BIG DATA PROJECT-2
BUAN 6346.504

Group members:

Barath Kumar Dhnasekar-bxd220033

Jagadeep Nandagopal-jxn220044

Rahul Sadineni-rxs230051

Kosuri Durga Sravya-dxk220083

Nikitha Masineni-nxm230033

Navya Sahithi Surapaneni-nxs230031

Data Analysis:

A. Data Analysis using Pig

A. Use Pig Latin scripting language to investigate and understand the data and provide the following summaries: [Deliverable: for each analysis include the Pig Latin code and results]

-- Load the dataset using text file data = LOAD '/project/Step-1A/blockchain_block_data.csv' USING JsonLoaner

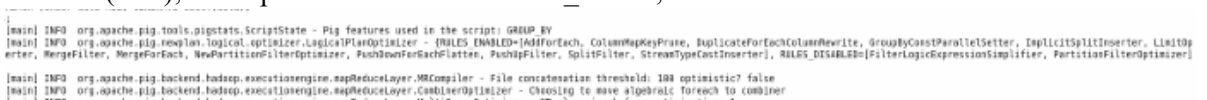
(' hash:chararray,
 ver:int,
prev_block:chararray,
 mrkl_root:chararray,
time:long,
bits:int,
 fee:long,
 nonce:long,
 n_tx:int,
 size:long,
block_index:int,
 main_chain:boolean,
height:long, weight:long ');

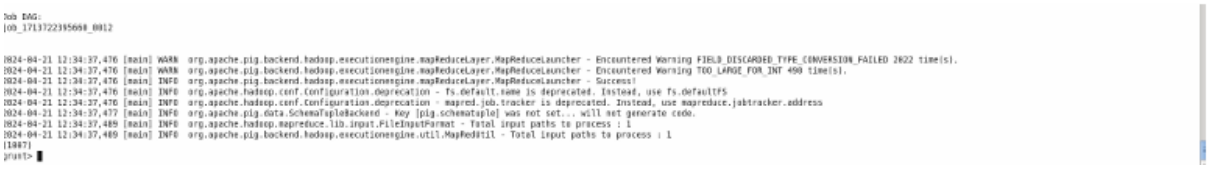
```
>> [training@localhost ~]$ pig -x mapreduce
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2024-04-19 11:52:12,746 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.3 (rexported) compiled Jun 24 2015, 19:3
6:38
2024-04-19 11:52:12,747 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/training/pig_1713552732708.log
2024-04-19 11:52:12,782 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/training/.pigbootup not found
2024-04-19 11:52:13,906 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 11:52:13,907 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:13,907 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file syst
em at: hdfs://localhost:8020
2024-04-19 11:52:16,140 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 11:52:16,140 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job t
racker at: localhost:8021
2024-04-19 11:52:16,141 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:16,269 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:16,271 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 11:52:16,397 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:16,399 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 11:52:16,549 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:16,551 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 11:52:16,654 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:16,656 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 11:52:16,803 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:16,804 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 11:52:16,951 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 11:52:16,952 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
```

1. How many total blocks are there in your dataset?

-- Loading the dataset data = LOAD '/user/training/output.csv' USING PigStorage(',') AS (hash:chararray,
ver:int,
 prev_block:chararray,
mrkl_root:chararray,
 time:long,
 bits:int,
 fee:long,
nonce:long,
 n_tx:int,
size:long,

block_index:int,
main_chain:boolean,
height:int,
weight:long);
-- 1) Count the total number of blocks total_blocks = FOREACH (GROUP data ALL) GENERATE COUNT(data); -- Output the result DUMP total_blocks;





1027

A.2 What is the largest block height among the blocks in your dataset?
-- 2) Find the largest block height max_height = FOREACH (ORDER data BY height DESC) GENERATE height; largest_block_height = LIMIT max_height 1;

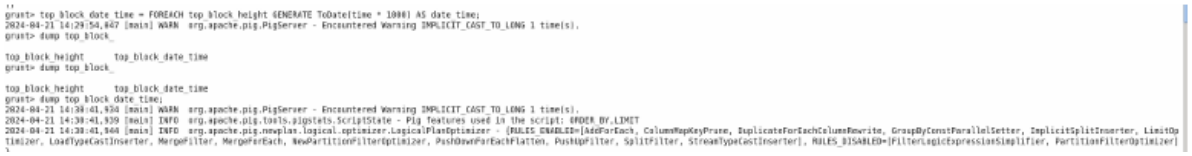
-- Output the result
DUMP largest_block_height;

835509

A.3 What is the date and time for that block?
-- 3) Find the largest block height top_block_date_time = FOREACH top_block_height GENERATE ToDate(time * 1000) AS date_time;

-- Output the result

DUMP top_block_date_time;

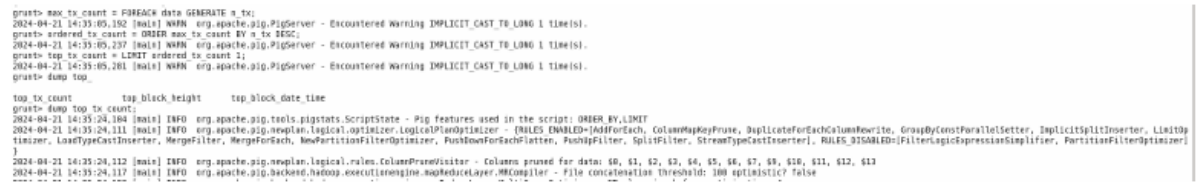


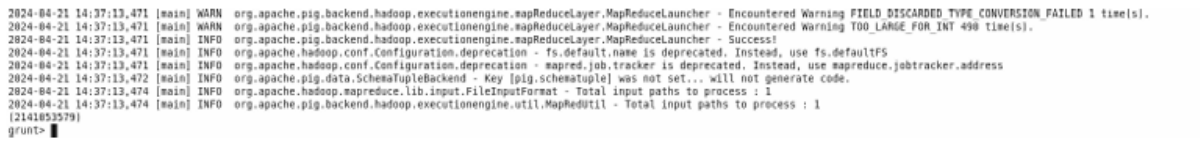
2024-04-24T22:04:56.000-08:00

A.4 What is the highest number of transactions in your blocks?

-- 4) Group data by block height and count transactions max_tx_count = FOREACH data GENERATE n_tx; ordered_tx_count = ORDER max_tx_count BY n_tx DESC; top_tx_count = LIMIT ordered_tx_count 1;

-- Output the result
DUMP top_tx_count;

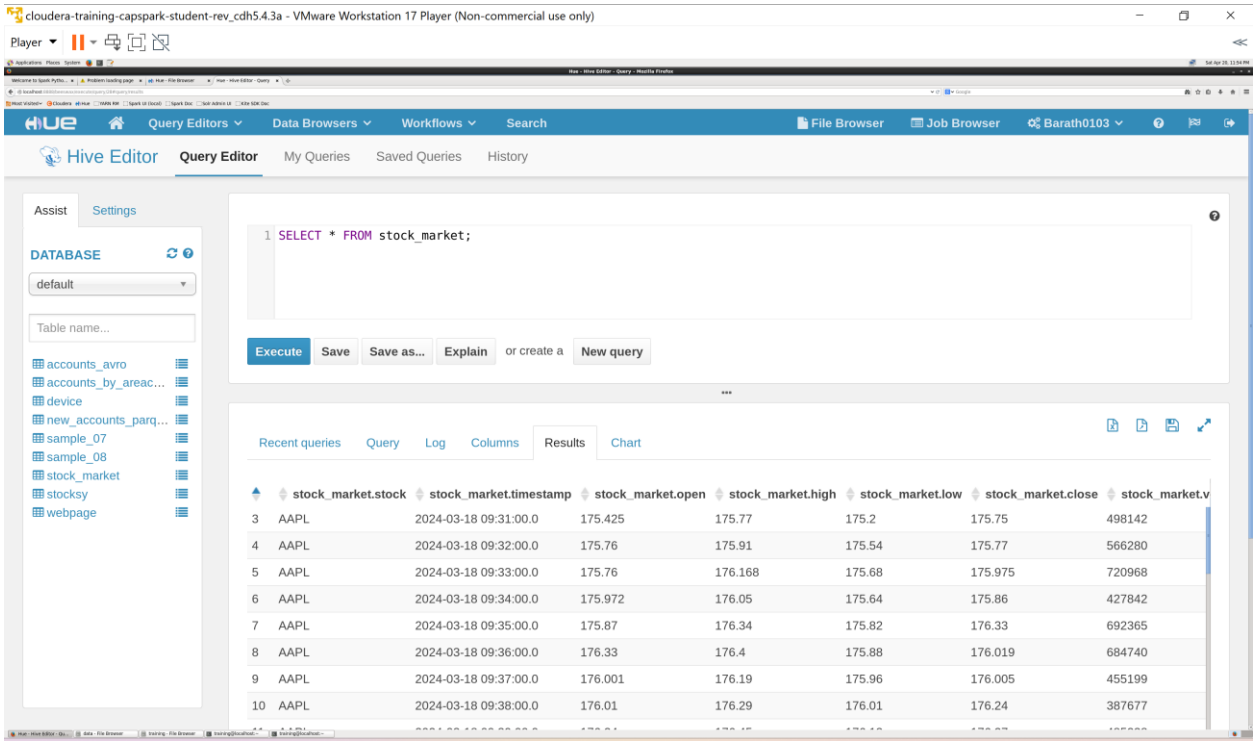




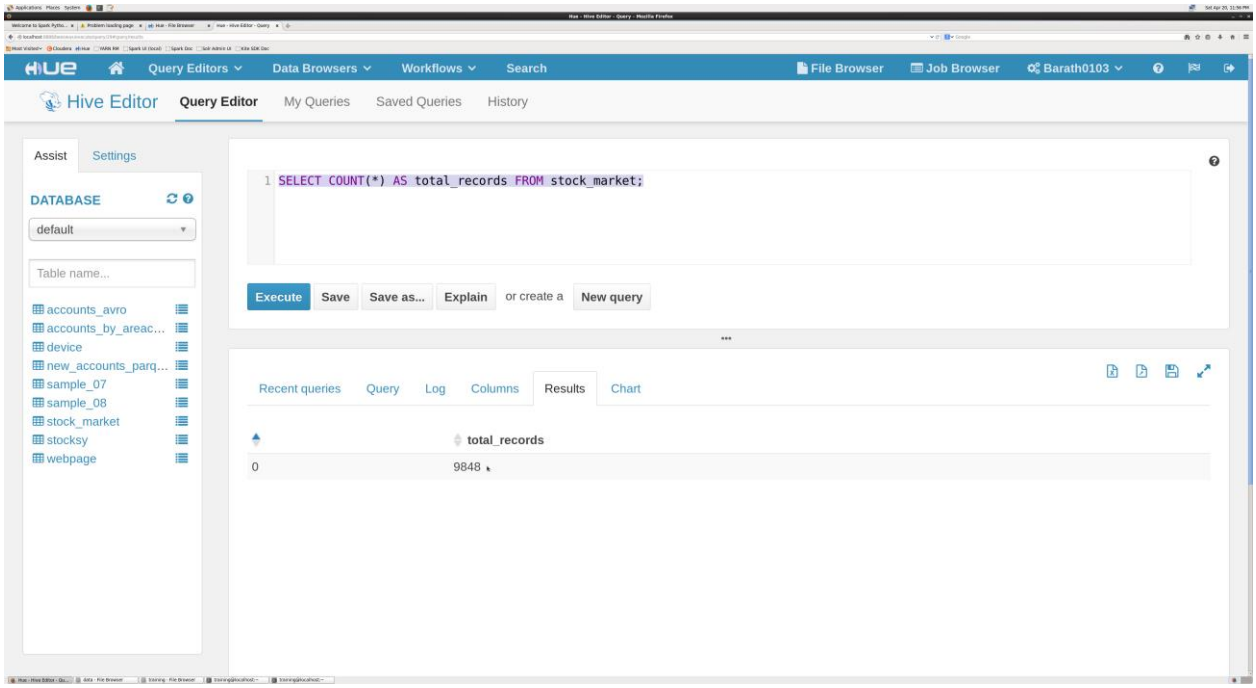
2141853579

B. Data Analysis using Hive – Part 1

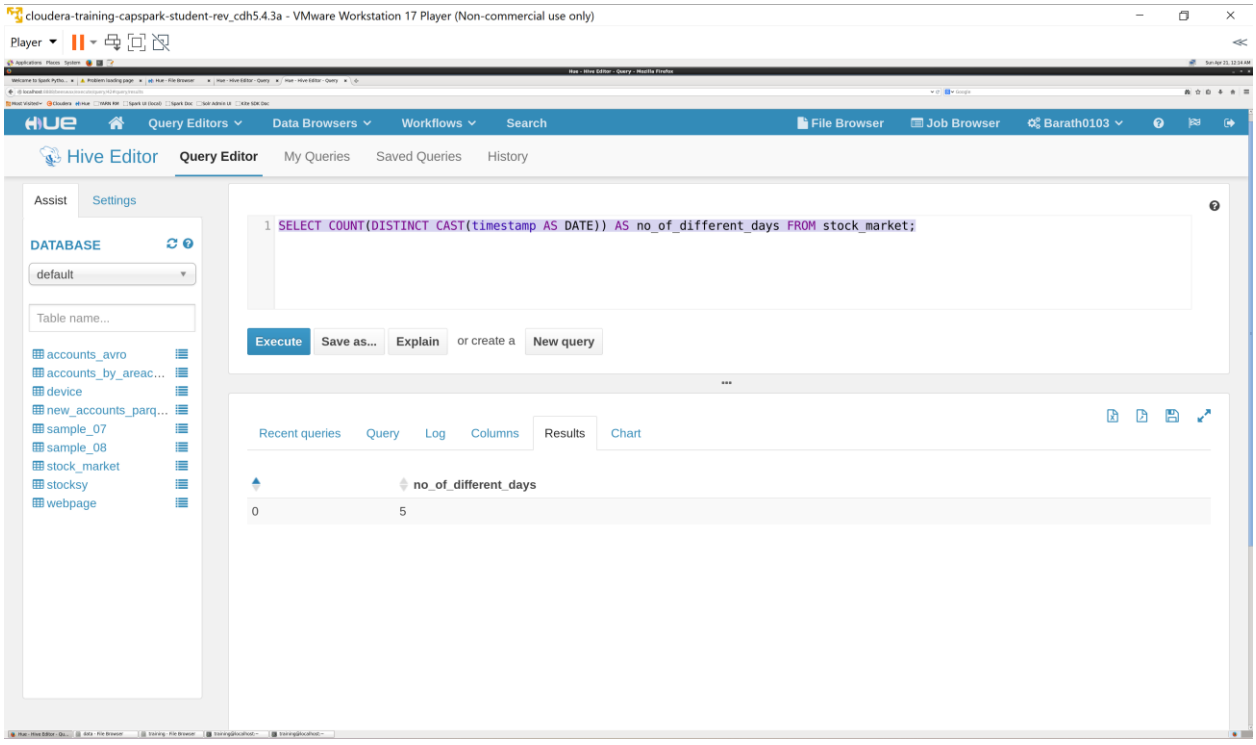
```
CREATE TABLE stock_market (
   stock STRING,
   timestamp TIMESTAMP,
   open DECIMAL(10,4),
   high DECIMAL(10,4),
   low DECIMAL(10,4),
   close DECIMAL(10,4),
   volume BIGINT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

LOAD DATA INPATH 'hdfs://localhost:8020/flume/data/stock_market_data.txt'
INTO TABLE stock_market;

SELECT * FROM stock_market;
```



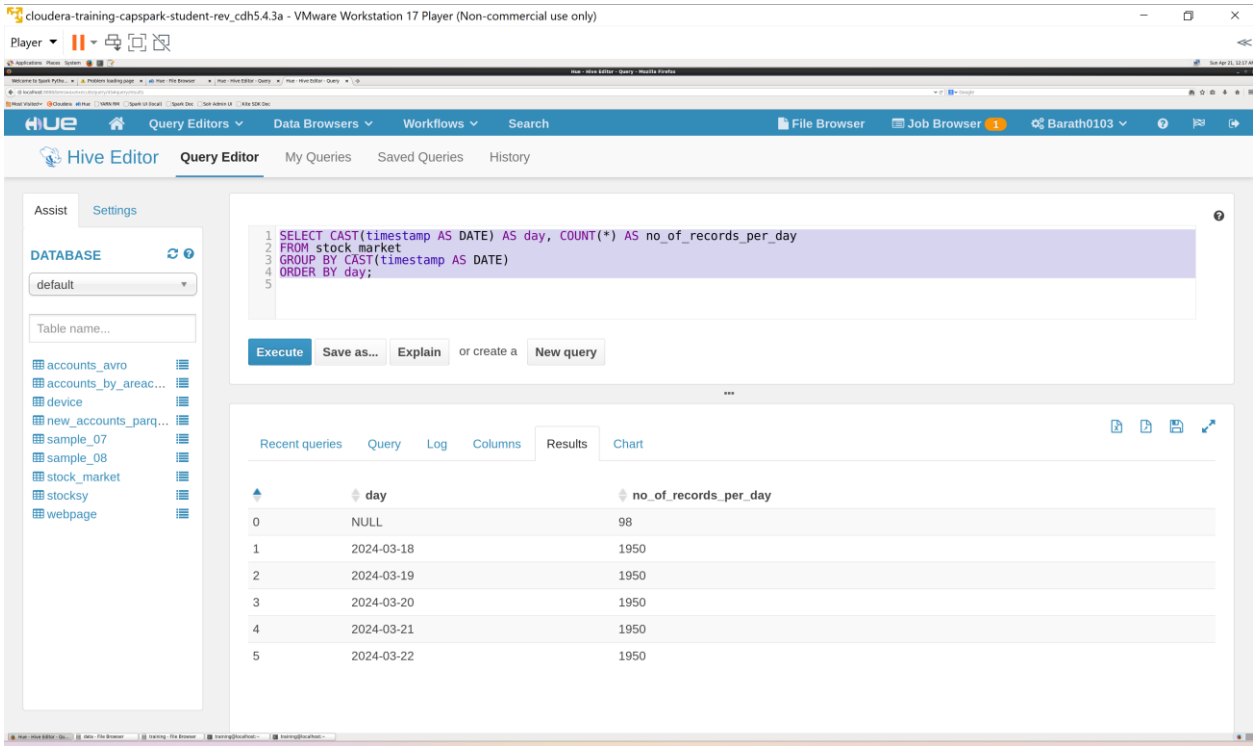
2. A. SELECT COUNT(*) AS total_records FROM stock_market;



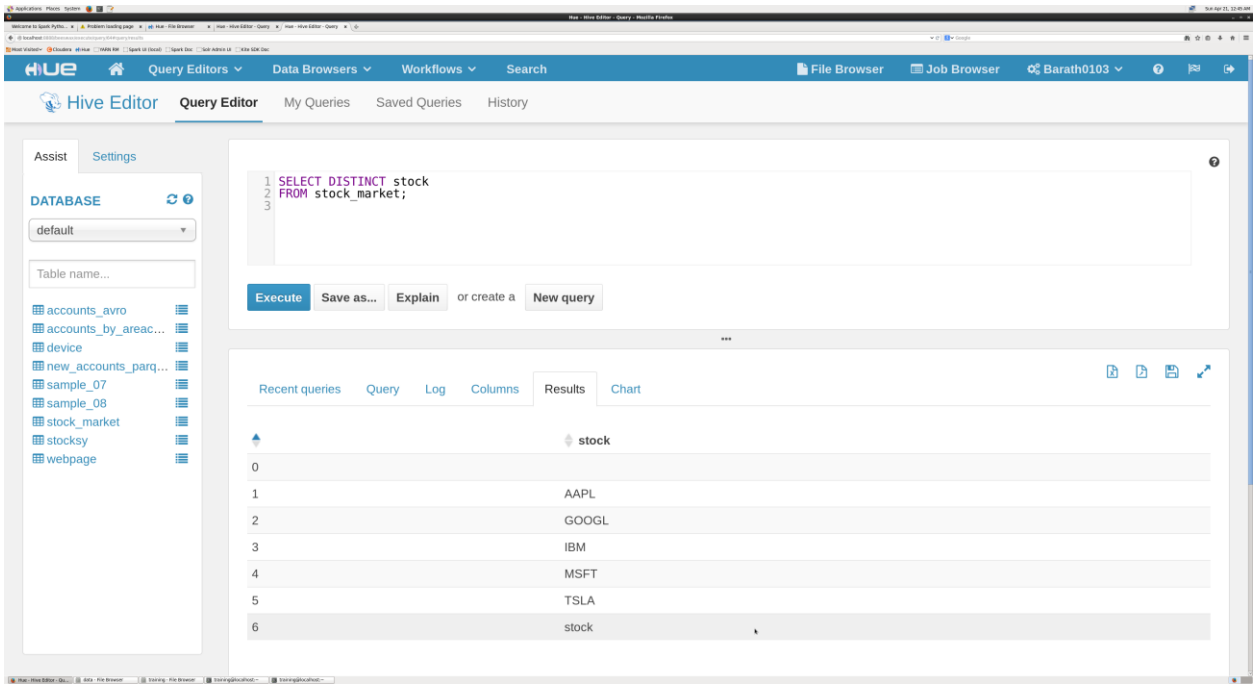
B. SELECT COUNT(DISTINCT CAST(timestamp AS DATE)) AS no_of_different_days FROM stock_market;



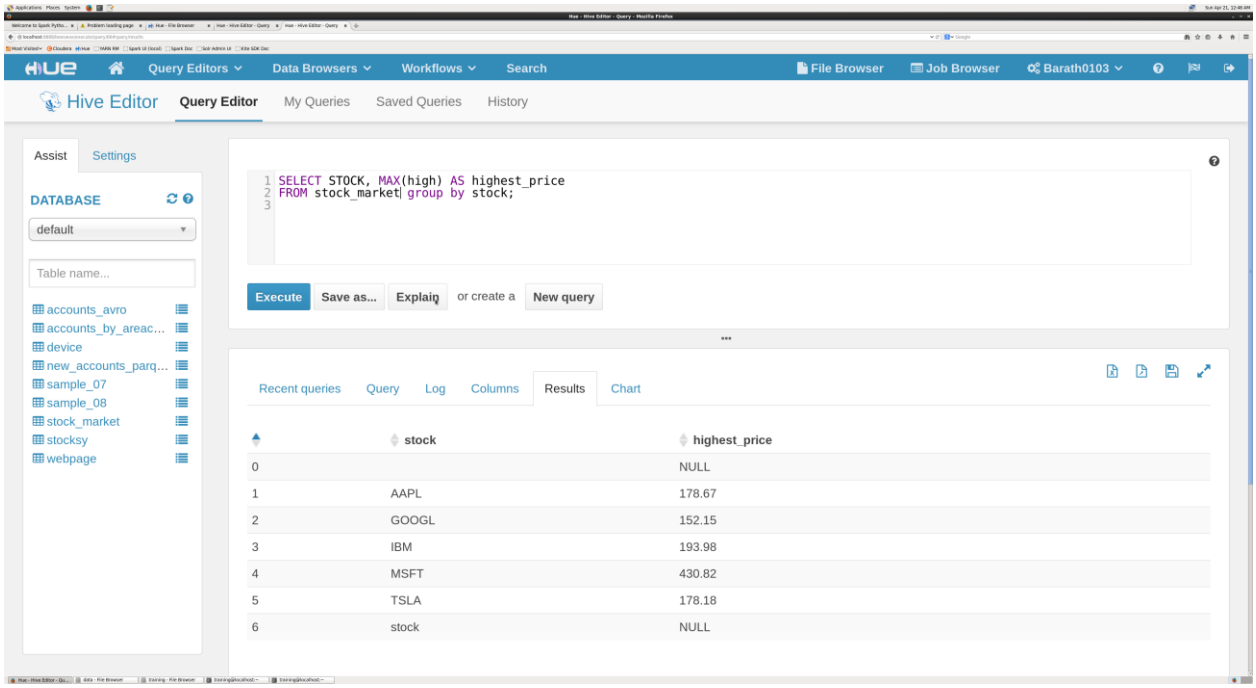
C. SELECT CAST(timestamp AS DATE) AS day, COUNT(*) AS no_of_records_per_day
FROM stock_market
GROUP BY CAST(timestamp AS DATE)
ORDER BY day;



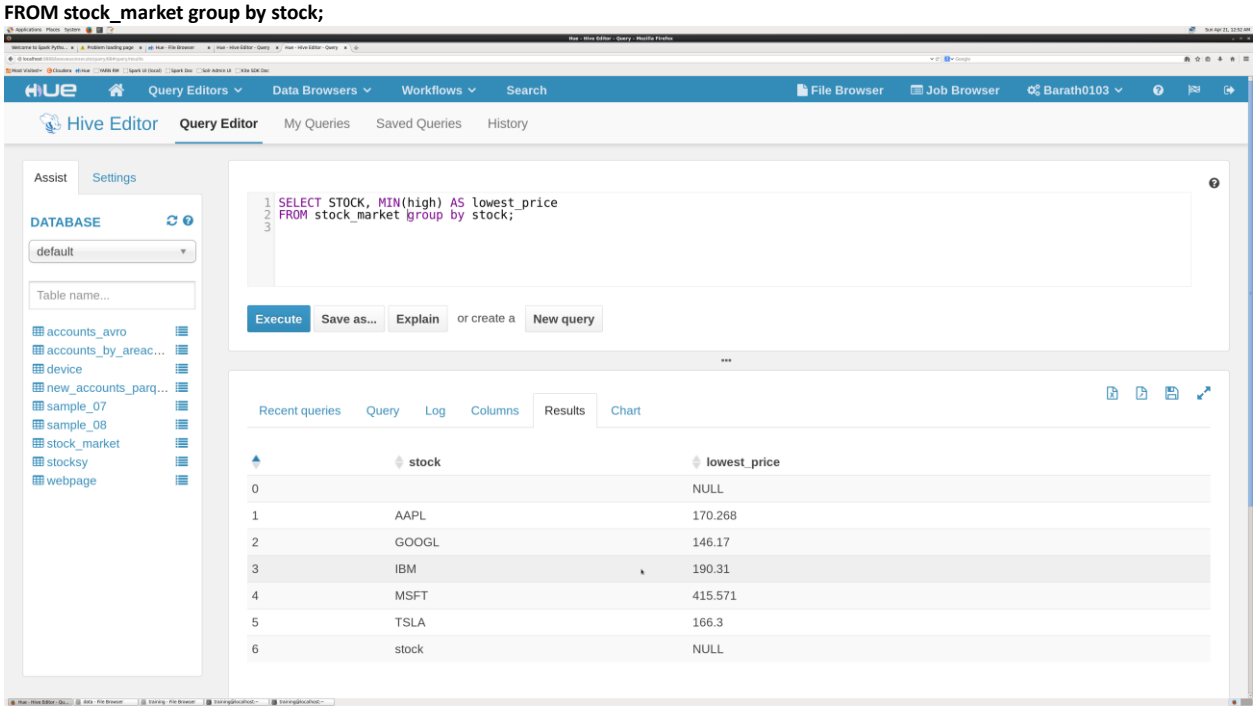
D. SELECT DISTINCT stock FROM stock_market;



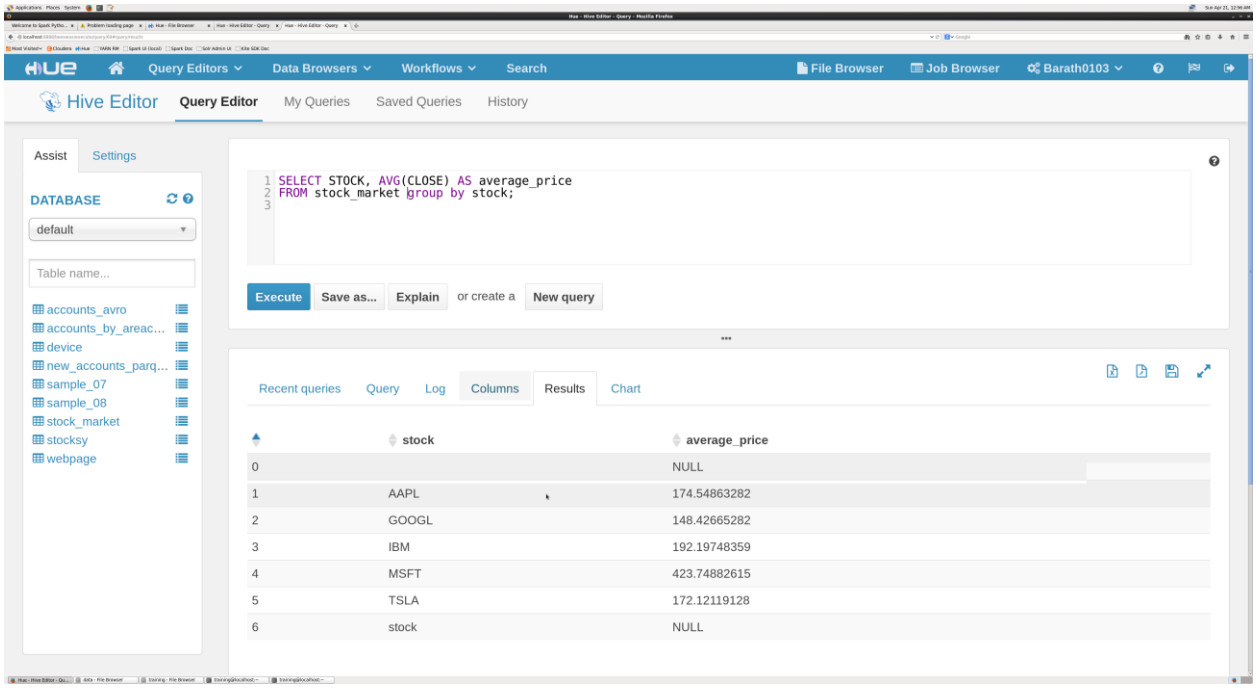
E. SELECT STOCK, MAX(high) AS highest_price
FROM stock_market group by stock;



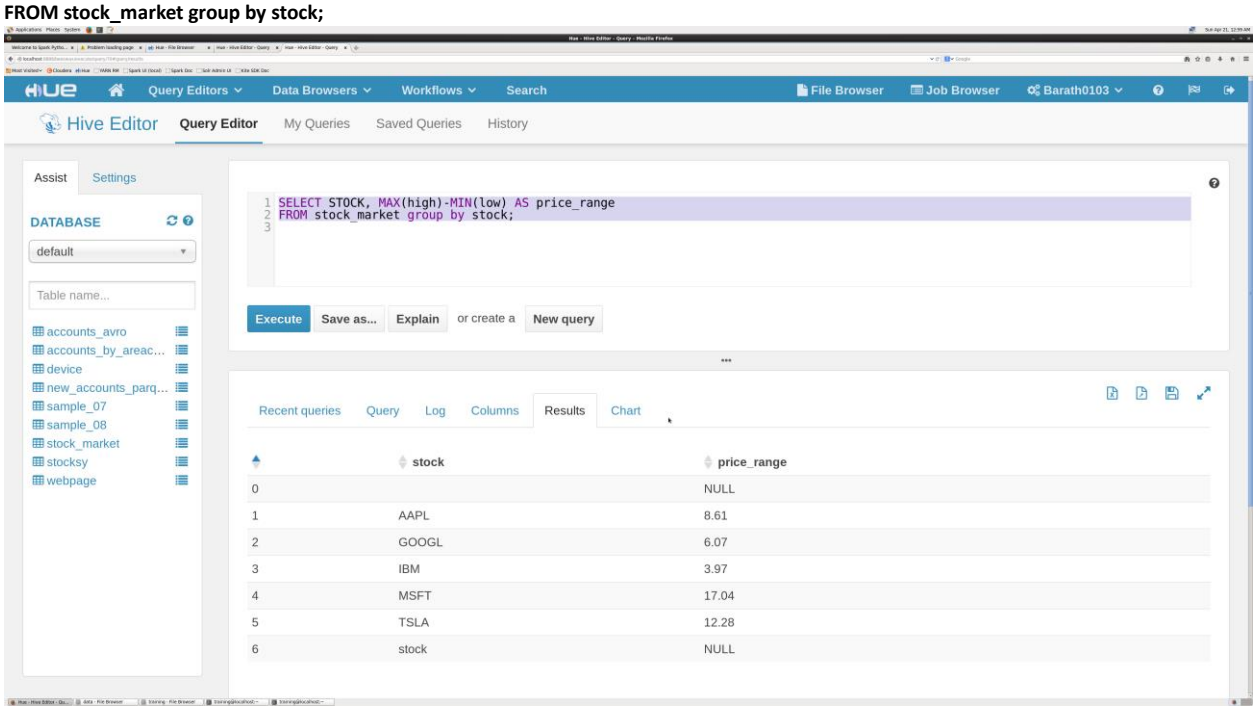
F. SELECT STOCK, MIN(high) AS lowest_price
FROM stock_market group by stock;



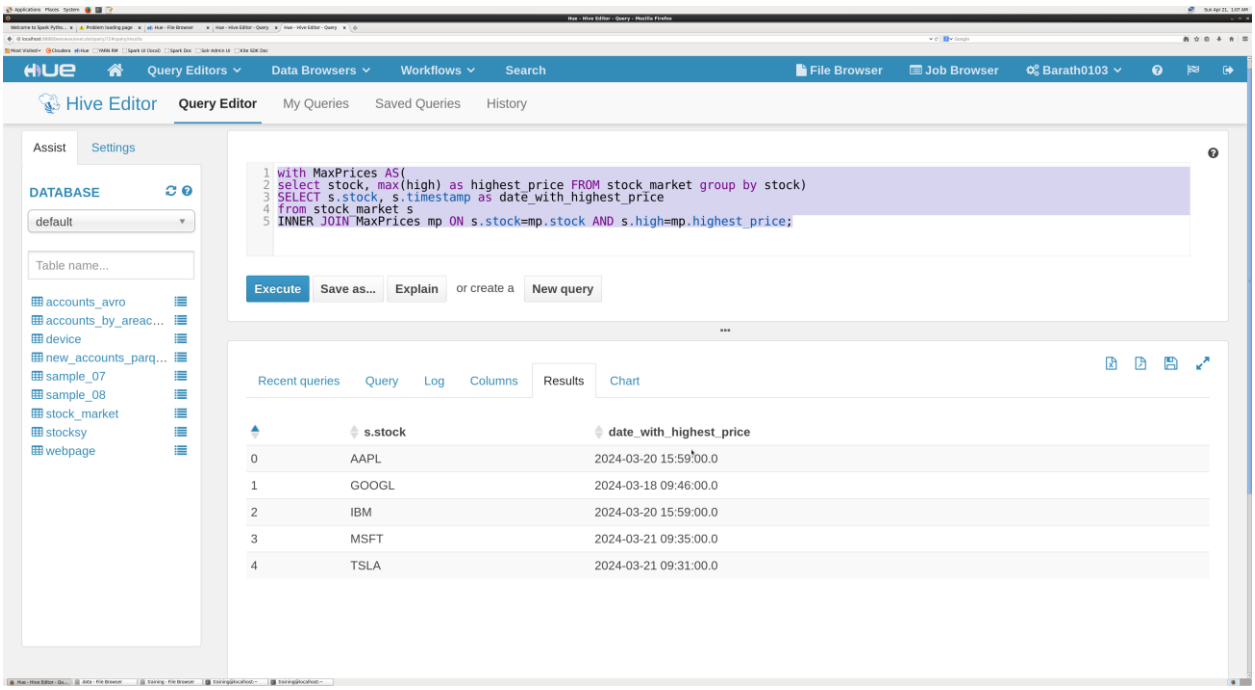
G. SELECT STOCK, AVG(CLOSE) AS average_price
FROM STOCK_MARKET
GROUP BY STOCK;



H. SELECT STOCK, MAX(high)-MIN(low) AS price_range
FROM stock_market group by stock;



I. with MaxPrices AS(
select stock, max(high) as highest_price FROM stock_market group by stock)
SELECT s.stock, s.timestamp as date_with_highest_price
from stock_market s
INNER JOIN MaxPrices mp ON s.stock=mp.stock AND s.high=mp.highest_price;



C. Data Analysis using Hive – Part 2

Use HiveQL to query the table you created in step 1.C and provide the following information:

a. How many total blocks are there in your blocks table?

-- select count(*) as total_blocks from blocks;

```
[training@localhost ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> use project1;
OK
Time taken: 2.199 seconds
hive> show tables;
OK
blocks_2023_sep_10_to_15
blocks_info_2023_sep_10_to_15
stock_data
tx_info_2023_sep_10_to_15
Time taken: 0.815 seconds, Fetched: 4 row(s)
hive> select count(*) as total_blocks from blocks_2023_Sep_10_to_15;
Query ID = training_20240420163939_114d3a71-29a1-495c-82e3-637f038af264
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1711325591056_0108, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0108/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1711325591056_0108
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-04-20 16:40:30,466 Stage-1 map = 0%,  reduce = 0%
2024-04-20 16:40:51,304 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.53 sec
2024-04-20 16:41:15,193 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.53 sec
MapReduce Total cumulative CPU time: 8 seconds 530 msec
Ended Job = job_1711325591056_0108
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.53 sec   HDFS Read: 97242 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 530 msec
OK
920
Time taken: 82.807 seconds, Fetched: 1 row(s)
hive>
```

920

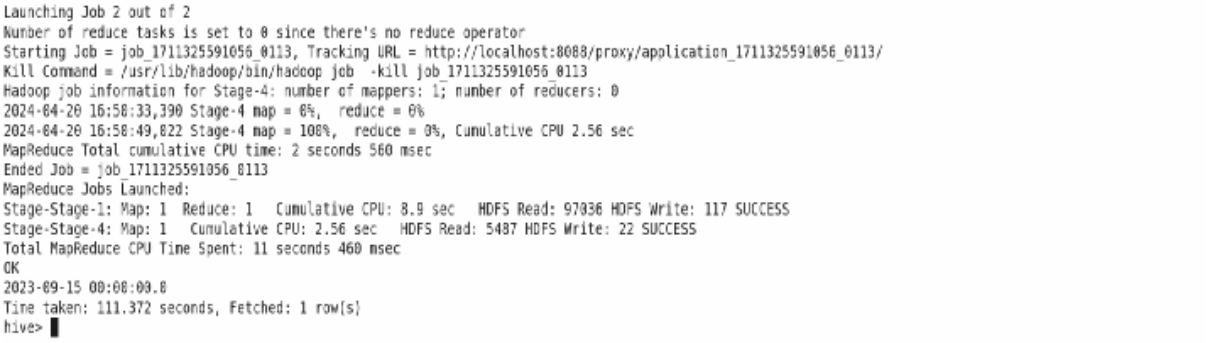
C.1.b What is the largest block height among the blocks in your blocks table?

-- SELECT MAX(height) AS largest_block_height FROM blocks_info;

807751

C.1.c What is the date and time for that block?

SELECT b.time FROM blocks b JOIN (SELECT MAX(block_index) AS max_index FROM blocks) maxblocksq ON b.block_index = maxblocksq.max_index;



d. What is the largest number of transactions in your blocks?

SELECT block_hash, COUNT(tx_hash) as num_transactions FROM tx_info_2023_Sep_10_to_15 GROUP BY block_hash ORDER BY num_transactions DESC LIMIT 1;

```
hive> SELECT block_hash, COUNT(tx_hash) as num_transactions
    > FROM tx_info_2023_Sep_10_to_15
    > GROUP BY block_hash
    > ORDER BY num_transactions
    > DESC LIMIT 1;
Query ID = training_20240420165252_e2e76a65-b905-43e2-84d9-96e2cd55150d
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1711325591056_0114, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0114/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1711325591056_0114
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 2
2024-04-20 16:53:30,091 Stage-1 map = 0%,  reduce = 0%
2024-04-20 16:54:06,515 Stage-1 map = 17%,  reduce = 0%, Cumulative CPU 10.34 sec
2024-04-20 16:54:08,950 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 12.31 sec
2024-04-20 16:54:12,467 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 14.17 sec
2024-04-20 16:54:13,934 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.16 sec
2024-04-20 16:54:43,880 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 19.63 sec
2024-04-20 16:55:11,391 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 25.14 sec
MapReduce Total cumulative CPU time: 25 seconds 140 msec
Ended Job = job_1711325591056_0114
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1711325591056_0115, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0115/
```

```
Starting Job = job_1711325591056_0115, Tracking URL = http://localhost:8088/proxy/application_17113255910
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1711325591056_0115
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-04-20 16:55:48,893 Stage-2 map = 0%,  reduce = 0%
2024-04-20 16:56:15,135 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.51 sec
2024-04-20 16:56:43,244 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.27 sec
MapReduce Total cumulative CPU time: 8 seconds 270 msec
Ended Job = job_1711325591056_0115
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 2   Cumulative CPU: 25.75 sec   HDFS Read: 256690135 HDFS Write: 26697 SUC
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 8.27 sec   HDFS Read: 31420 HDFS Write: 70 SUCCESS
Total MapReduce CPU Time Spent: 34 seconds 20 msec
OK
000000000000000000002dbf9d0bc1c743ac17bdb60d5c6abc8cd94f2d253621d        7252
Time taken: 236.629 seconds, Fetched: 1 row(s)
hive>
```