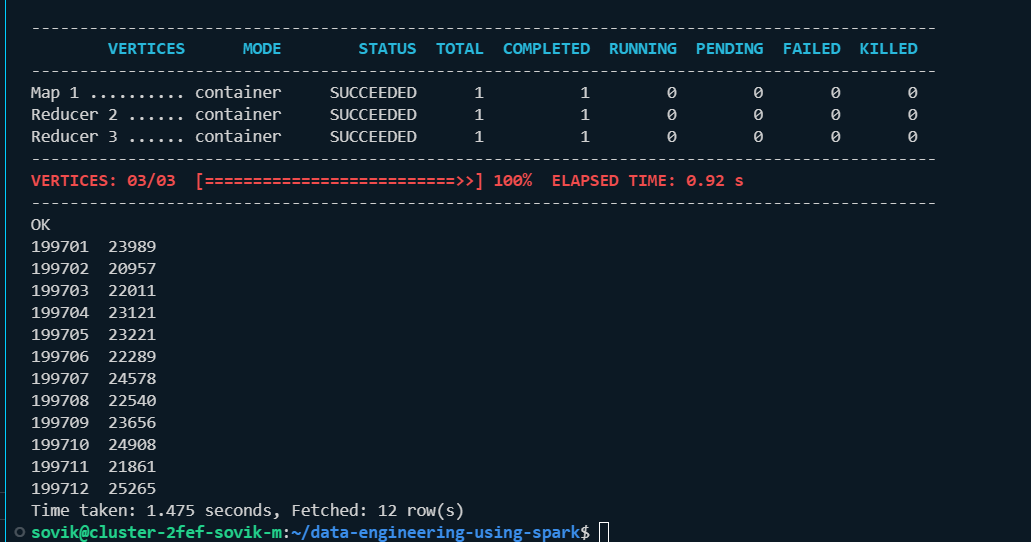
**to fix it we do as below:**

****

**now we drop db and use below command:**

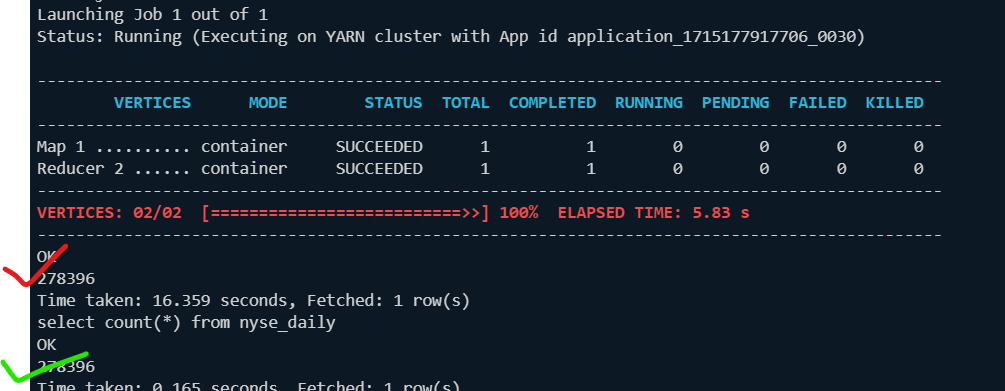
**hive -f hive\_scripts/nyse\_dbloader\_script\_v2.hql --hivevar username=`whoami` --verbose**

**Validated:**

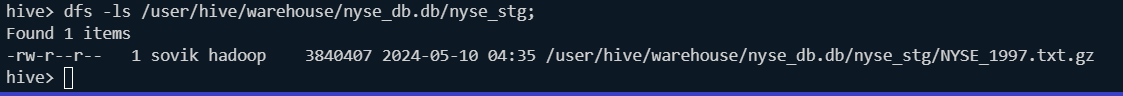
****

**Commad validation:**

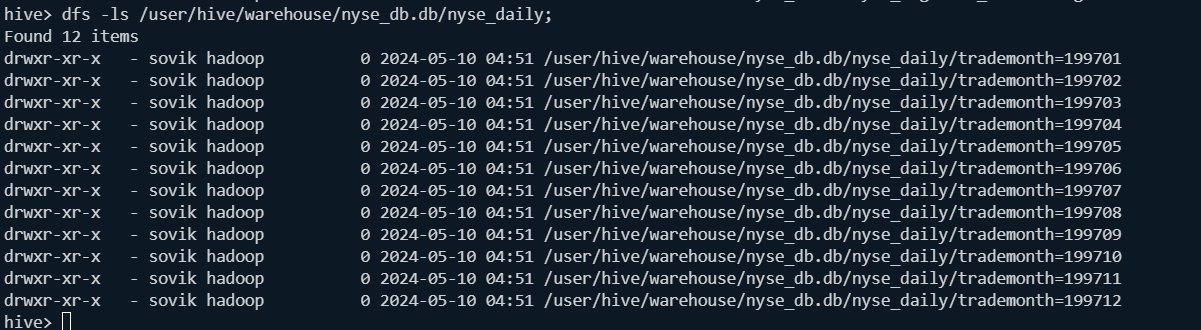
**hive -e "USE nyse\_db;select count(\*) from nyse\_stg;select count(\*) from nyse\_daily;" –verbose**

**o/P:  
**

**stg table location**

****

**Daily table location:**

****

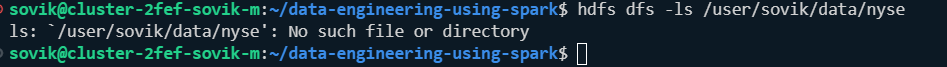
**Now we design the shell script to deply the script in HDFS for distributed worker nodes as this is now only present in the master node:**

**We create the script folder in hdfs folder**

****

**Copy hdfs\_script to hdfs:**

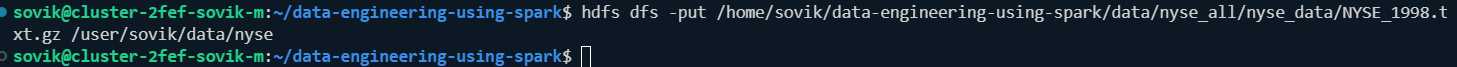
****

**No file in **

**Lets copy 1998 data:**

****

**Data sent from local to hive directory:**

****

**Code:** **hive -f hdfs://cluster-2fef-sovik-m/user/`whoami`/scripts/nyse\_dbloader\_script\_v2.hql --hi**

**vevar user\_name=`whoami` --verbose**

**Validated:**

****

**Now we create shell script:**

**Shell wrapper:**

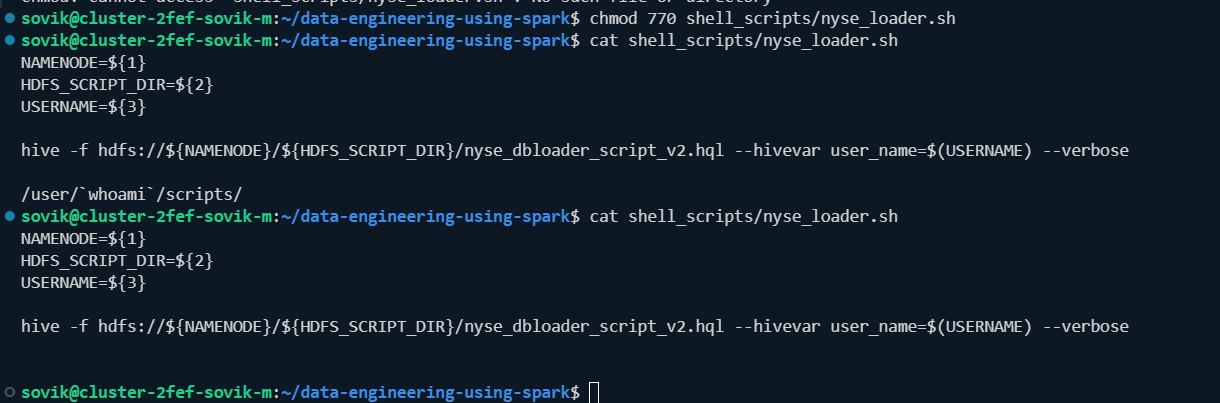
NAMENODE=*${1}*

HDFS\_SCRIPT\_DIR=*${2}*

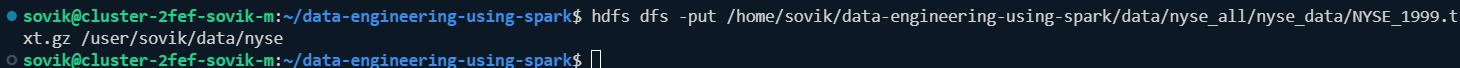
USERNAME=*${3}*

hive -f hdfs://${NAMENODE}/${HDFS\_SCRIPT\_DIR}/nyse\_dbloader\_script\_v2.hql --hivevar user\_name=${USERNAME} --verbose

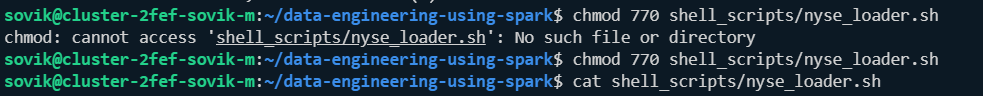
**Command:**

****

**Lets first load 1999 data:**

****

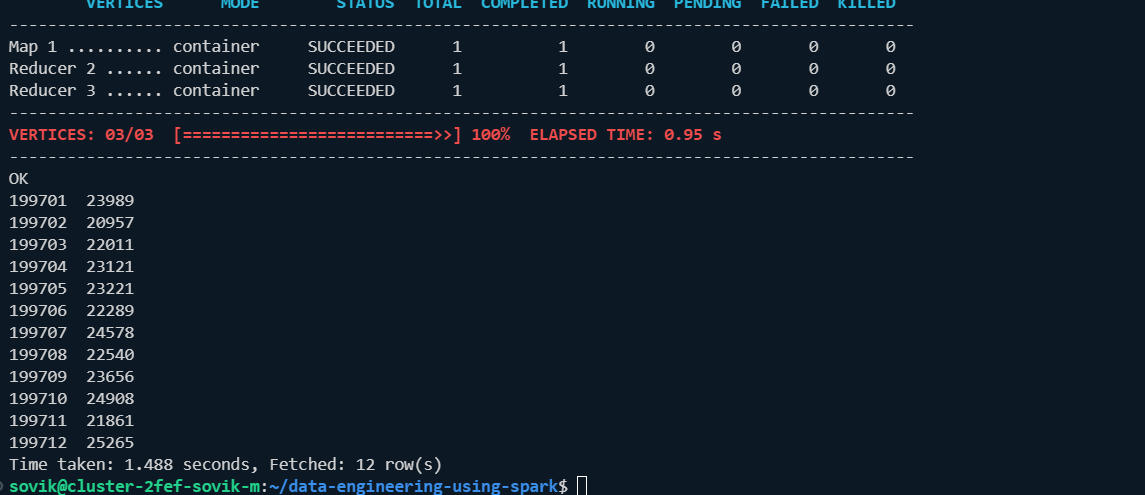
**Give u+x to user and groups for the shell script:**

****

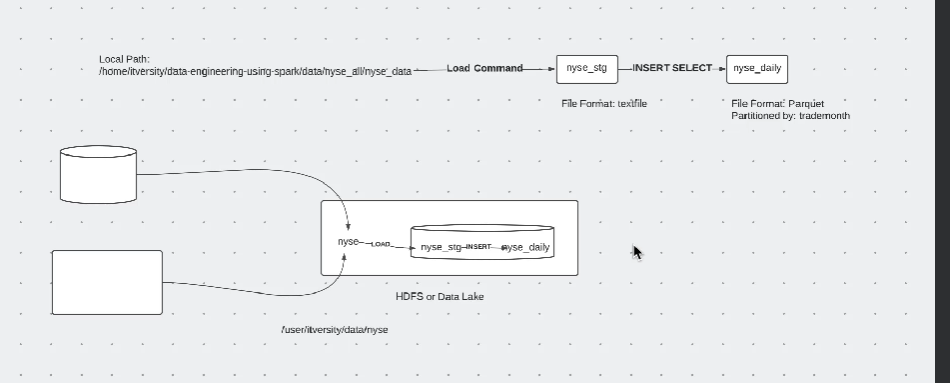
**Lets hit the command**

**./shell\_scripts/nyse\_loader.sh cluster-2fef-sovik-m /user/sovik/scripts sovik**

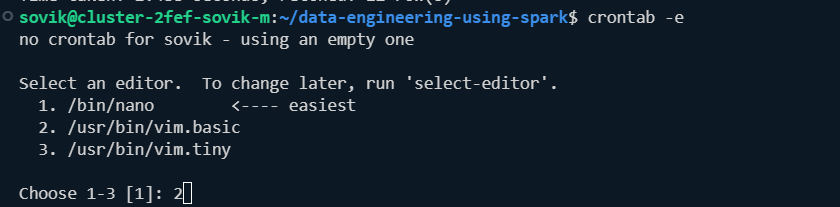
**Command is successful:**

****

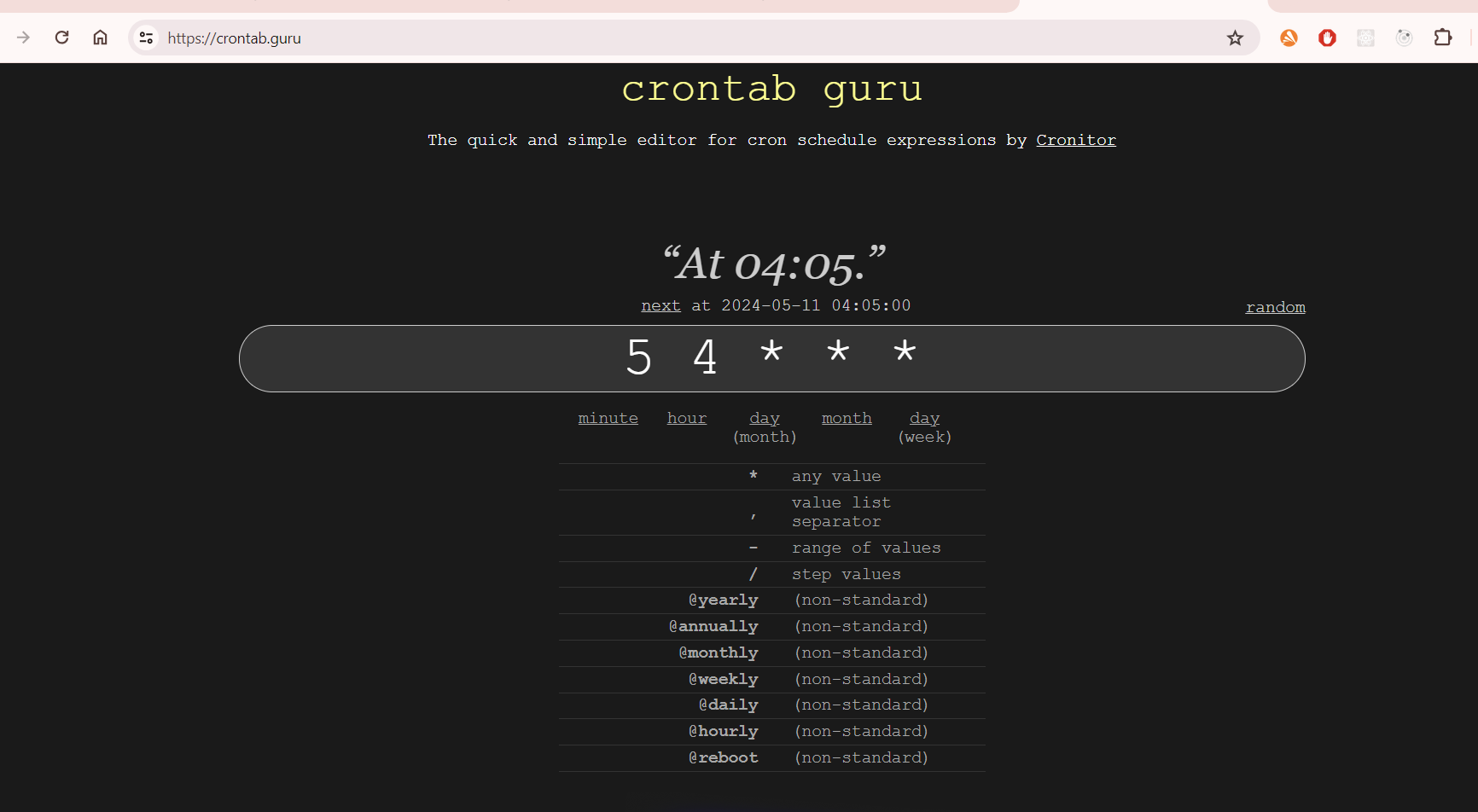
**Scheduling with cron:**

****

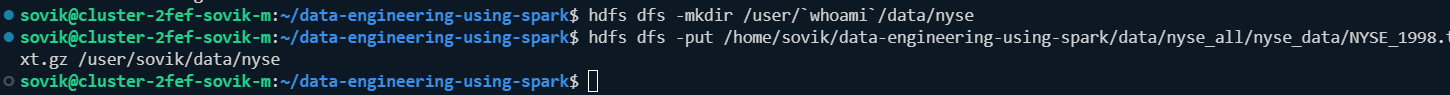
**Cron job schedule tasks as per intervals:daily,weekly,hourly etc:**

****

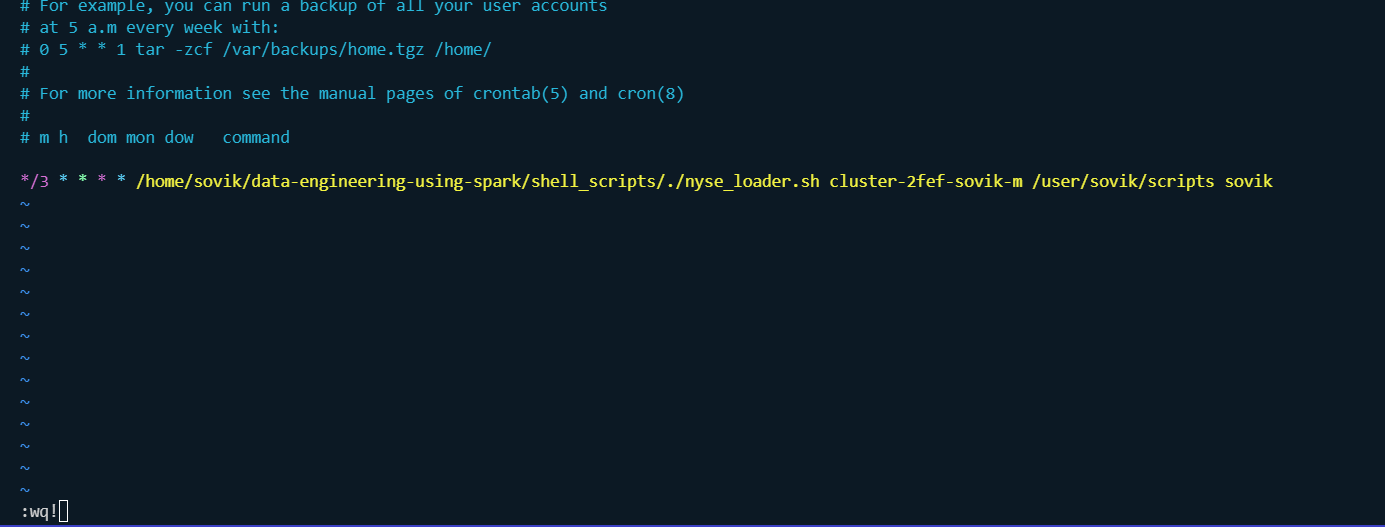
**Crontab example:**

****

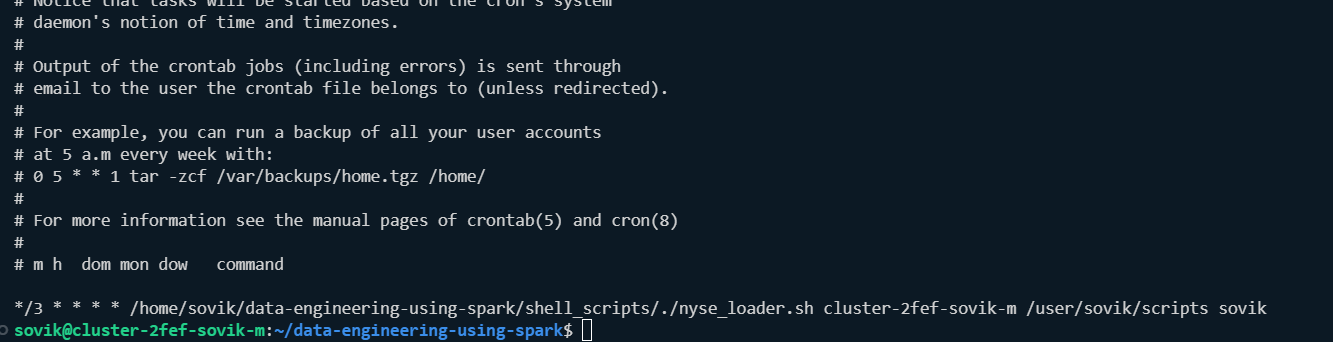
**Cleanup for setting crontab for 1998 data:**

****

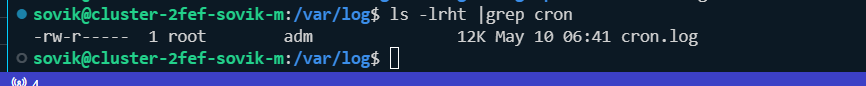
**Every third minute scheduling:**

****

**Validate :**

****

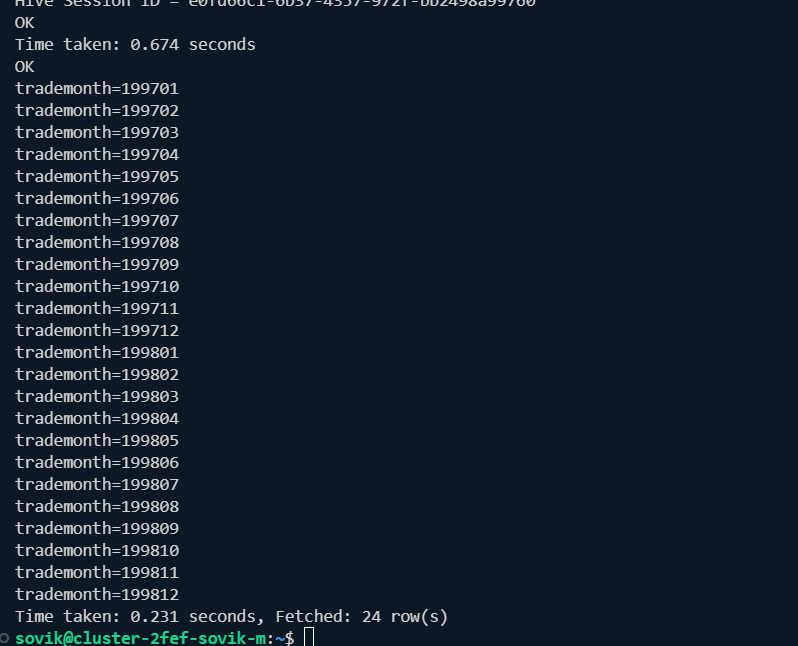
Check in var/log:

****

**Validate:**

****

**O/P:**

****

**Now we disable the cron/delete**

**We can use a log file to log creation in var/log/**

**Command:**

**# : \*/3 \* \* \* \* /home/sovik/data-engineering-using-spark/shell\_scripts/./nyse\_loader.sh cluster-2fef-sovik-m /user/sovik/scripts sovik>> var/log/nyse/nyse\_loader.log**

**Note:create the folder first:**