Big Data Project

BUAN 6346.504

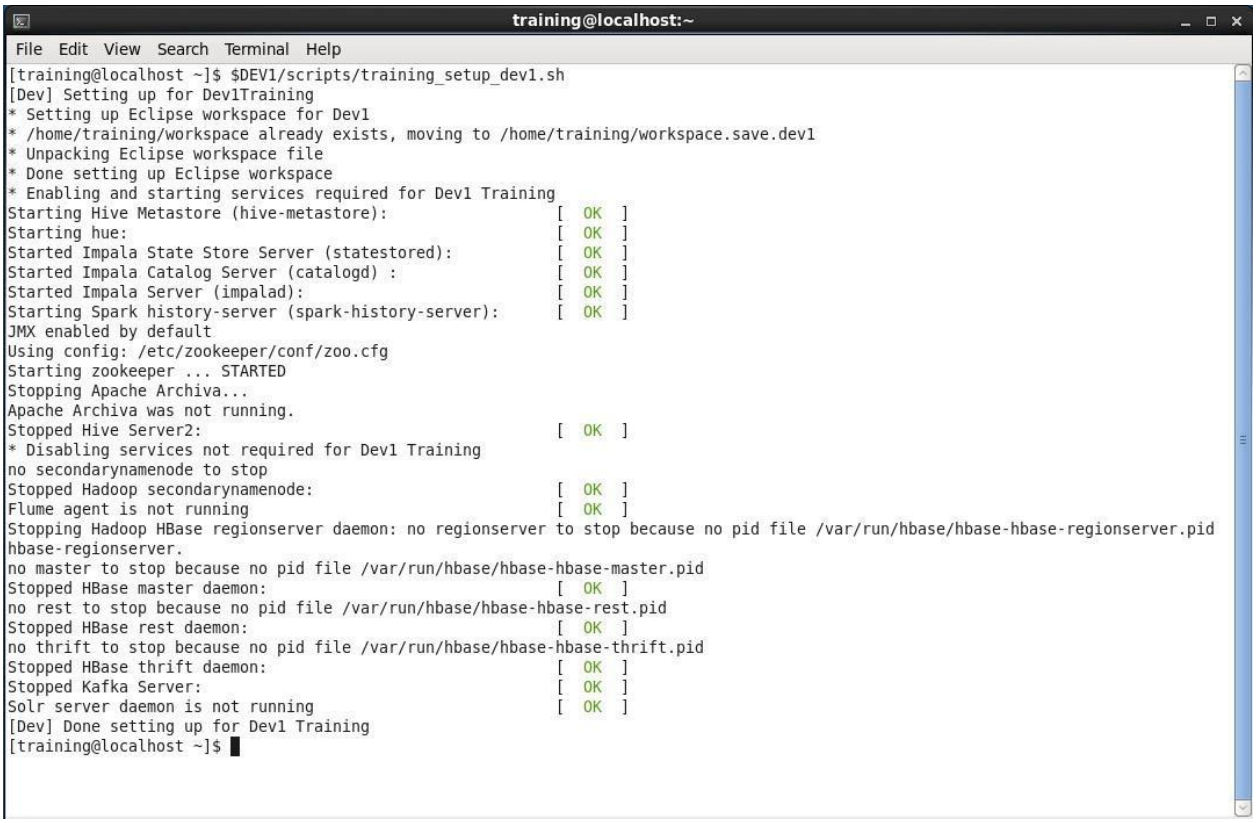
Group 14

Group Members:

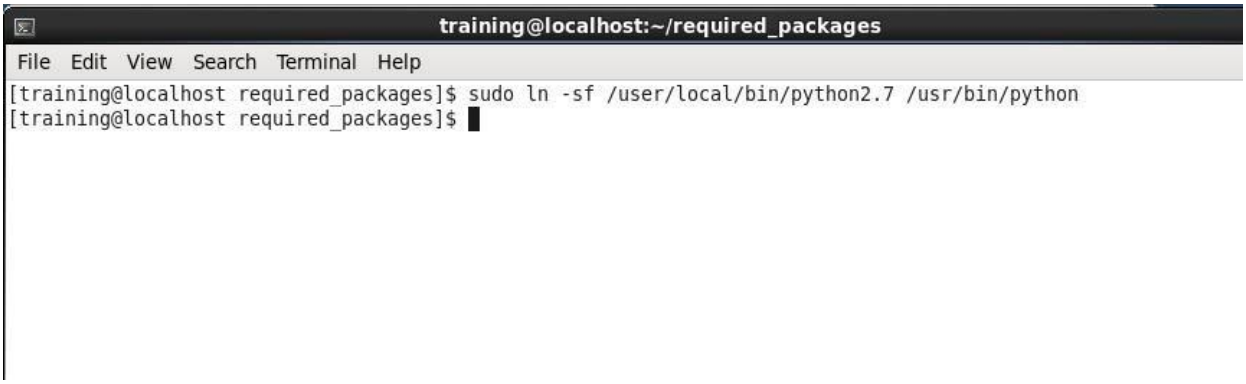
- Sai Madhan Muthyam
- Amol Bhadane
- Venkata Rama Anirudh Vikram Kesapragada
- Karan Raturi
- Lakshmi Vinanya Yenumula
- Neelima Nandigam

Step 0: VM Setup

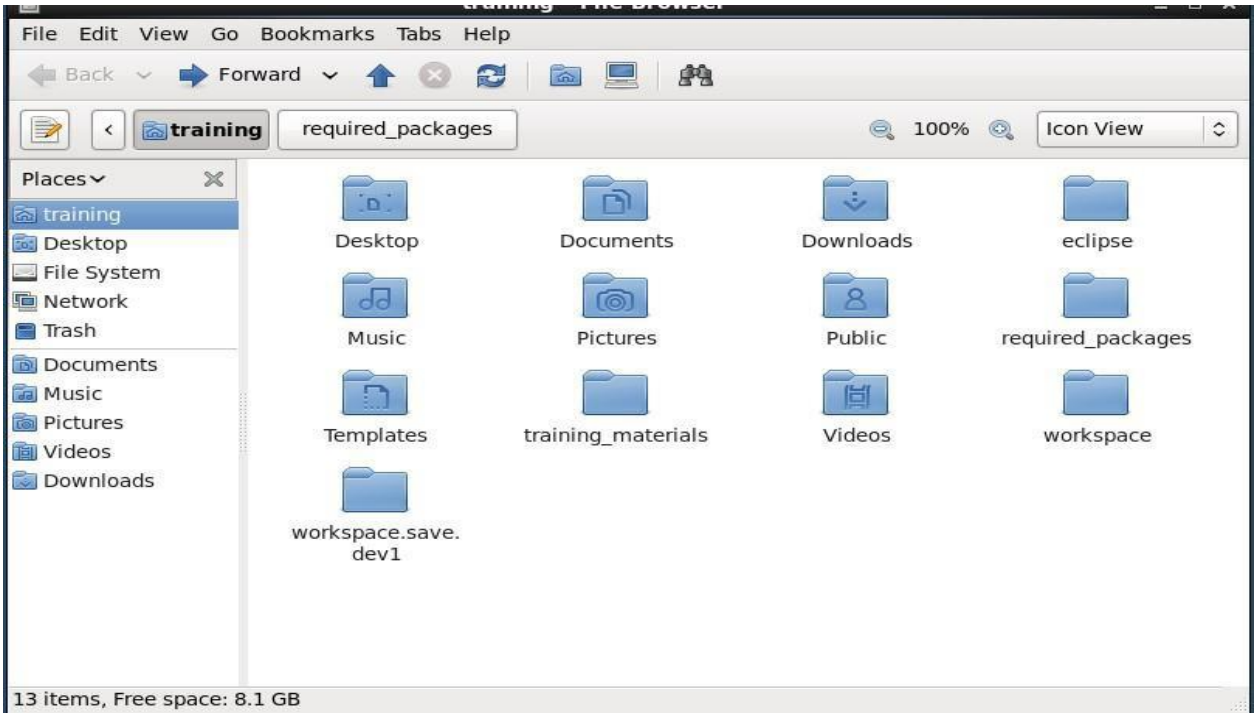
2. Follow assignment 1 to enable all the services and set up your Hadoop.



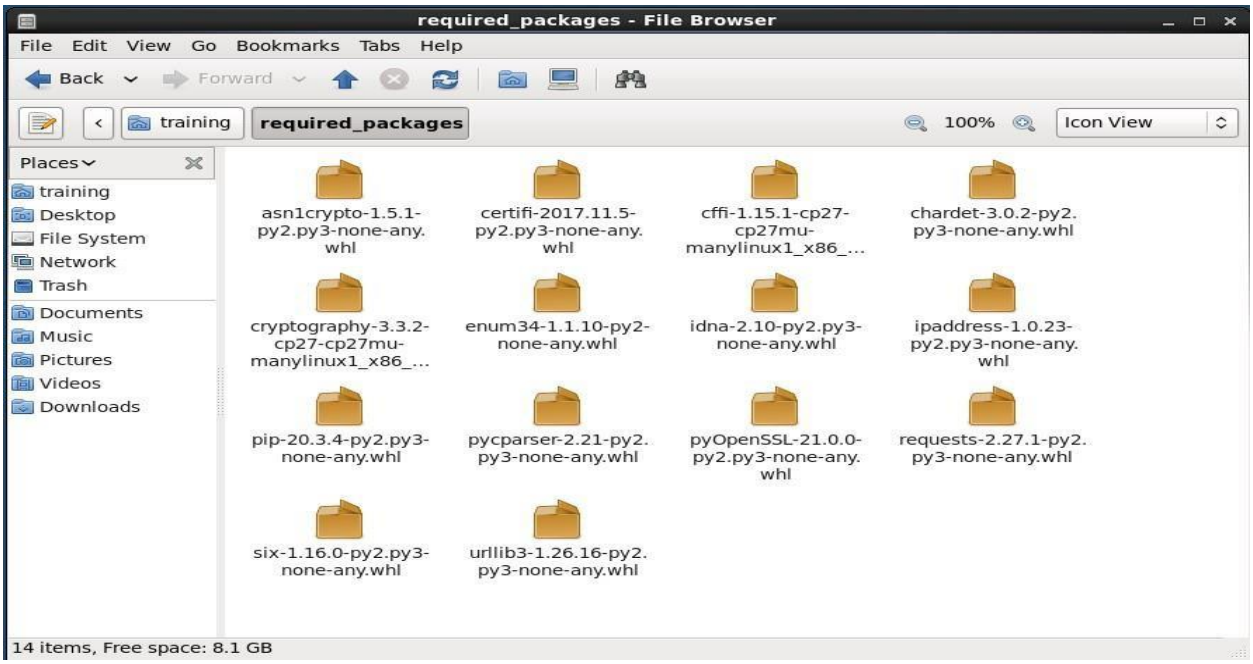
3. Change the Python default interpreter to Python2.7 in your VM by executing the following command in your Terminal:



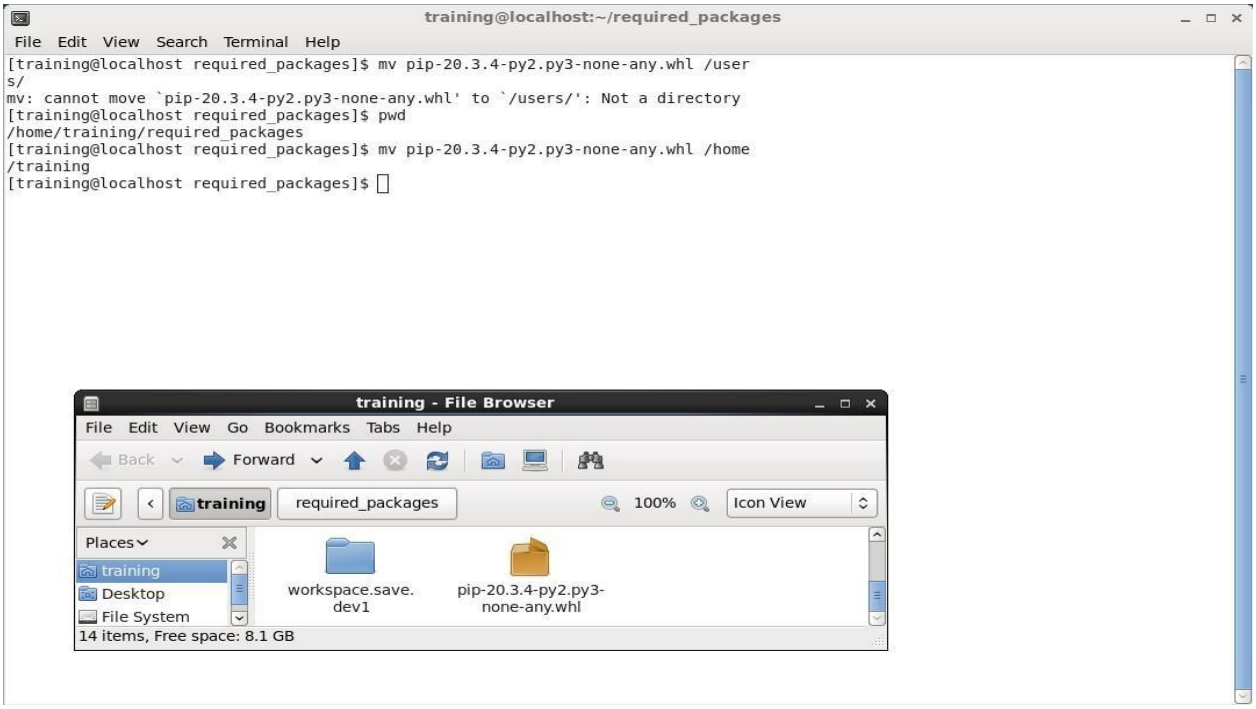
5. Create a directory in the user directory of your VM and call it required_packages: /home/training/required_packages



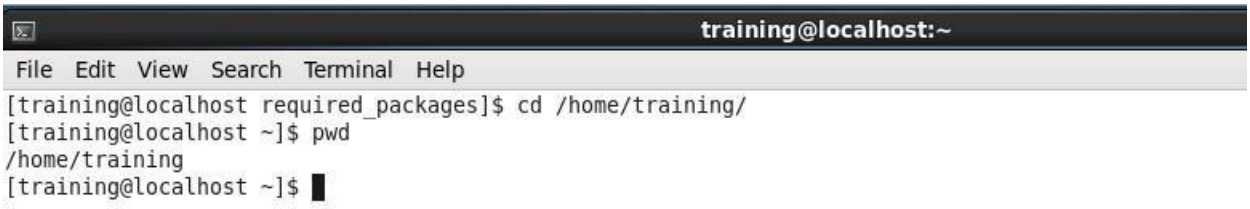
6. Copy the contents of the Required Packages on eLearning to your VM to the directory you created in the previous step (required_packages)



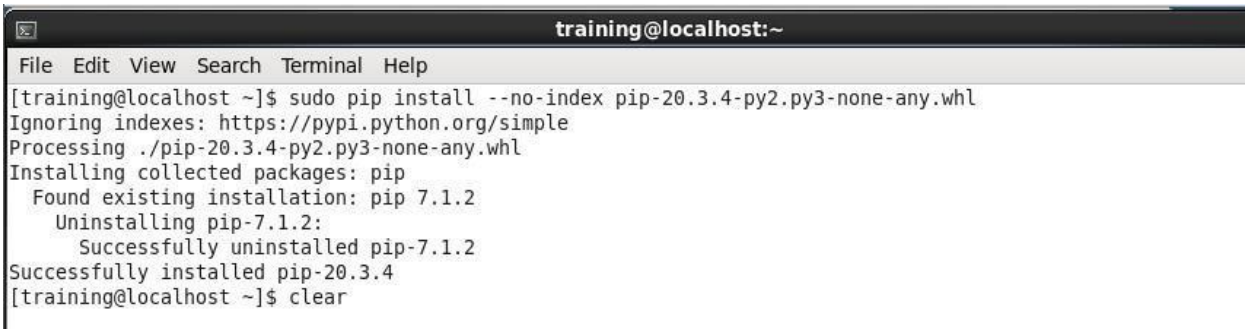
7. Move the file pip-20.3.4-py2.py3-none-any.whl from required_packages to the following directory:



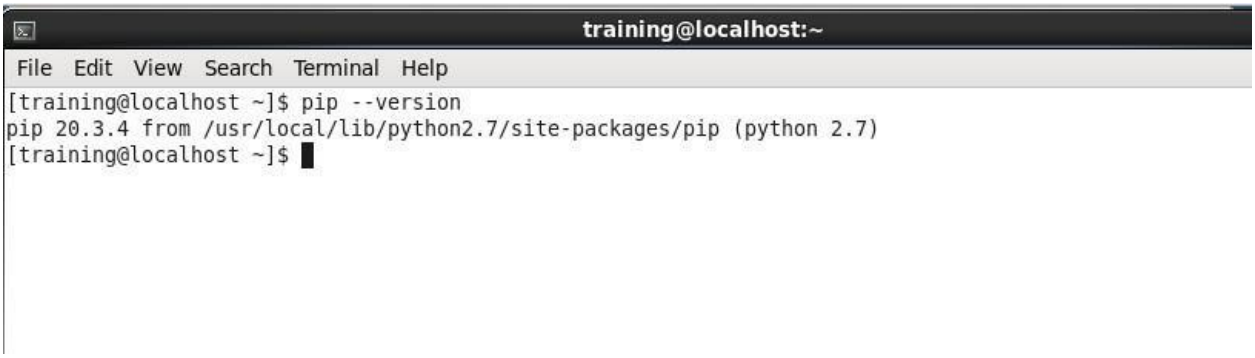
8. Change your working directory to /home/training



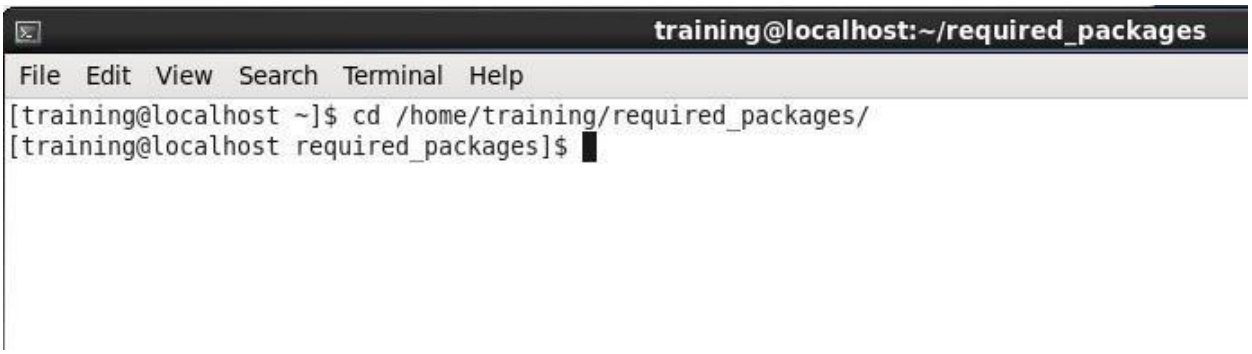
9. Upgrade the pip on your machine by executing the following command in Terminal:



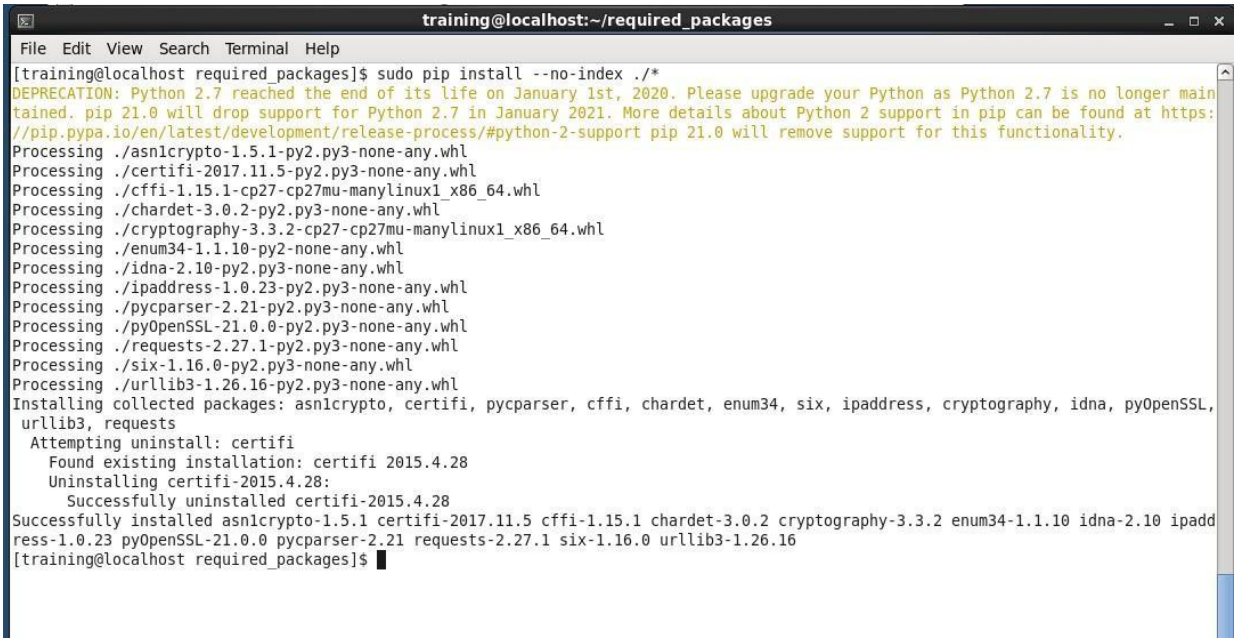
10. Verify that the pip has been upgraded by executing the following:



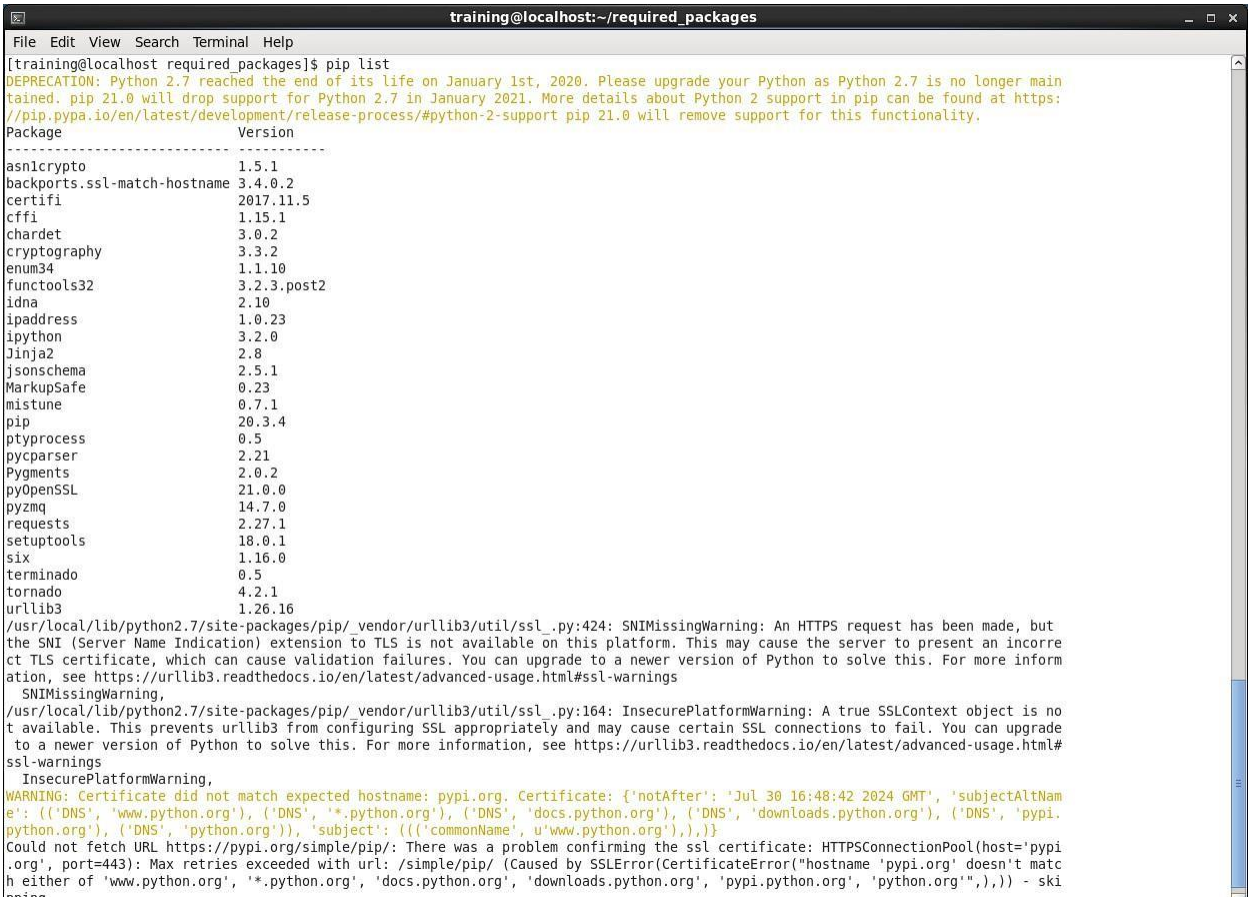
11. Change your directory to /home/training/required_packages



12. Install all the packages in the directory by executing the following in Terminal:



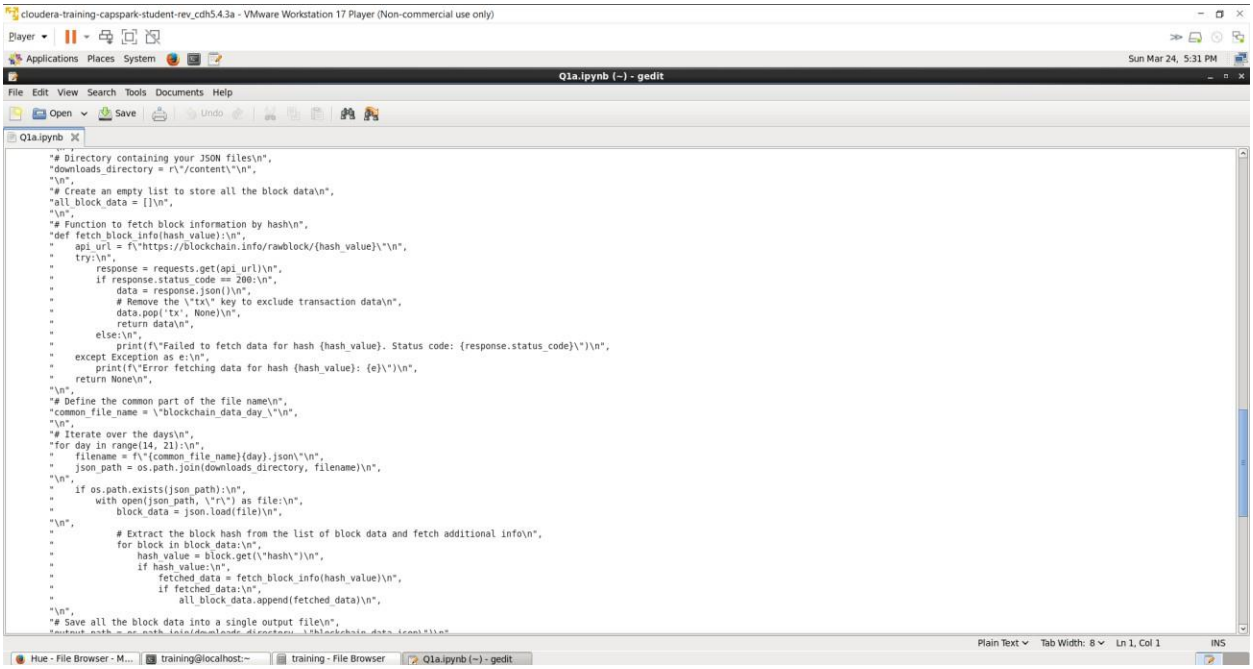
13. Verify that all the packages have been installed by running the following:



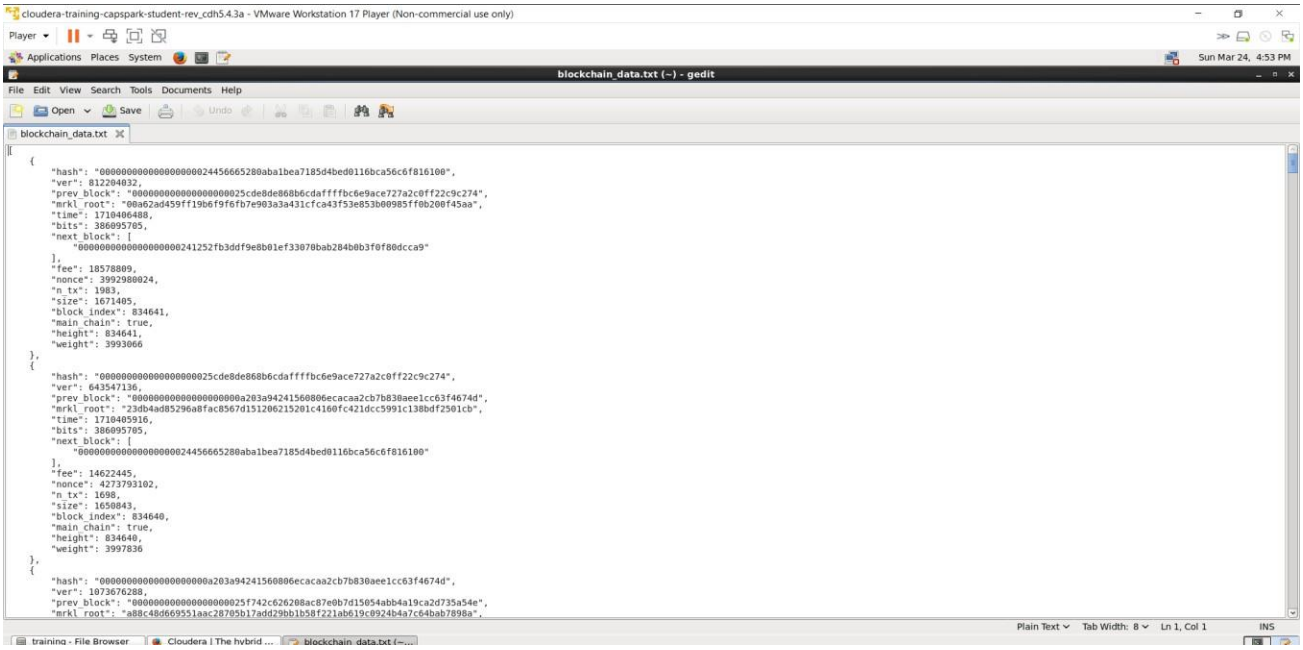
Step 1: Data Ingestion

A. Direct File Transfer

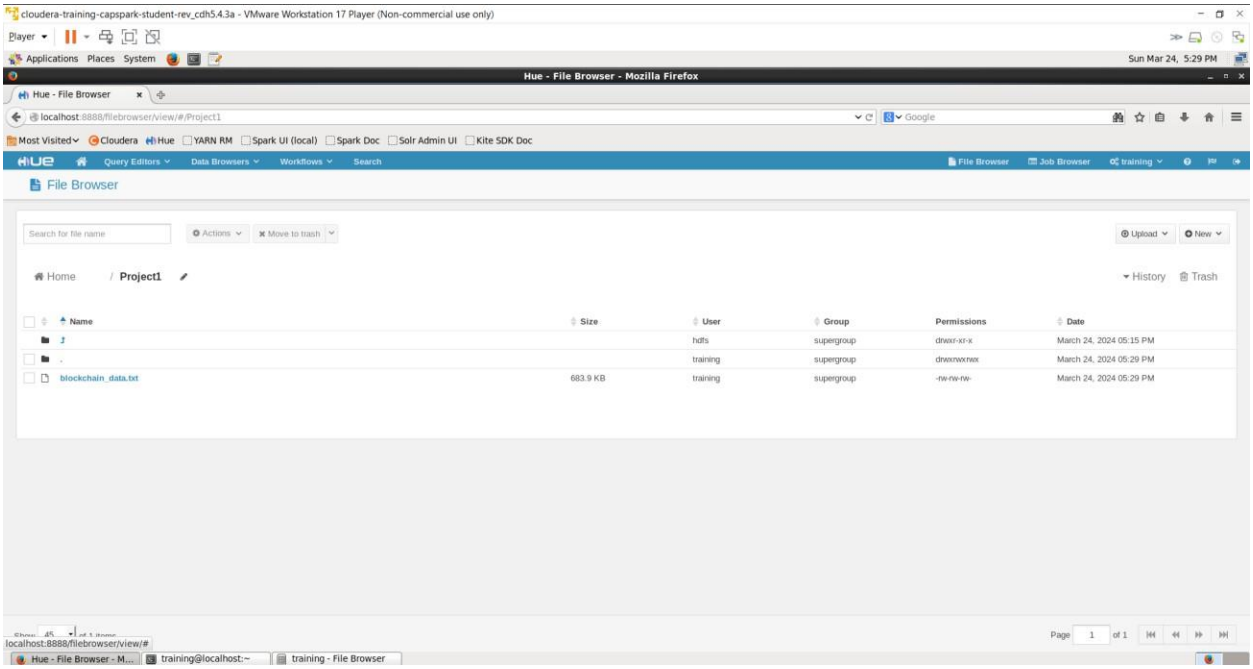
1. Setup your Python code using the following information to pull data from the API and store it in a text file on your local storage.



e. Your code should store all this information in a text file on your local storage [Deliverable: screenshots of pulled data in a text file]



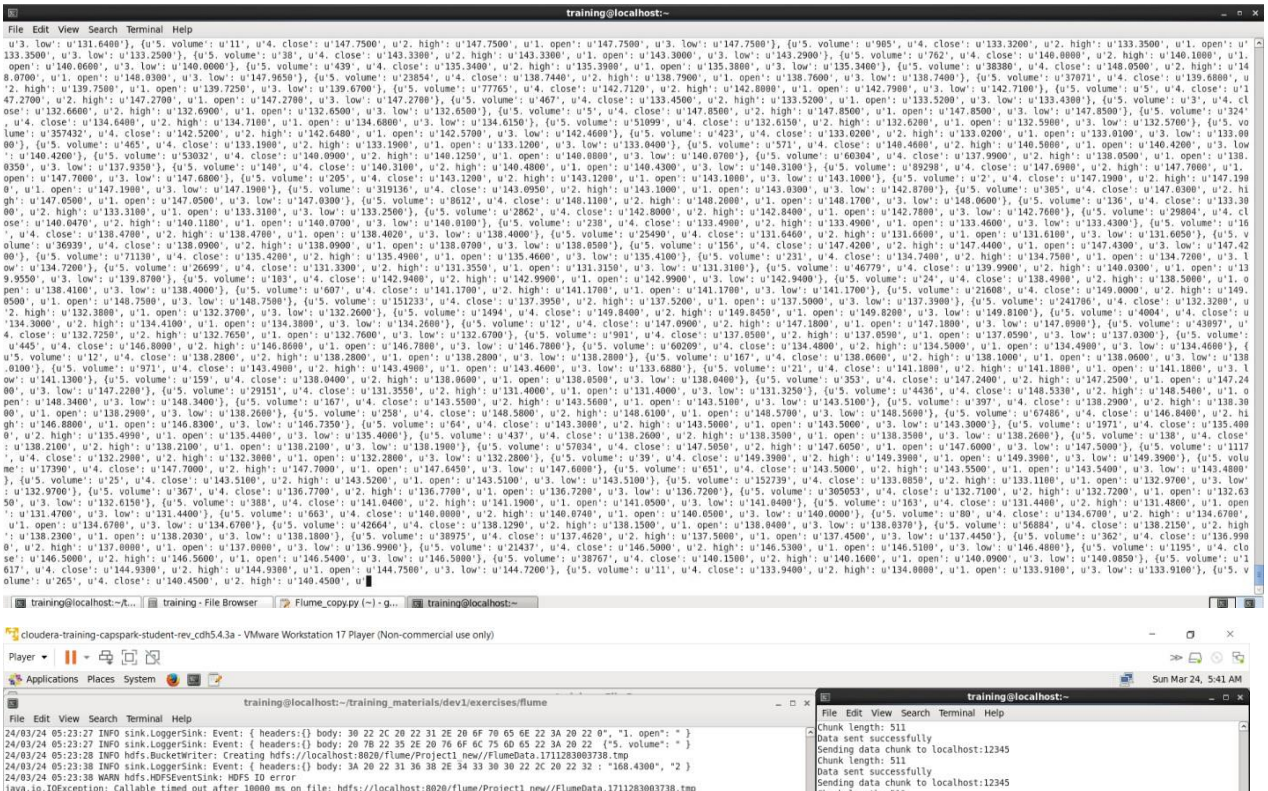
2. Transfer the data from your VM's local storage to HDFS [Deliverable: screenshot of data in HDFS]



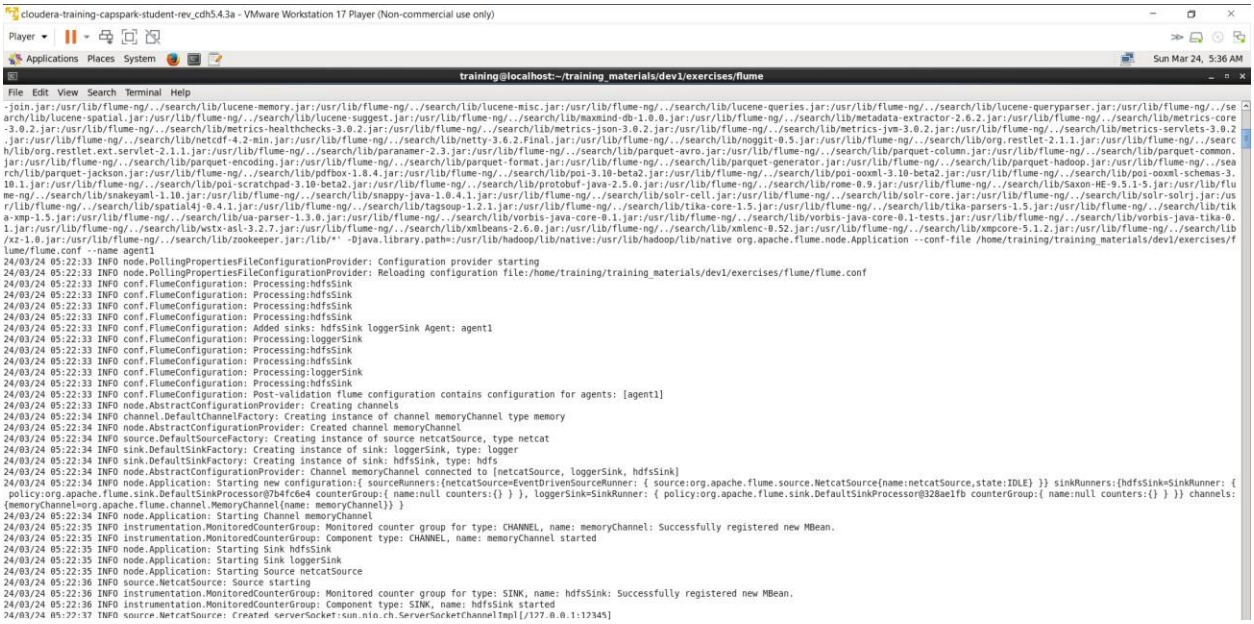
B. Stream Ingestion using Flume:

2. Setup your Python code using the following information to pull data from the API and send it to localhost on a specific port

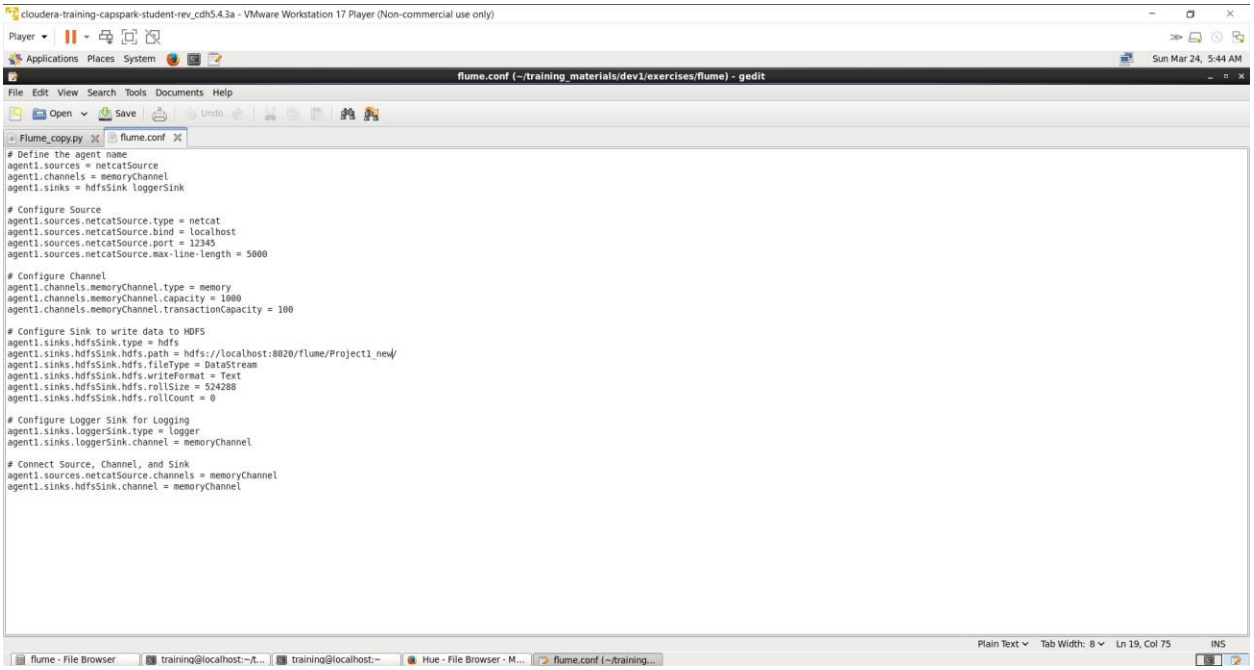
d. Your code should print this data in the output



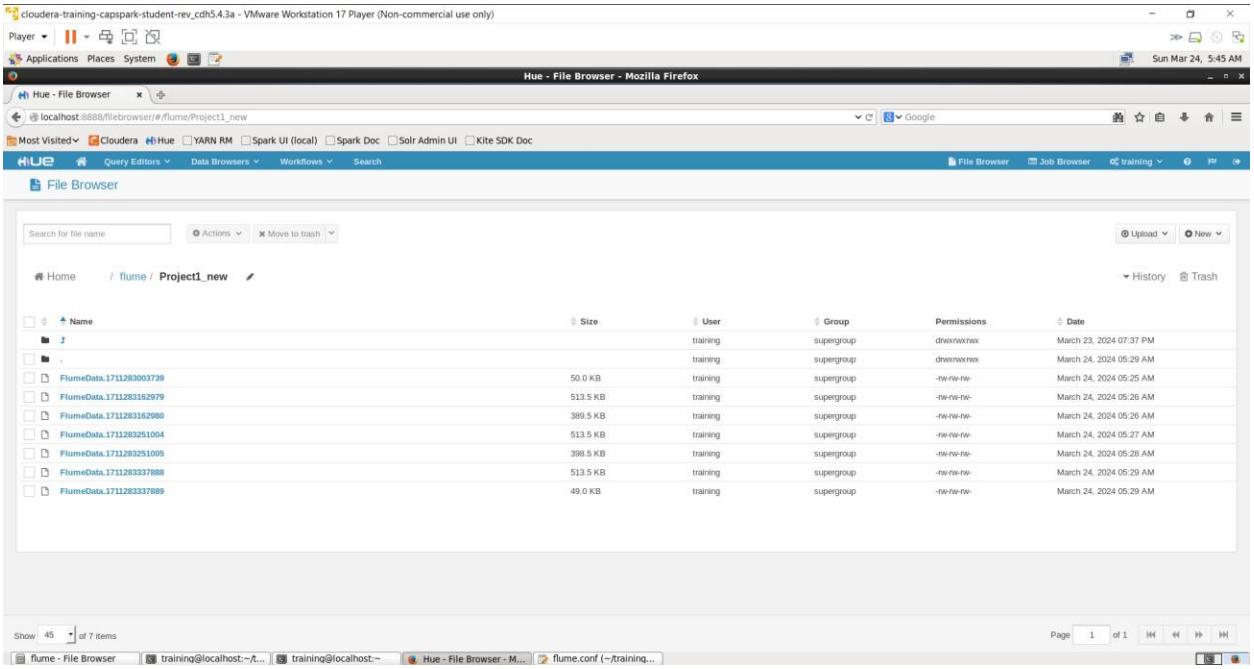
3. Setup and configure you Flume agent so that it captures the data from the host localhost and stores it on HDFS



a. Configure Flume by creating a Flume configuration file with the characteristics mentioned below. [Deliverable: Flume configuration]



e. The data should be directly stored on HDFS [Deliverable: screenshot of data in HDFS]



C. Data Ingestion using Sqoop:

2. Upload each file from your VM's local storage to MySQL database on your VM

CREATE TABLE blocks_2023_Sep_10_to_15 (

id INT PRIMARY KEY,

hash VARCHAR(255),

time DATETIME,

block_index INT

);

```
mysql> use project1;
Database changed
mysql> create table blocks_2023_Sep_10_to_15(
    -> id INT PRIMARY KEY,
    -> hash VARCHAR(255),
    -> time DATETIME,
    -> block_index INT
    -> );
Query OK, 0 rows affected (0.04 sec)
```

LOAD DATA LOCAL INFILE
'/home/training/Desktop/Project1/Project1Data/blocks_2023_Sep_10_to_15.csv'

INTO TABLE blocks_2023_Sep_10_to_15

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

```
mysql> LOAD DATA LOCAL INFILE '/home/training/sqoop_datasets/blocks_2023_Sep_10_to_15.csv'
    -> INTO TABLE blocks_2023_Sep_10_to_15
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 920 rows affected (0.05 sec)
Records: 920  Deleted: 0  Skipped: 0  Warnings: 0
```

CREATE TABLE blocks_info_2023_Sep_10_to_15 (

id INT PRIMARY KEY,

hash CHAR(64) NOT NULL,

ver INT NOT NULL,

bits INT NOT NULL,

fee INT NOT NULL,

nonce BIGINT UNSIGNED NOT NULL,

size INT NOT NULL,

block_index INT NOT NULL,

main_chain BOOLEAN NOT NULL,

height INT NOT NULL,

weight INT NOT NULL

);

```
mysql> create table blocks_info_2023_Sep_10_to_15 (
    -> id INT PRIMARY KEY,
    -> hash CHAR(64) NOT NULL,
    -> ver INT NOT NULL,
    -> bits INT NOT NULL,
    -> fee INT NOT NULL,
    -> nonce BIGINT UNSIGNED NOT NULL,
    -> size INT NOT NULL,
    -> block_index INT NOT NULL,
    -> main_chain BOOLEAN NOT NULL,
    -> height INT NOT NULL,
    -> weight INT NOT NULL
    -> );
Query OK, 0 rows affected (0.04 sec)
```

LOAD DATA LOCAL INFILE
'/home/training/Desktop/Project1/Project1Data/blocks_info_2023_Sep_10_to_15.csv'

INTO TABLE blocks_info_2023_Sep_10_to_15

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

```
mysql> LOAD DATA LOCAL INFILE '/home/training/sqoop_datasets/blocks_info_2023_Sep_10_to_15.csv'
    -> INTO TABLE blocks_info_2023_Sep_10_to_15
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 310 rows affected, 310 warnings (0.03 sec)
Records: 310  Deleted: 0  Skipped: 0  Warnings: 0
```

```
CREATE TABLE tx_info_2023_Sep_10_to_15 (
id INT PRIMARY KEY,
tx_hash CHAR(64) NOT NULL,
block_hash CHAR(64) NOT NULL,
ver INT NOT NULL,
vin_sz INT NOT NULL,
vout_sz INT NOT NULL,
size INT NOT NULL,
weight INT NOT NULL,
fee BIGINT NOT NULL,
relayed_by VARCHAR(15) NOT NULL,
lock_me BIGINT NOT NULL,
tx_index BIGINT NOT NULL,
double_spend BOOLEAN NOT NULL,
me BIGINT NOT NULL,
block_index INT NOT NULL,
block_height INT NOT NULL
);
```

```
mysql> create table tx_info_2023_Sep_10_to_15 (
    -> id INT PRIMARY KEY,
    -> tx_hash CHAR(64) NOT NULL,
    -> block_hash CHAR(64) NOT NULL,
    -> ver INT NOT NULL,
    -> vin_sz INT NOT NULL,
    -> vout_sz INT NOT NULL,
    -> size INT NOT NULL,
    -> weight INT NOT NULL,
    -> fee BIGINT NOT NULL,
    -> relayed_by VARCHAR(15) NOT NULL,
    -> lock_time BIGINT NOT NULL,
    -> tx_index BIGINT NOT NULL,
    -> double_spend BOOLEAN NOT NULL,
    -> time BIGINT NOT NULL,
    -> block_index INT NOT NULL,
    -> block_height INT NOT NULL
    -> );
Query OK, 0 rows affected (0.01 sec)
```

LOAD DATA LOCAL INFILE
'/home/training/Desktop/Project1/Project1Data/tx_info_2023_Sep_10_to_15.csv'

INTO TABLE tx_info_2023_Sep_10_to_15

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

```
mysql> LOAD DATA LOCAL INFILE '/home/training/sqoop_datasets/tx_info_2023_Sep_10_to_15.csv'
    -> INTO TABLE tx_info_2023_Sep_10_to_15
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 1197104 rows affected, 65535 warnings (25.25 sec)
Records: 1197104  Deleted: 0  Skipped: 0  Warnings: 0
```

3. Verify the upload by using both MySQL and Sqoop to explore the database

Using MySQL, verifying upload:

show tables;

```
[training@localhost ~]$ mysql -u training -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 9508
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use project1;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+------------------------------+
| Tables_in_project1           |
+------------------------------+
| blocks_2023_Sep_10_to_15     |
| blocks_info_2023_Sep_10_to_15 |
| tx_info_2023_Sep_10_to_15    |
+------------------------------+
3 rows in set (0.00 sec)
```

select * from blocks_2023_Sep_10_to_15 LIMIT 5;

```
mysql> select * from blocks_2023_Sep_10_to_15 limit 5;
+----+------------------------------------------------------------------+---------------------+-------------+
| id | hash                                                             | time                | block_index |
+----+------------------------------------------------------------------+---------------------+-------------+
|  0 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 2023-09-10 00:00:00 |      806982 |
|  1 | 00000000000000000003f3a19f16e20cc5f50f108ed7dabd5957f2e61e90868d | 2023-09-10 00:00:00 |      806981 |
|  2 | 00000000000000000000372109b9a114633512587c8b074910a4bc02921828b59 | 2023-09-10 00:00:00 |      806980 |
|  3 | 00000000000000000000438def0efe8257f6ff025665074dbcf7e1fda070aff30 | 2023-09-10 00:00:00 |      806979 |
|  4 | 00000000000000000000022e8847528953ebe30fc81867dc9ab70b50ce2660d5df | 2023-09-10 00:00:00 |      806978 |
+----+------------------------------------------------------------------+---------------------+-------------+
5 rows in set (0.01 sec)
```

select * from blocks_info_2023_Sep_10_to_15 LIMIT 5;

```
mysql> select * from blocks_info_2023_Sep_10_to_15 limit 5;
+----+------------------------------------------------------------------+-----------+-----------+----------+------------+----------+-------------+------------+--------+---------+
| id | hash                                                             | ver       | bits      | fee      | nonce      | size     | block_index | main_chain | height | weight  |
+----+------------------------------------------------------------------+-----------+-----------+----------+------------+----------+-------------+------------+--------+---------+
|  0 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 545259520 | 386216622 | 17380913 | 3862147940 | 1733081  |      806982 |          0 | 806982 | 3993482 |
|  1 | 00000000000000000003f3a19f16e20cc5f50f108ed7dabd5957f2e61e90868d | 536879104 | 386216622 | 17655615 |  822115101 | 1747360  |      806981 |          0 | 806981 | 3993154 |
|  2 | 00000000000000000000372109b9a114633512587c8b074910a4bc02921828b59 | 547356672 | 386216622 | 15397499 | 2453648697 | 1797458  |      806980 |          0 | 806980 | 3993365 |
|  3 | 00000000000000000000438def0efe8257f6ff025665074dbcf7e1fda070aff30 | 887709696 | 386216622 | 14746325 | 2618089852 | 1913649  |      806979 |          0 | 806979 | 3993468 |
|  4 | 00000000000000000000022e8847528953ebe30fc81867dc9ab70b50ce2660d5df | 536870912 | 386216622 | 17157713 |  700034109 | 1695531  |      806978 |          0 | 806978 | 3993549 |
+----+------------------------------------------------------------------+-----------+-----------+----------+------------+----------+-------------+------------+--------+---------+
5 rows in set (0.02 sec)
```

select * from tx_info_2023_Sep_10_to_15 LIMIT 5;

```
mysql> select * from tx_info_2023_Sep_10_to_15 limit 5;
+----+------------------------------------------------------------------+------------------------------------------------------------------+-----+--------+---------+------+--------+-----------+-----------+------------------------------------------------------------------+-----------+
| id | tx_hash                                                          | block_hash                                                       | ver | vin_sz | vout_sz | size | weight | fee       | relayed_by | lock_time | tx_index                                                         | double_spend | time       | block_index | block_height |
+----+------------------------------------------------------------------+------------------------------------------------------------------+-----+--------+---------+------+--------+-----------+-----------+------------------------------------------------------------------+-----------+
|  0 | ee6337e93ffc48b7ef212e0069d6ddb19148eac2c52c7642d36c3e417e9ca119 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 |   1 |      1 |       4 |  351 |   1296 |         0 | 0.0.0.0   |         0 | 901820990367693                                                  |            0 | 1694319239 |      806982 |       806982 |
|  1 | cec929f3eace7db4c4ff4108e4caf141a1abe718040b7580af6b4e8cf3c2e8c6 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 |   1 |      1 |       1 |  192 |    438 |     33148 | 0.0.0.0   |         0 | 6998496174508493                                                 |            0 | 1694310785 |      806982 |       806982 |
|  2 | 6790aee6b2cc9af0b581a3e7b32bc241ddbb9134a99d5533e00eed34505f1911 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 |   1 |      1 |       2 |  380 |    758 |     50000 | 0.0.0.0   |         0 | 601621470289313                                                  |            0 | 1694319047 |      806982 |       806982 |
|  3 | 4cef857be937250098114e0fc6ab22cc71bf3f3d73e1b627f8fb6c95c2add06c | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 |   2 |      1 |       1 |  193 |    442 |     24400 | 0.0.0.0   |         0 | 3828592774655391                                                 |            0 | 1694319236 |      806982 |       806982 |
|  4 | 79418046ff1075979648918251515321fca5e688a2b6f16e9c372279bca971a | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 |   2 |      1 |       1 |  191 |    437 |     20618 | 0.0.0.0   |         0 | 935655729589848                                                  |            0 | 1694318800 |      806982 |       806982 |
+----+------------------------------------------------------------------+------------------------------------------------------------------+-----+--------+---------+------+--------+-----------+-----------+------------------------------------------------------------------+-----------+
5 rows in set (0.04 sec)
```

Using Sqoop, verifying upload:

sqoop list-tables \--connect jdbc:mysql://localhost/project1 \--username training --password training

```
[training@localhost ~]$ sqoop list-tables \--connect jdbc:mysql://localhost/project1 \--username training --password training
24/03/18 15:23:48 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/18 15:23:49 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/03/18 15:23:50 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
blocks_2023_Sep_10_to_15
blocks_info_2023_Sep_10_to_15
tx info 2023 Sep 10 to 15
```

sqoop eval \--connect jdbc:mysql://localhost/project1 \--username training --password training --query "SELECT * FROM blocks_2023_Sep_10_to_15 LIMIT 5"

```
[training@localhost ~]$ sqoop eval \--connect jdbc:mysql://localhost/project1 \--username training --password training --query "SELECT * FROM blocks_2023_Sep_10_to_15 LIMIT 5"
24/03/18 15:25:02 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/18 15:25:02 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/03/18 15:25:03 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
------------------------------------------------------------------------------------------------
| id          | hash                                                             | time                | block_index |
------------------------------------------------------------------------------------------------
| 0           | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 2023-09-10 00:00:00.0 | 806982      |
| 1           | 00000000000000000003f3a19f16e20cc5f50f108ed7dabd5957f2e61e90868d | 2023-09-10 00:00:00.0 | 806981      |
| 2           | 00000000000000000000372109b9a114633512587c8b074910a4bc02921828b59 | 2023-09-10 00:00:00.0 | 806980      |
| 3           | 00000000000000000000438def0efe8257f6ff025665074dbcf7e1fda070aff30 | 2023-09-10 00:00:00.0 | 806979      |
| 4           | 00000000000000000000022e8847528953ebe30fc81867dc9ab70b50ce2660d5df | 2023-09-10 00:00:00.0 | 806978      |
------------------------------------------------------------------------------------------------
```

sqoop eval \--connect jdbc:mysql://localhost/project1 \--username training --password training --query "SELECT * FROM blocks_info_2023_Sep_10_to_15 LIMIT 5"

```
[training@localhost ~]$ sqoop eval \--connect jdbc:mysql://localhost/project1 \--username training --password training --query "SELECT * FROM blocks_info_2023_Sep_10_to_15 LIMIT 5"
24/03/18 15:26:32 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/18 15:26:32 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/03/18 15:26:33 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
-------------------------------------------------------------------------------------------------------------------------------------------------------------
| id          | hash                                                             | ver         | bits        | fee         | nonce                | size        | block_index | main_chain | height      | weight      |
-------------------------------------------------------------------------------------------------------------------------------------------------------------
| 0           | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 545259520   | 386216622   | 17380913    | 3862147940           | 1733081     | 806982      | 0 | 806982      | 3993482     |
| 1           | 000000000000000000003f3a19f16e20cc5f50f108ed7dabd5957f2e61e90868d | 536879104   | 386216622   | 17655615    | 822115101            | 1747360     | 806981      | 0 | 806981      | 3993154     |
| 2           | 000000000000000000372109b9a1146335125 87c8b074910a4bc02921828b59 | 547356672   | 386216622   | 15397499    | 2453648697           | 1797458     | 806980      | 0 | 806980      | 3993365     |
| 3           | 000000000000000000438def0efe8257f6ff025665074dbcf7e1fda070aff30  | 887709696   | 386216622   | 14746325    | 2618089852           | 1913649     | 806979      | 0 | 806979      | 3993468     |
| 4           | 00000000000000000022e8847528953ebe30fc81867dc9ab70b50ce2660d5df  | 536870912   | 386216622   | 17157713    | 700034109            | 1695531     | 806978      | 0 | 806978      | 3993549     |
-------------------------------------------------------------------------------------------------------------------------------------------------------------
```

sqoop eval \--connect jdbc:mysql://localhost/project1 \--username training --password training --query "SELECT * FROM tx_info_2023_Sep_10_to_15 LIMIT 5"

```
[training@localhost ~]$ sqoop eval \--connect jdbc:mysql://localhost/project1 \--username training --password training --query "SELECT * FROM tx_info_2023_Sep_10_to_15 LIMIT 5"
24/03/18 15:27:40 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/18 15:27:40 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/03/18 15:27:40 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
-------------------------------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------
| id          | tx_hash                                                          | block_hash                                                       | ver         | vin_sz      | vout_sz     | size        | weight      | fee                 | relayed_by      | lock_time           | tx_index            | double_spend | tim
e                | block_index | block_height |
-------------------------------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------
| 0           | ee6337e93ffc48b7ef212e0069d6ddb19148eac2c52c7642d36c3e417e9ca119 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 1           | 1           | 4           | 351         | 1296        | 0                   |
    | 0.0.0.0         | 0                   | 901820990367693       | 0 | 1694319239            | 806982      | 806982      |
| 1           | cec929f3eace7db4c4ff4108e4caf141a1abe718040b7580af6b4e8cf3c2e8c6 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 1           | 1           | 1           | 192         | 438         | 33148               |
    | 0.0.0.0         | 0                   | 6998496174508493      | 0 | 1694318785            | 806982      | 806982      |
| 2           | 6798aee6b2cc9af0b581a3e7b320c241ddbb9134a99d5533e00eed345051f1911 | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 1           | 1           | 2           | 380         | 758         | 50000               |
    | 0.0.0.0         | 0                   | 601621470289313       | 0 | 1694319047            | 806982      | 806982      |
| 3           | 4cef857be937250098114e0fc6ab22cc71bf3f3d73e1b627f8fb6c95c2add06c | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 2           | 1           | 1           | 193         | 442         | 24400               |
    | 0.0.0.0         | 0                   | 382859277465539 1     | 0 | 1694319236            | 806982      | 806982      |
| 4           | 79416846ff1075979648918251 51c5321fca5e688a2b6f16e9c372279bca971a | 00000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 2           | 1           | 1           | 191         | 437         | 20618               |
    | 0.0.0.0         | 0                   | 935655729589848       | 0 | 1694318800            | 806982      | 806982      |
-------------------------------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------
```

4. Use Sqoop to import each table from MySQL directly into Hive

sqoop import \

--connect jdbc:mysql://localhost/project1 \

--username training \

--password training \

--table blocks_2023_Sep_10_to_15 \

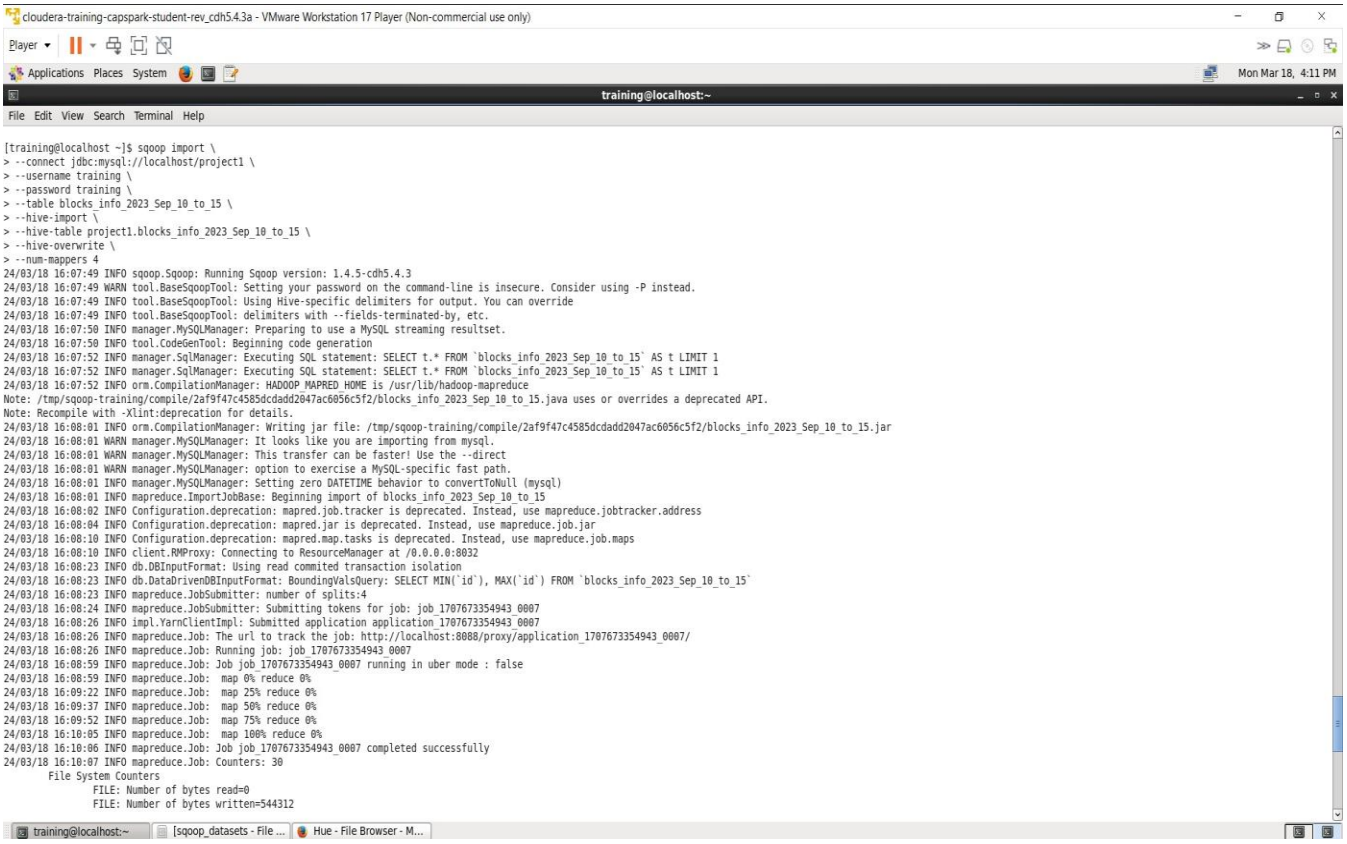
--hive-import \

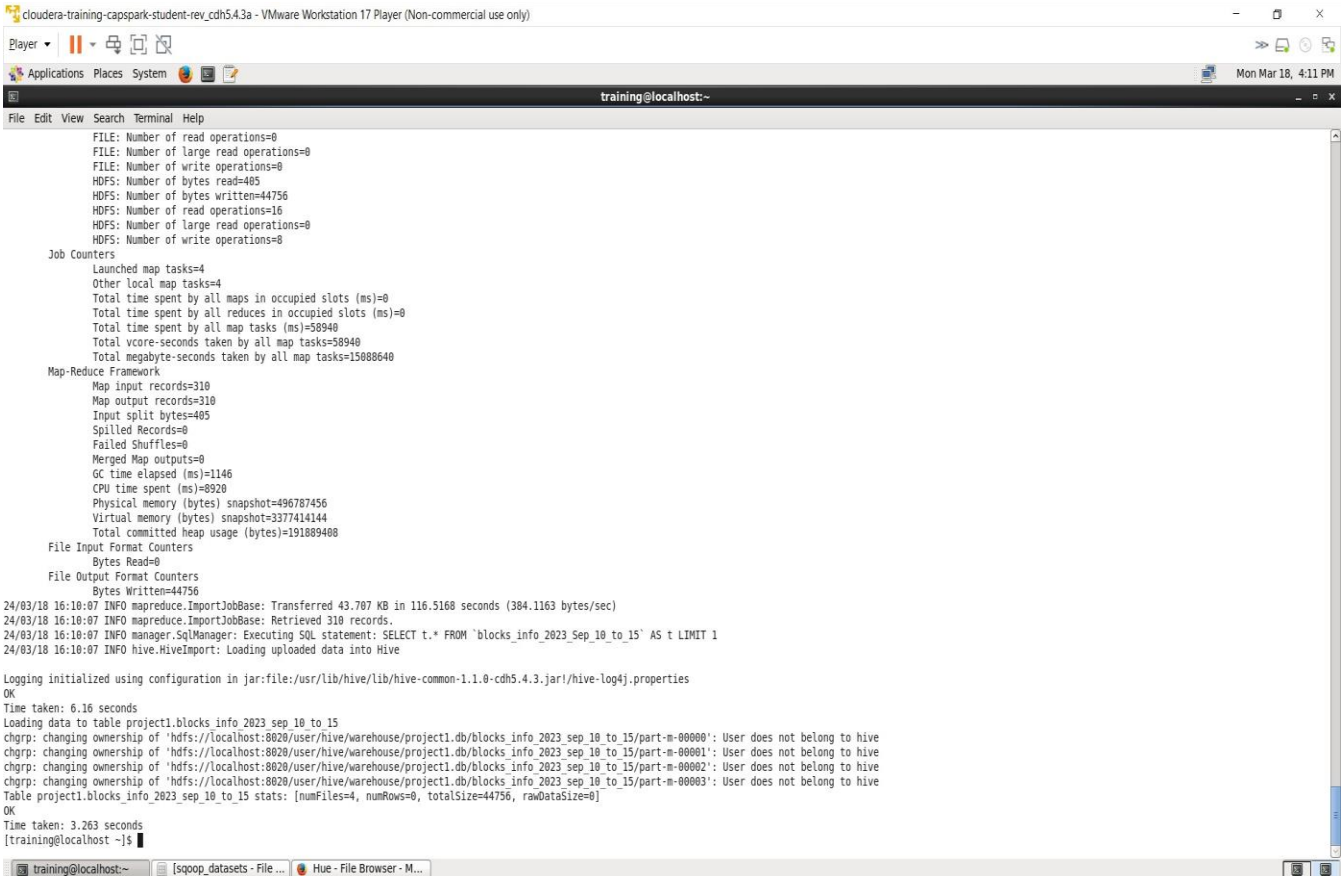
--hive-table project1.blocks_2023_Sep_10_to_15 \

--hive-overwrite \

--num-mappers 4

```
sqoop import \
--connect jdbc:mysql://localhost/project1 \
--username training \
--password training \
--table blocks_info_2023_Sep_10_to_15 \
--hive-import \
--hive-table project1.blocks_info_2023_Sep_10_to_15 \
--hive-overwrite  \
--num-mappers 4
```






```
sqoop import \
--connect jdbc:mysql://localhost/loudacre \
--username training \
--password training \
--table tx_info_2023_Sep_10_to_15 \
--hive-import \
--hive-table project.tx_info_2023_Sep_10_to_15 \
--hive-overwrite  \
--num-mappers 4
```

claudera-traini11g-capspark-student-rev_cdh5.4.3a VMware Workstatian 17 Player (Non-rnmmercial use only)

Elaye, • Applications Places System • MonMarIB, 4:24PM

tra,mng@localhost

File Edit View Searrh Terminal Help

```
[training@localhost -JS clear

[training@localhost -JS sqoop import\
  --connectjdbc:mysql://localhost/projectl\
  --usernametraining\
  --password training\
    tabletxinfo2B23SepHltol5\
) --hive-impOrt \
    h1vetablepr0Jectl.txinfo2823SepI8tol5\
) --hive-overwrite\
  --num-mappers 4
24/83/1816:16:22 INFOsqoop.Sqoop: Running sqoop version: l.4.5-cdh5.4.3
24/03/18 16:16:22 WARN tool.BaseSqoopTool: Setting your password on the coffllland-line is insecure. Consider using -P instead.
24/03/18 16:16:22 INFO tool.BasesqoopTool: Using Hive-specific delimiters for output. You can override
24/83/18 16:16:22 INFO tool.BaseSqoopTool: delimiters with --fields-tenninated-by, etc.
24/03/18 16:16:23 INFO manager.MySOLHanager: Preparing to use a MySOL streaming resultset.
24/83/18 16:16:23 INFO tool.CodeGenTool: Beginning code generation
24/83/18 16:16:25 INFO manager.SqlHanager: Executing SOL statement: SELECT t, FROM 'tx info 2023 Sep 18 to 15' AS t LIMIT l
24/83/18 16 16 25 INFO manager SqlHanager Executing SOL statement SELECT t * FROM 'tx-info-2023-Sep-18-to-15' AS t LIMIT 1
24/83/18 16:16:25 INFO orm.compilationHanager: HADOOP 11APREO HOME is /usr/lib/hadoop-mal)reduCe
Note: /tmp/sqoop-training/cornpile/d285413f568359619ab5dd3849'iif8753/tx info 2823 Sep 18 to 15.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/83/181616 34 INFOormComp1lationHanager \rlnting Jar file /tmp/sqoop trainrng/cornp1le/d285413f568359619ab5dd38498f8753/tx info 2823 Sep 18 to 15 Jar
24/83/18 16:16:34 \r/MN manager.MySOLHanager: It looks like you are importing from mysql.
24/83/18 16:16:34 WARN manager.MySOL.Mam1ger: This transfer rnn be faster! Use the --direct
24/83/18 16:16:34 WARN manager.MySOLHanager: option to exercise a HySOL-specific fast path.
24/83/18 16:16:34 INFO manager.MySOLHanager: Setting zero DATETIME behavior to convertToNull {mysql)
24/83/18 16:16:34 INFO mapreduce.ImportJobBase: Beginning import of tx info 2023 Sep IB to 15
24/83/18 16:16:34 INFO configuration.deprecation: mapred.job.tracker iS dep"recat"ed. instead, use mapreduce.jobtracker.address
24/83/18 16:16:35 INFO Configuration.deprecation: mapred.jar iS deprecated. Instead, use mapreduce.job.jar
24/83/1816:16:39INFOConfiguration.deprecation:mapred.map.taskSisdeprecated. Instead, usemapreduce.job.maps
24/83/18 16:16:39 INFO client.RMProxy: Connecting to ResourceManager at /0.8.0.8:8832
24/83/18 16:16:49 INFO db.DBinputFormat: Using read commited transaction isolation
24/83/18 16:16:49 INFO db.DataDrivenDBinputformat: BoundingValsOuery: SELECT MIN('id'), MAX('id') FROM tx_info_2823 Sep 18 to 15
24/83/18 16:16:49 INFO mapreduce.JobSubmitter: number of splits:4
24/83/18 16:16:58 INFD mapreduce.JobSubmitter: Submitting tokens for job: job 1787673354943 8808
24/83/18 16:16:52 INFO impl.YarnclientImpl: Submitted application application-1787673354943-8808
24/93/18 16:16:52 INFO mapreduce.Job: The url to track the job: http://localhiist:8888/proxy/application 1787673354943 8988/
24/83/18 16:16:52 INFO mapreduce.Job: Running job: job 1787673354943 8EI88
24/83/18 16:17:28 INFO mapreduce.Job: Job job 1787673354943 8888 run'iiing in uber mode : false
24/83/18 16:17:28 INFO mapreduce.Job: map B\-reduce EI%.
24/83/18 16:18:13 INFO mapreduce.Job: map 25\ reduce 8\
24/83/18 16:18:42 INFO mapreduce.Job: map SB\ reduce B\
24/83/18 16:19:55 INFO mapreduce.Job: map 75\ reduce 8\
24/83/18 16:21:22 INFO mapreduce.Job: map 188\ reduce 8%
24/83/18 16:21:24 INFO mapreduce.Job: Job job 1787673354943 8888 completed successfully
24/83/18 16:21:26 INFO mapreduce.Job: counters": 3EI
        File system counters
                FILE: Numberofbytesread=9
```

lii trainl_)g@localhost:- [sqoop_datasets-File ... li Hue File Browser - M...

claudera-traini11g-capspark-student-rev_cdh5.4.3a VMware Workstatian 17 Player (Non-commercial useonly)

Elayer • Applications Places System Mon Mar IS, 4:24 PM

tra1nrng@localhost:-

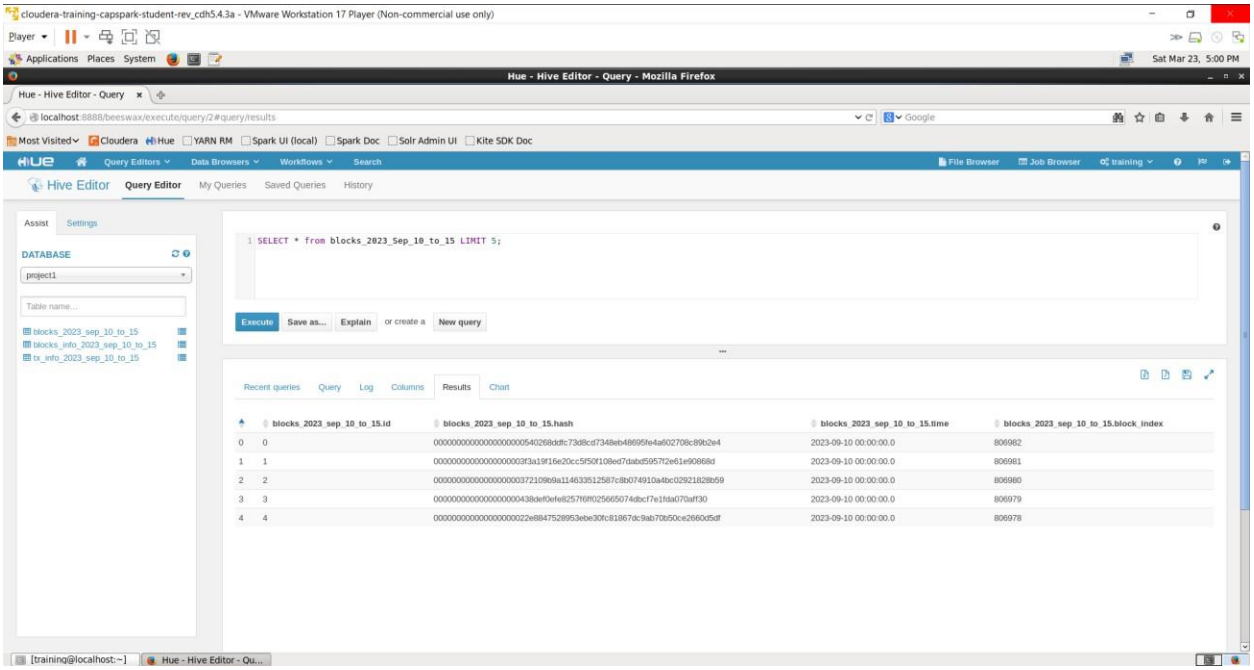
File Edit View searrh Terminal Help

```
                FILE: Number of read operations=8
                FILE: Number of large read operations=8
                FILE: Number of write operations=B
                HDFS: Number of byte:s read=429
                HDFS: Number of bytes written=256678856
                HDFS: Number of read operations=l6
                HDFS: Number of large read operations=9
                HDFS: Number of write operations:8
        Job Counters
                Launched map tasks=4
                Other local map tasks=4
                Total time spent by all maps in occupied slots {ms}=B
                Total time spent by all reduces in occupied slots {msl=B
                Total time :spent by all map tasks (msl=234737
                Total vcore-seconds taken by all map tasks,,,234737
                Total megabyte-seconds taken by all map tasks=60992672
        Map-Reduce Framework
                Hap input records=ll97184
                Hap output records=ll971B4
                Inputsplitbytes=429
                SpilledRecords=0
                FailedShuffles=0
                Merged Map outputs=B
                GC time elapsed (msl=3421
                CPUtimespent{ms}=981}1B
                Physical memory (bytes) snapshot=51773449B
                Virtual memory (byte:sl snapshot=3382149128
                Total comitted heap usage {bytes)=l91889488
        File Input Format Counters
                BytesRead=B
        File Output Format counters
                Bytes Written=256678856
24/83/18 16:21:26 INFO mapreduce.ImportJobBase: Transferred 244.788 MB in 287.1819 seconds (873.El799 KB/sec)
24/83/18 16:21:26 INFO mapreduce.ImportJobBase: Retrieved 1197104 records.
24/B3/18 16 21 26 INFO manager Sql11anager Executing SOL statement SELECT t • FROM tx rnfo 2823 Sep 18 to 15 AS t LIMIT 1
24/B3/18 16:21:27 INFO hive.Hiveimport: Loading uploaded data into Hive

Logginginitializedusingconfigurationinjar:file:/usr/lib/hive/lib/hive-co!Mlon-l.l.B-cdh5.4.3.jar!/hive-log4j.properties
0K
Time taken: 18.818 seconds
Loadingdatatotableprojectl.txinfo2823sepl8to 15
chgrp: changing ownership of 'hdfS://liicalhiist:Se2B;uSer/hive/warehouse/projectl.db/tx info 2B23 sep 1B to l5/part-m-0EIBBB' User does not belong to hive
chgrp: changing ownership of 'hdfs: //localhost:8B20/user/hive/warehouse/projectl .db/tx-info-2823-sep-18- t-o15/part-m-0EIB81' User does not belong to hive
chgrp: changing ownership of 'hdfs: //localhost:8820/user/hive/warehouse/projectl .db/t-xinfo-2823-sep-18- t-o l5/part-m-0EI882' User does not belong to hive
chgrp: changing ownership of 'hdfs: //localhost:8B20/user/hive/warehouse/project .db/tx-info-2B23-sep-1B-t-o15/part-m-0EIBB3' User does not belong to hive
Table proJectl tx rnfo 2823 sep 18 to 15 stats. [nurnF1lese:c4, numRows El, totalS1ze 256678856 raW OataS1ze"eT
0K
Time taken: 5.896seconds
[training@localhost -JS
```

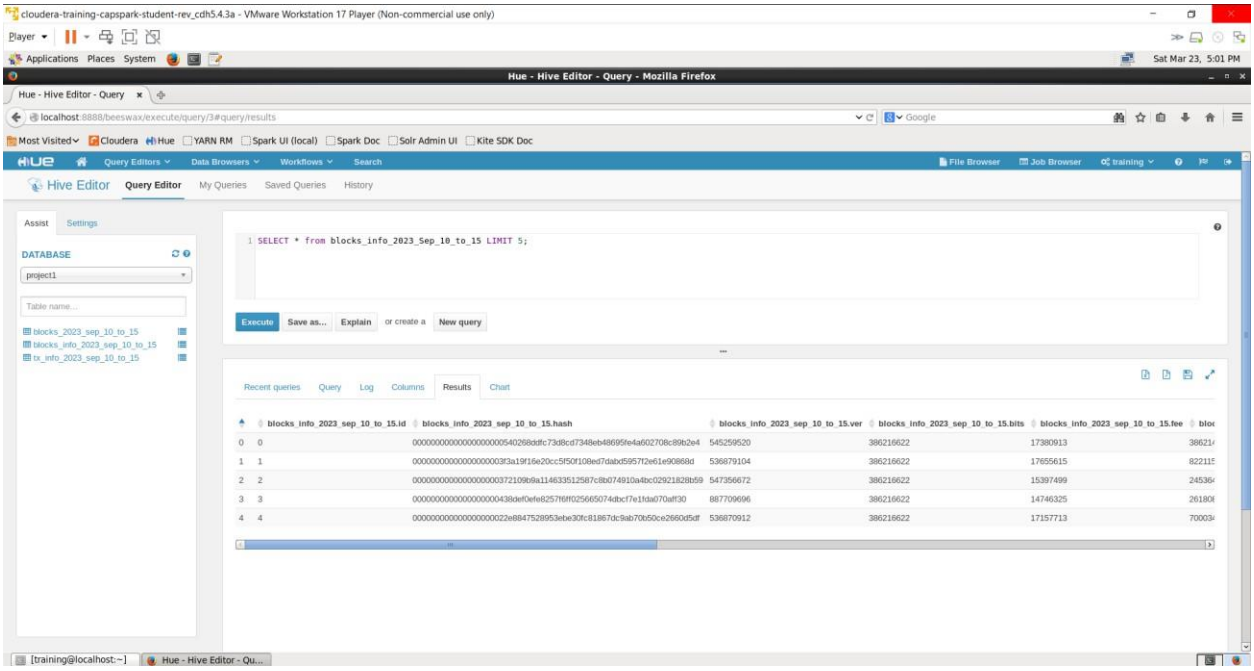
(jitraini™.l@localhos [sqoop_datasets File ... Hue- File Browser- M...

5. Verify the import by browsing Hive

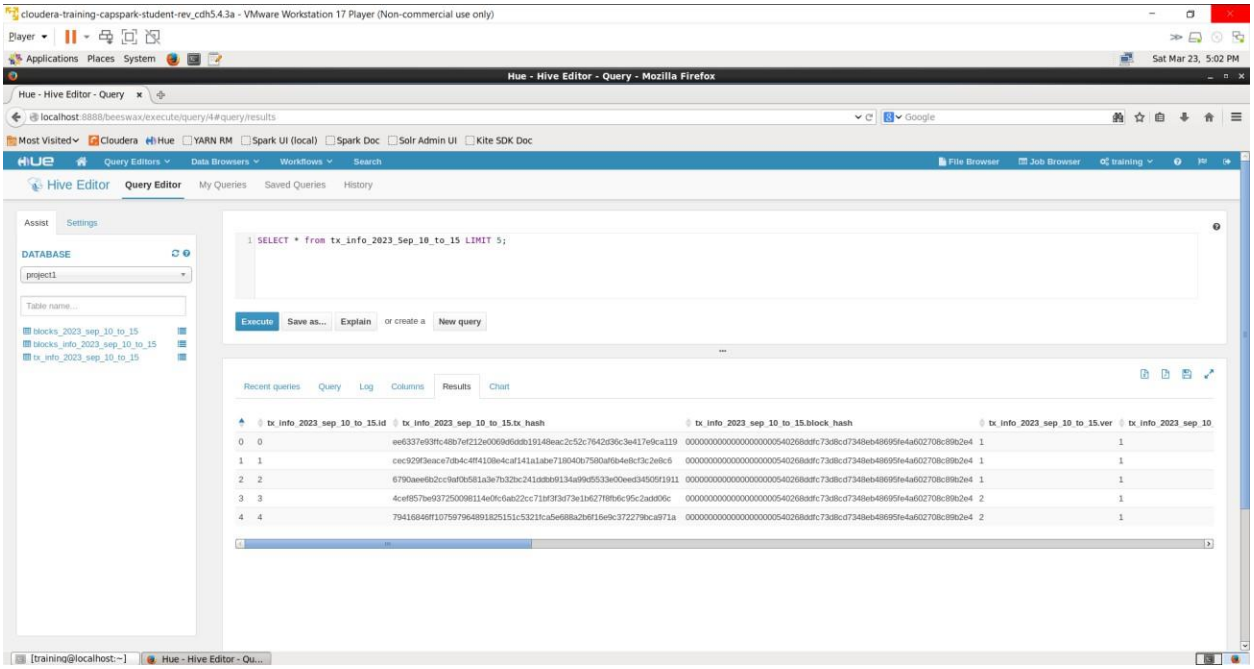
SELECT * from blocks_2023_Sep_10_to_15 LIMIT 5;



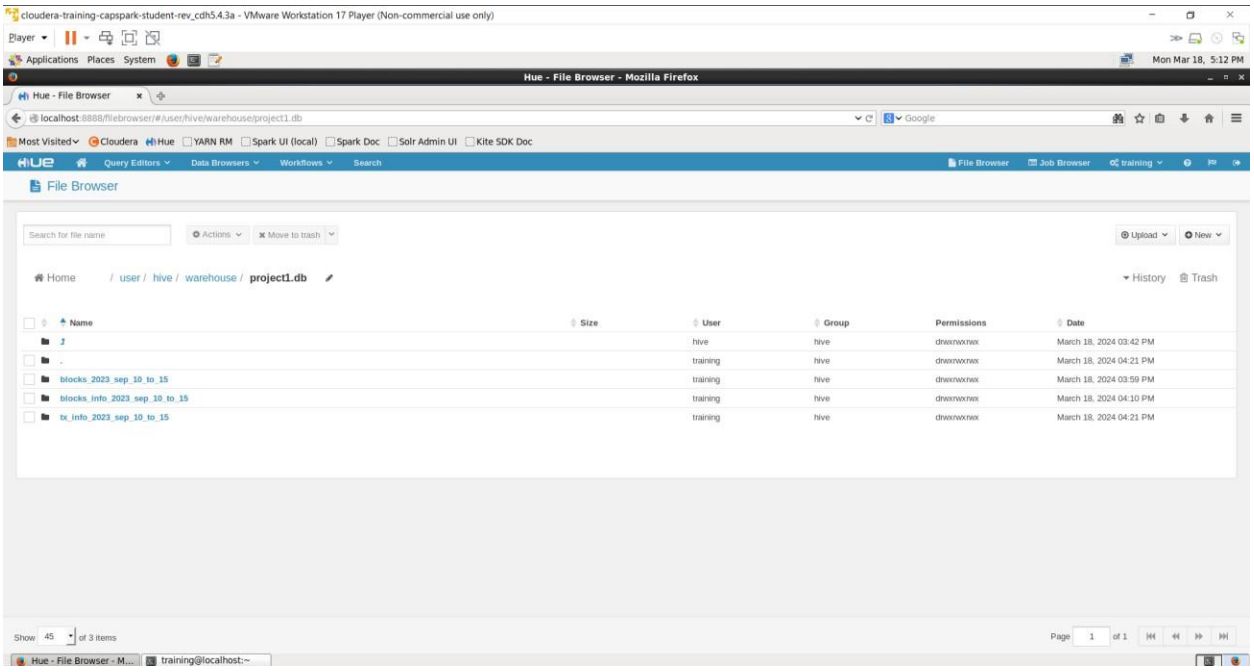
SELECT * from blocks_info_2023_Sep_10_to_15 LIMIT 5;



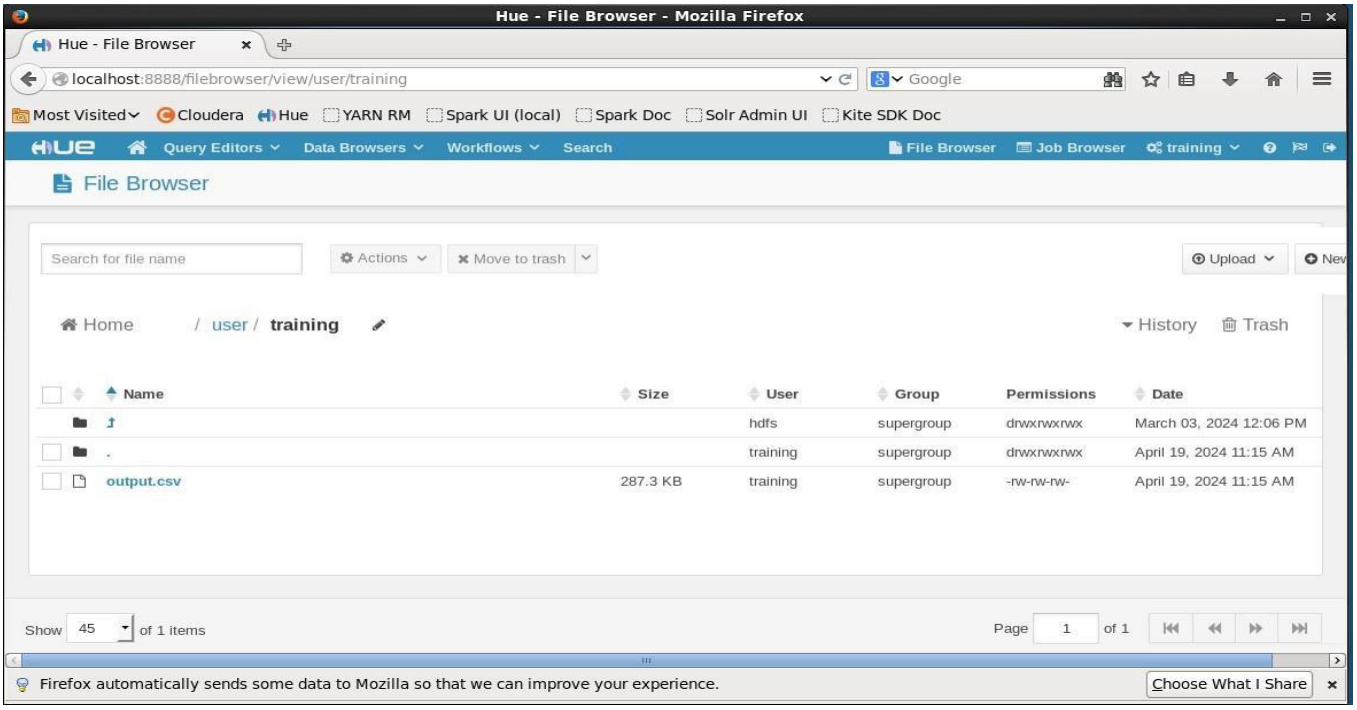
SELECT * from tx_info_2023_Sep_10_to_15 LIMIT 5;

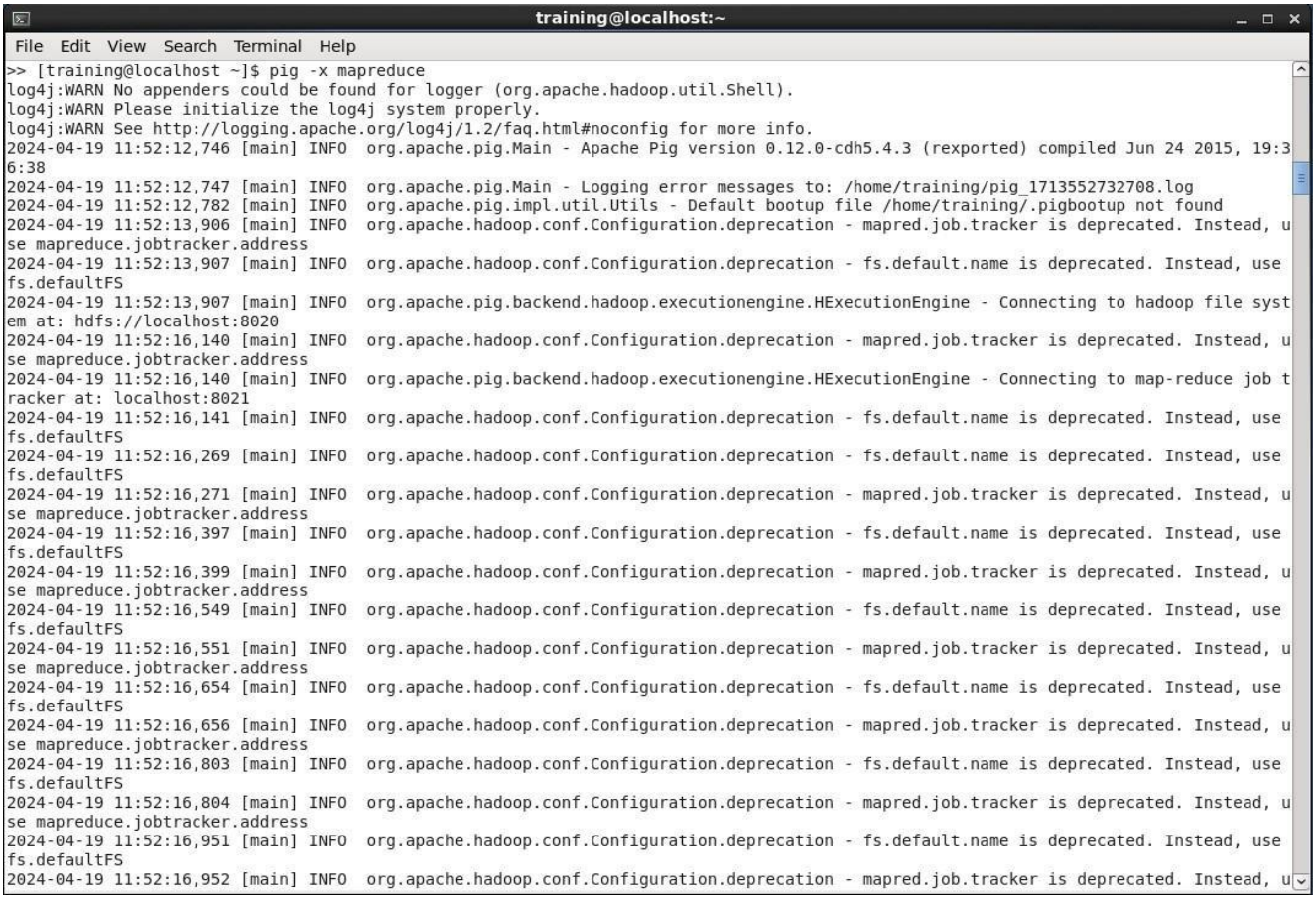


6. Show where the actual data is stored on HDFS



A. Data Analysis using Pig





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;
```

```
training@localhost:~
File Edit View Search Terminal Help
grunt> 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
);
grunt> total blocks= FOREACH (GROUP data ALL) GENERATE COUNT(data);
grunt> DUMP total blocks;
2024-04-19 11:58:46,855 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY
2024-04-19 11:58:46,962 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES ENABLED [AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-04-19 11:58:47,488 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-04-19 11:58:47,518 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
2024-04-19 11:58:47,583 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-04-19 11:58:47,583 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2024-04-19 11:58:47,767 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2024-04-19 11:58:48,325 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-04-19 11:58:48,486 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2024-04-19 11:58:48,486 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-04-19 11:58:48,487 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2024-04-19 11:58:48,492 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating# of required reducers.
2024-04-19 11:58:48,492 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
```

```
training@localhost,-
File Edit View Search Terminal Help
HadoopVersion   PigVersion      Userid  StartedAt       FinishedAt      Features
2.6.0-cdhS.4.3  0.12.0-cdhS.4.3 training        2024-04-19 12:06:11     2024-04-19 12:07:01     GROUP BY

Success!

Job Stats (time in seconds):
JobId   Maps    Ft.educes MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime
e       MedianReducetime        Alias   Feature outputs
job 1709496148644 0002  1       1       10      10      10      10      8       8       8       8       1-10,data,total blocks  GROUP_BY,COMBINER  hdfs://localhost:8020/tmp/temp-112855710/tmp1891525910,

Input(s):
successfully read 1027 records (294515 bytes) from: "/user/training/output.csv"

output(s):
successfully stored 1 records (7 bytes) in: "hdfs://localhost:8020/tmp/temp-112855710/tmp1891525910"

Counters:
Total records written : 1
Total bytes written : 7
Spillable Memory Manager spill count  0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
Job 1709496148644 0002


2024-04-19 12:07:02,128 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered \'larning FIELO DISCARDED TYPE CONVERSION FAILED 10 time(s).
2024-04-19 12:07:02,128 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - success!
2024-04-19 12:07:02,129 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-04-19 12:07:02,129 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-04-19 12:07:02,130 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-04-19 12:07:02,145 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileinputFormat - Total input paths to process
2024-04-19 12:07:02,145 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
1
(1027)
grunt> clear
```

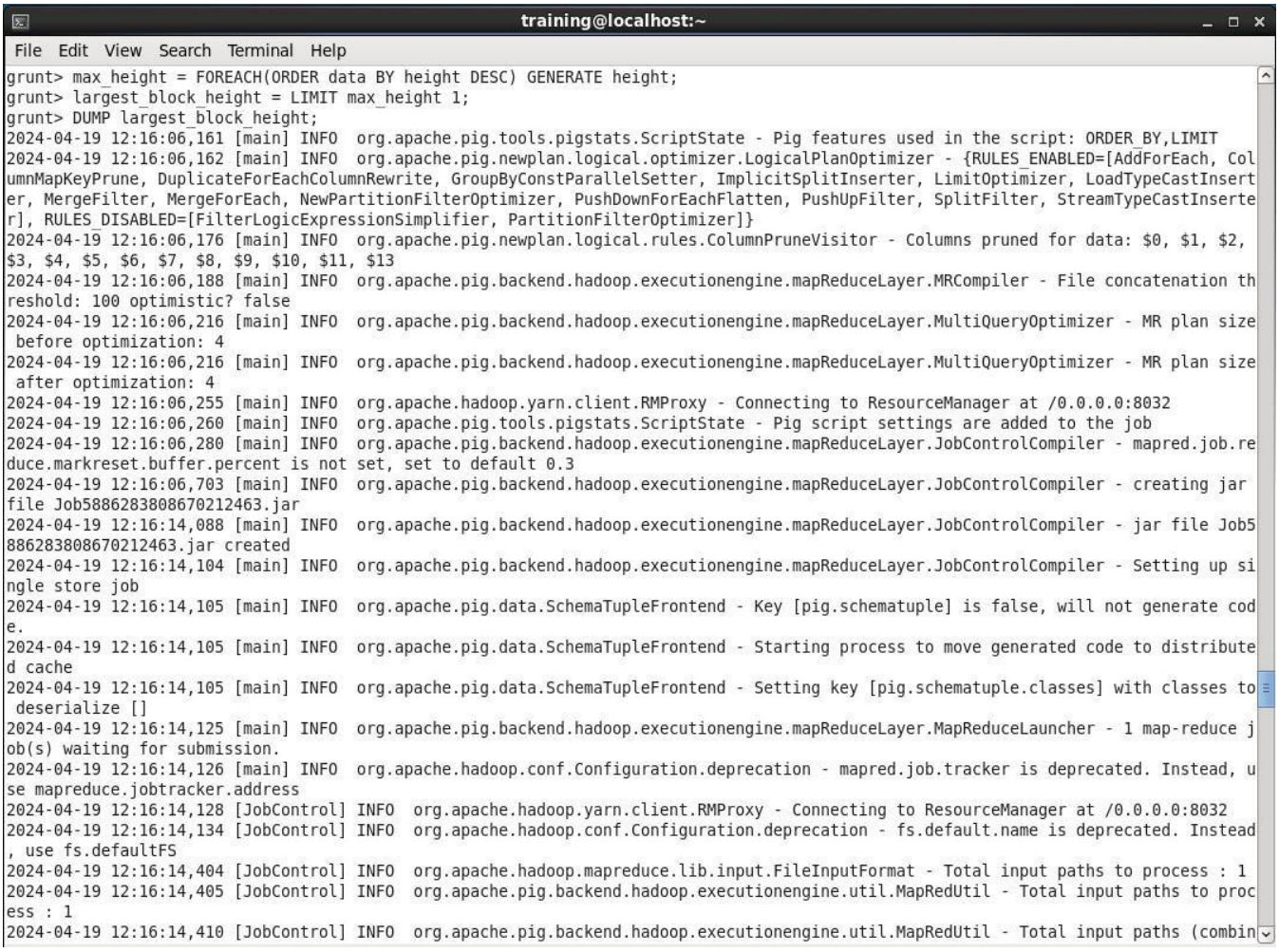
-- 2) Find the largest block height

max_height = FOREACH (ORDER data BY height DESC) GENERATE height;

largest_block_height = LIMIT max_height 1;

-- Output the result

DUMP largest_block_height;



```
training@ acalhast:-
File Edit View Search Terminal Help
HadoopVersion   PigVersion      Userid  StartedAt       FinishedAt      Features
2.6.0-cdh5.4.3  0.12.0-cdh5.4.3 training        2024-04-19 12:16:06     2024-04-19 12:18:52     OROER_BY,LIMIT

Success!

Job Stats (time in seconds):
Jobid   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTim
e       MedianReducetime        Alias   Feature Outputs
job 1709496148644 0003  1       0       8       8       8       n/a     n/a     n/a     n/a     data    MAP ONLY
job-1709496148644-0004  1       1       7       7       7       7       8       8       8       8       1-26    SAMPLER
job-1709496148644-0005  1       1       8       8       8       8       9       9       9       9       1-26    ORDER BY.COMBINER
job-1709496148644-0006  1       1       7       7       7       7       8       8       8       8       1-26,max height         hdfs
://localhost:8020/tmp/temp-112855710/tmp1544766529,

Input(sl:
Successfully read 1027 records (294515 bytes) from: "/user/training/output.csv"

Output(sl:
Successfully stored 1 records (9 bytesl in: "hdfs://localhost:8020/tmp/temp-112855710/tmp1544766529"

Counters:
Total records written : 1
Total bytes written : 9
Spillable Memory Manager spill count    0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job 1709496148644 0003  ->      ]Ob 1709496148644 0004,
job-1709496148644-0004  ->      job-1709496148644 0005,
job-1709496148644-0005  ->      JOb-1709496148644-0006,
JOb-1709496148644-0006


2024-04-19 12:18:53,105 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58171. Already t
ried 0 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:54,110 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58171. Already t
ried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:55,115 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58171. Already t
ried 2 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:55,232 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onstatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:18:57,797 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:51714. Already tv
```

```
training@lacalhast:~
File Edit View Search Terminal Help
ried 1 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:55,115 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58171. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:55,232 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onstatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:18:57,797 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:51714. Already t
ried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:58,802 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:51714. Already t
ried 1 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:59,806 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:51714. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:59,920 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onstatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:19:01,314 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58982. Already t
ried 0 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:02,415 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58982. Already t
ried 1 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:03,418 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58982. Already t
ried 2 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:03,530 [main] INFO  org.apache.hadoop.mapred.ClientService□elegate - Application state is completed. FinalApplicati
onstatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:19:04,972 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:60722. Already t
ried 0 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:05,977 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:60722. Already t
ried 1 time(s); retry policy is RetryUpToMaximumcountWithfixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:06,982 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:60722. Already t
ried 2 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:07,096 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:19:07,305 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReducelauncher - Encountered Wa
rning FIELD DISCARDED TYPE CONVERSION FAILED 1 time(s).
2024-04-19 12:19:07,305 [m8in] INFO  Org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReducelauncher - Success!
2024-04-19 12:19:07,306 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation    ts.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 12:19:07,306 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 12:19:07,307 [main] INFO   org.apache.pig.data.SchemaTupleBackend -  Key [pig.schematuple] was  not set... will not generate
 code.
2024-04-19 12:19:07,320 [main] INFO   org.apache.hadoop.mapreduce.lib.input.Fileinputformat   Total input paths to process : 1
2024-04-19 12:19:07,320 [main] INFO   org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process :
1
(835509)
grunt>
```

-- 3) Find the largest block height

max_height = FOREACH (ORDER data BY height DESC) GENERATE height;

largest_block_height = LIMIT max_height 1;

-- Filter data for the largest block

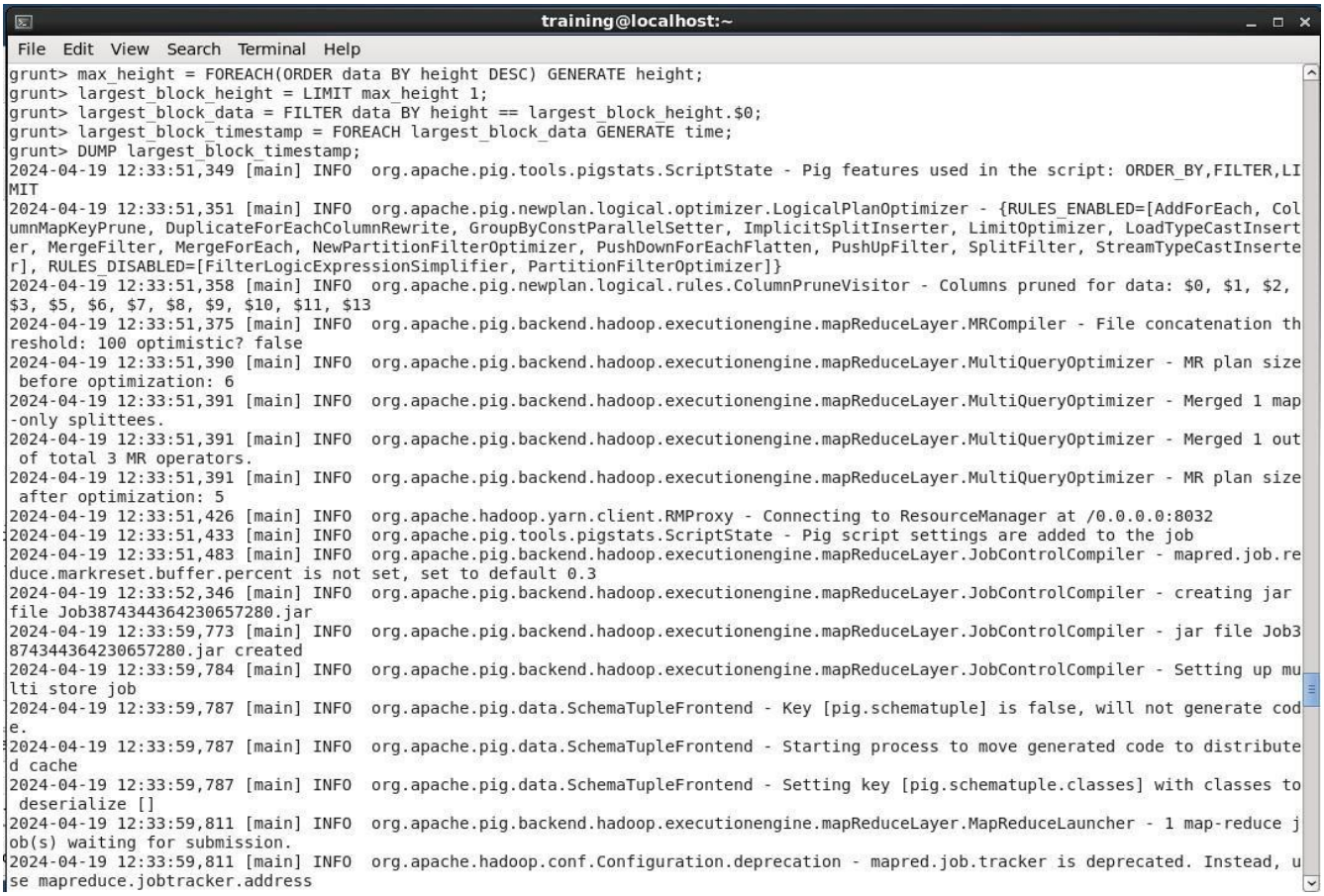
largest_block_data = FILTER data BY height == largest_block_height.$0;

-- Extract date and time for the largest block

largest_block_timestamp = FOREACH largest_block_data GENERATE time;

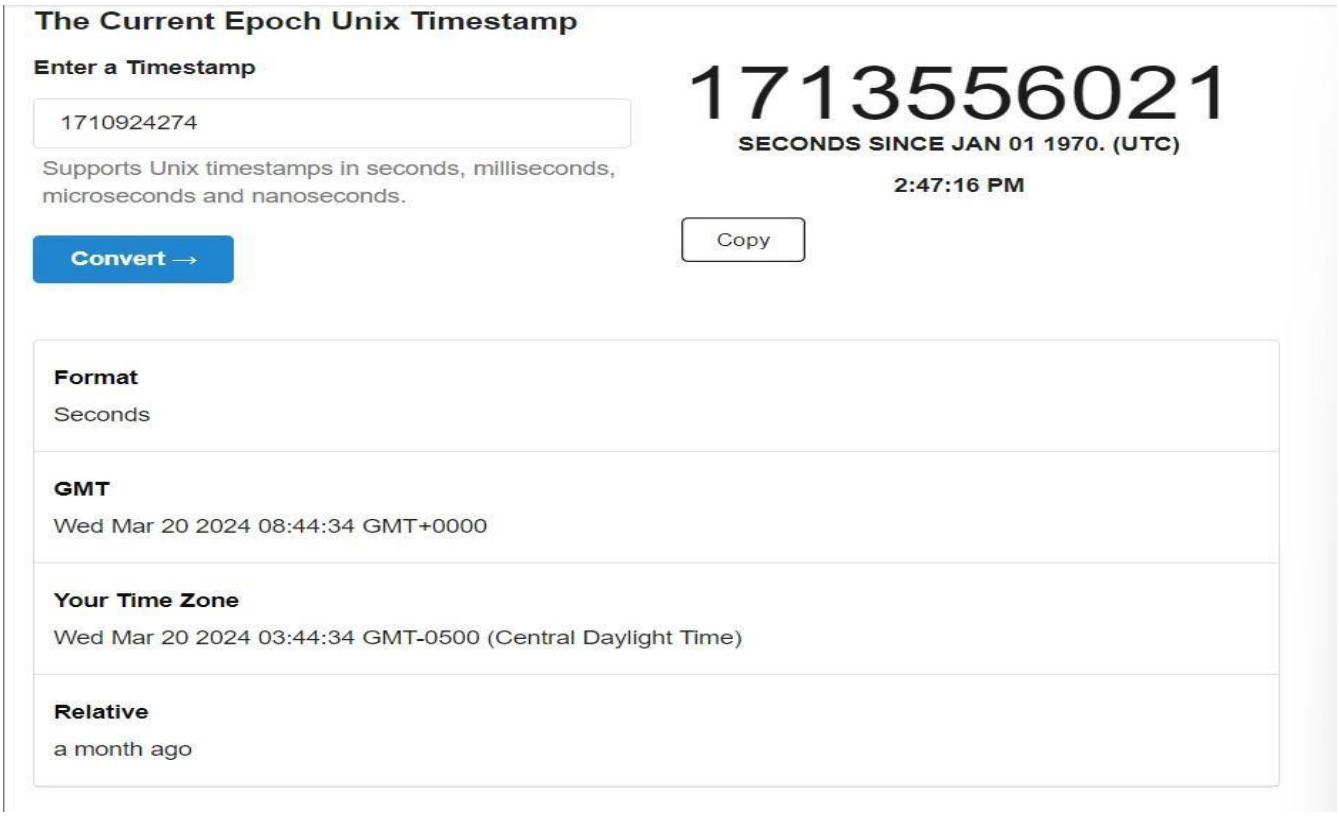
-- Output the result

DUMP largest_block_timestamp;




training@localhost:~
File Edit View Search Terminal Help
HadoopVersion PigVersion Userid StartedAt FinishedAt Features
2.6.0-cdh5.4.3 0.12.0-cdh5.4.3 training 2024-04-19 12:33:51 2024-04-19 12:37:12 ORDER_BY,FILTER,LIMIT
Success!
Job Stats (time in seconds):
Jobid Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgR.educeTim
e MedianReducetime Alias Feature Outputs
job 1709496148644 0007 1 0 8 8 8 8 n/a n/a n/a n/a data MULTI OUERY,MAP ONLY
job-1709496148644-0008 1 1 7 7 7 7 8 8 8 8 1-68 SAMPLER -
job-1709496148644-0009 1 1 7 7 7 7 8 8 8 8 1-68 ORDER BY,COMBINER
job-1709496148644-0010 1 1 7 7 7 7 8 8 8 8 1-68,max height
job-1709496148644-0011 1 0 7 7 7 7 n/a n/a n/a n/a largest block data.largest b
lock_timestamp 1-IAP_ONLY hdfs://localhost:8020/tmp/temp-112855710/tmp-2010907623,
InputIsI:
Successfully read 1027 records (294515 bytes) from: u/user/training/output.csvu
Output (s):
Successfully stored 1 records 19 bytes) in: "hdfs://localhost:8020/tmp/temp-112855710/tmp-2010907623"
Counters:
Total records written: 1
Total bytes written : 9
Spillable Memory Manager spill count 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job 1709496148644 0007]Ob 1709496148644 0008,]0b 1709496148644 0011,
job-1709496148644-0008 job-1709496148644 0009, -
job-1709496148644-0009 -> job-1709496148644-0010,
job-1709496148644-0010 -> Job-1709496148644=0011,
job=1709496148644=0011
2024-04-19 12:37:13,291 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:35782. Already t
ried 0 time(s); retry policy is RetryUpToMaximumcountWithfixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:14,296 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:35782. Already t
ried 1 time(s); retry policy is R.etryUpToMaximumcountWithfixedSleep{maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:15,302 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:35782. Already t
ried 2 time(s); retry policy is R.etryUpToMaximumCountWithfixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:15,418 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati


training@localhost:-
File Edit View Search Terminal Help
ried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:18,667 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:46423. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:18,780 [main] INFO org.apache.hadoop.mapred.ClientserviceDelegate - Application state is completed. FinalApplicati
onstatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:37:20,082 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:36102. Already t
ried 0 time(sl; retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:21,085 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:36102. Already t
ried 1 time(sl; retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:22,090 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:36102. Already t
ried 2 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:22,203 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:37:23,441 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:44079. Already t
ried 0 timels); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:24,445 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:44079. Already t
ried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:25,449 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:44079. Already t
ried 2 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:25,561 [main] INFO org.apache.hadoop.mapred.ClientserviceDelegate - Application state is completed. FinalApplicati
onstatus=SUCCEEDEO. Redirecting to job history server
2024-04-19 12:37:26,783 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:42563. Already t
ried 0 time(sl; retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:27,787 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:42563. Already t
ried 1 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:28,792 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:42563. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:37:28,905 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:37:29,080 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Wa
rning FIELD DISCARDED TYPE CONVERSION FAILED 2 time(s).
2024-04-19 12:37:29,080 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - success!
2024-04-19 12:37:29,080 [main] INFO org.apache.hadoop.conf.Configuration.deprecation fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 12:37:29,080 [main] INFO org.apache.hadoop.conf.Configuration.deprecation mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-04-19 12:37:29,081 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
code.
2024-04-19 12:37:29,093 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileinputFormat - Total input paths to process 1
2024-04-19 12:37:29,093 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
1
(1710924274)
grunt>



-- 4) Group data by block height and count transactions

transactions_per_block = FOREACH (GROUP data BY height) GENERATE group AS block_height, SUM(data.n_tx) AS total_transactions;

-- Find the block with the highest number of transactions

max_transactions = FOREACH (ORDER transactions_per_block BY total_transactions DESC) GENERATE block_height, total_transactions;

highest_transactions_block = LIMIT max_transactions 1;

-- Output the result

DUMP highest_transactions_block;

training@localhost:-

File Edit View Search Terminal Help

```
grunt> transactions per block= FOREACH (GROUP data BV height) GENERATE group AS block height, SUM(data.n tx) AS total transactions;
grunt> max transactions-= FOREACH (ORDER transactions per block BV total transactions DESC) GENERATE block_height, total transactions;
grunt> highest transactions block= LIMIT max transactions 1;
grunt> DUMP highest transactions block;
2024-04-19 12:57:15 226 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BV,ORDER_BV,LIMIT
2024-04-19 12:57:15,228 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES ENABLEO=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitinserter, LimitOptiffiizer, LoadTypeCastinserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastinserter], RULES DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-04-19 12:57:15,271 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReducelayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-04-19 12:57:15,282 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReducelayer.CombinerOptimizer  Choosing to move algebraic foreach to combiner
2024-04-19 12:57:15,292 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiOueryOptimizer  MR plan size before optimization: 4
2024-04-19 12:57:15,292 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReducelayer.MultiQueryOptimizer  MR plan size after optimization: 4
2024-04-19 12:57:15,338 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2024-04-19 12:57:15,344 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-04-19 12:57:15,368 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReducelayer.JobControlCompiler  mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2024-04-19 12:57:15,369 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReducelayer.JobControlCompiler  Reduce phase detected, estimating# of required reducers.
2024-04-19 12:57:15,369 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler  Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReducelayer.InputSizeReducerEstimator
2024-04-19 12:57:15,372 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReducelayer.InputSizeReducerEstimator - BytesPerReducer-1000000000 maxReducers=999 totalinputFileSize=294150
2024-04-19 12:57:15,372 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler  Setting Parallelism to 1
2024-04-19 12:57:15,776 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReducelayer.JobControlCompiler  creating jar file Job2448999998032650196.jar
2024-04-19 12:57:23,233 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler  jar file Job2448999998032650196.jar created
2024-04-19 12:57:23,248 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler  Setting up single store job
2024-04-19 12:57:23,250 [main] INFO  org.apache.pig.data.SchemaTupleFrontend  Key [pig.schematuple] is false, will not generate code.
2024-04-19 12:57:23,250 [main] INFO  org.apache.pig.data.SchemaTupleFrontend  Starting process to move generated code to distributed cache
2024-04-19 12:57:23,250 [main] INFO  org.apache.pig.data.SchemaTupleFrontend  Setting key [pig.schematuple.classes] with classes to deserialize [l
```

training@localhost:-

File Edit View Search Terminal Help

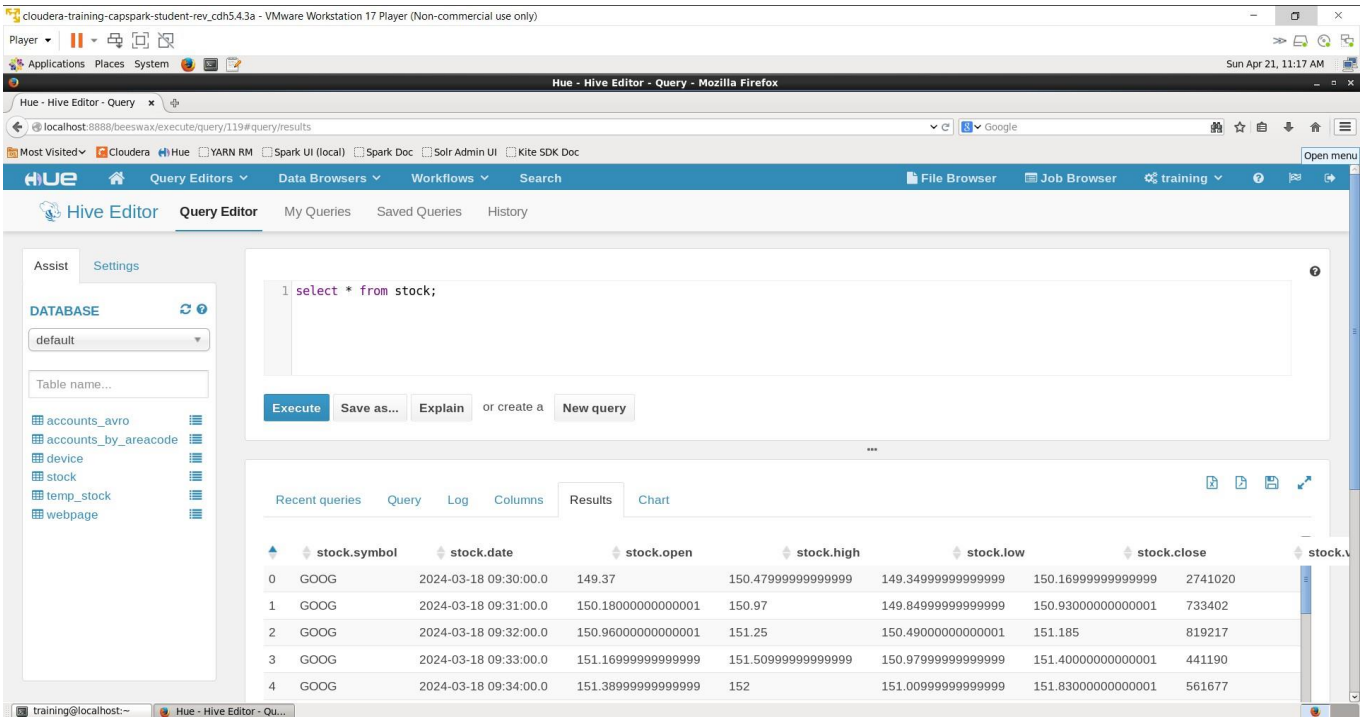
```
HadoopVersion   PigVersion      Userid  StartedAt       FinishedAt      Features
2.6.0-cdhS.4.3  0.12.0-cdhS.4.3 training        2024-04-19 12:57:15     2024-04-19 13:00:10     GROUP_BV,ORDER_BV,LIMIT

Success!

Job Stats (time in seconds):
Jobid   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime
e       MedianReducetime        Alias   Feature outputs
job 1709496148644 0012  1       1       8       8       8       8       8       8       1-112,data,transactions per_blo k   GROUP BV,fOMBINER
job 1709496148644 0013  1               7       7       7       7       7       7       7       1-113   SAMPLER
job-1709496148644-0014  1               8       8       8       8       8       8       8       1-113   ORDER BV,COMBINER
job-1709496148644-0015  1       1       7       7       7       7       7       7       7       1-113,max transactions  hdfs://localhost:8020/tmp/temp-112855710/tmp703680934,

Input(s):
Successfully read 1027 records (294515 bytes) from: "/user/training/output.csv"

Output(s):
successfully stored 1 records (12 bytes) in: "hdfs://localhost:8020/tmp/temp-112855710/tmp703680934"

Counters:
Total records written : 1
Total bytes written : 12
Spillable Memory Manager spill count  0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job 1709496148644 0012  ->      job 1709496148644 0013,
job-1709496148644-0013  ->      job-1709496148644-0014,
job-1709496148644-0014  ->      JOb-1709496148644-0015,
JOb-1709496148644-0015


2024-04-19 13:00:11,741 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:44582. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 13:00:12,747 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:44582. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 13:00:13,750 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:44582. Already tried 2 time(s); retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 13:00:13,870 [main] INFO  org.apache.hadoop.mapred.ClientserviceDelegate - Application state is completed. FinalApplicationstatus=SUCCEEDED. Redirecting to job history server
```


st:~
File Edit View Search Tenninal Help
ried 1 time(s);retry policy is RetryUpToMaximurnCountWithFixedSleep(maxRetries=3,sleepTime=l000 MILLISECONOS)
2024-04-19 13:00:13,750[main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:44582.Already t
ried 2 time(s);retry policy is RetryUpToMaximurnCountWithFixedSleep(maxRetries=3,sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:13,870[main]INFO org.apache.hadoop.rnapred.ClientServiceDelegate - Application state is completed.FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 13:00:15,101[main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:54974.Already t
ried 0 time(s);retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:16,106[main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:54974.Already t
ried 1 time(s);retry policy is RetryUpToMaximurncountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:17,110 [main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:54974.Already t
ried 2 time(s);retry policy is RetryUpToMaximurncountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:17,224[main]INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.FinalApplicati
onstatus=SUCCEEDED. Redirecting to job history server
2024-04-19 13:00:18,509[main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:55378.Already t
ried 0 time(s);retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:19,514 [main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:55378.Already t
ried 1 time(s);retry policy is RetryUpToMaximurnCountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:20,518 [main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:55378.Already t
ried 2 time(s);retry policy is RetryUpToMaximurnCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 13:00:20,631[main]INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 13:00:21,868[main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:49530.Already t
ried 0 time(s);retry policy is RetryUpToMaximumcountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:22,874 [main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:49530.Already t
ried 1 time(s);retry policy is RetryUpToMaximurncountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:23,879 [main]INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:49530.Already t
ried 2 time(s);retry policy is RetryUpToMaximurncountWithFixedSleep(maxRetries=3, sleepTime=l000 MILLISECONDS)
2024-04-19 13:00:23,991[main]INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.FinalApplicati
onstatus=SUCCEEDED. Redirecting to job history server
2024-04-19 13:00:24,199 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered \'la
rning FIELD DISCARDED TYPE CONVERSION FAILED 10 time(sl.
2024-04-1913:00:24,199 [main]INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-04-19 13:00:24,200[main]INFO org.apache.hadoop.conf.Configuration.deprecation ts.default.name is deprecated. Instead,use
fs.defaultFS
2024-04-19 13:00:24,200[main]INFO org.apache.hadoop.conf.Configuration.deprecation mapred.job.trackeris deprecated.Instead, u
se mapreduce.jobtracker.address
2024-04-19 13:00:24,200 [main]INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
code.
2024-04-19 13:00:24,213[main]INFO org.apache.hadoop.mapreduce.lib.input.FileinputFormat - Total input paths to process : 1
2024-04-19 13:00:24,213[main]INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
1
(834900,5058)
grunt>

B. Data Analysis using Hive – Part 1

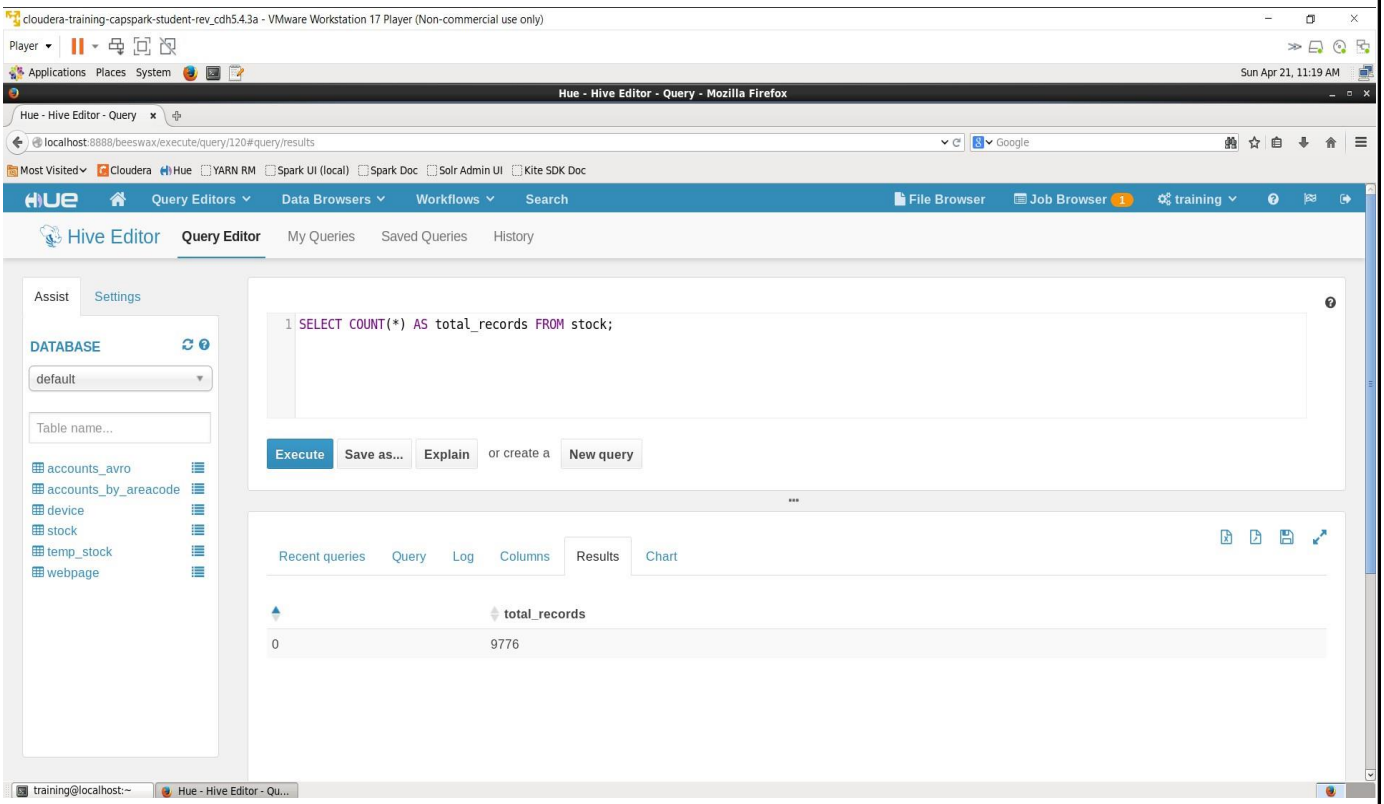
1. Create a table in Hive based on the data you stored on HDFS in step 1.B

```
CREATE TABLE stock (
    symbol STRING,
    date TIMESTAMP,
    open DOUBLE,
    high DOUBLE,
    low DOUBLE,
    close DOUBLE,
    volume BIGINT
)
ROW  FORMAT  DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
LOAD DATA INPATH 'hdfs://localhost:8020/flume/data/' INTO TABLE stock;
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
SELECT * FROM stock;
```



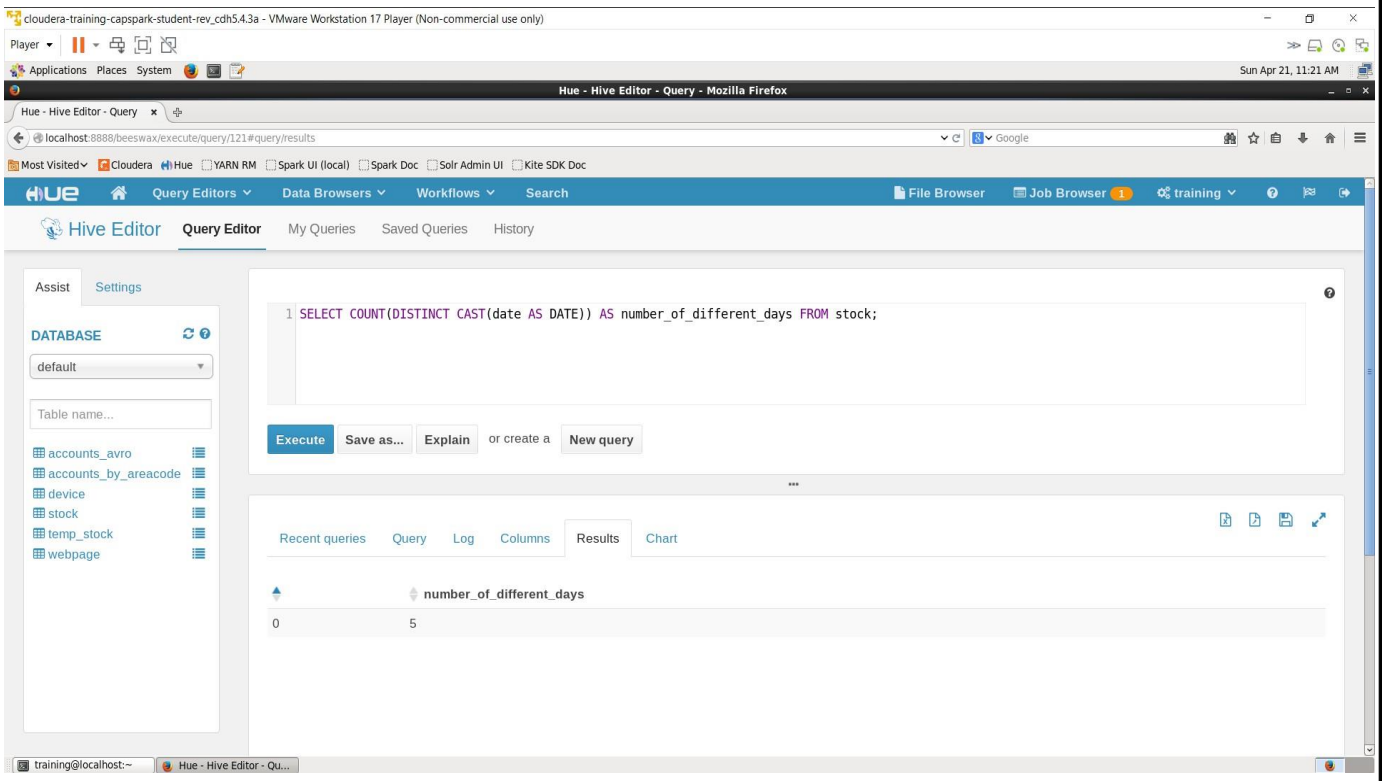
2. a. How many records are there in the table?

SELECT COUNT(*) AS total_records FROM stock;



b. How many different days are there in the table?

SELECT COUNT(DISTINCT CAST(date AS DATE)) AS number_of_different_days FROM stock;



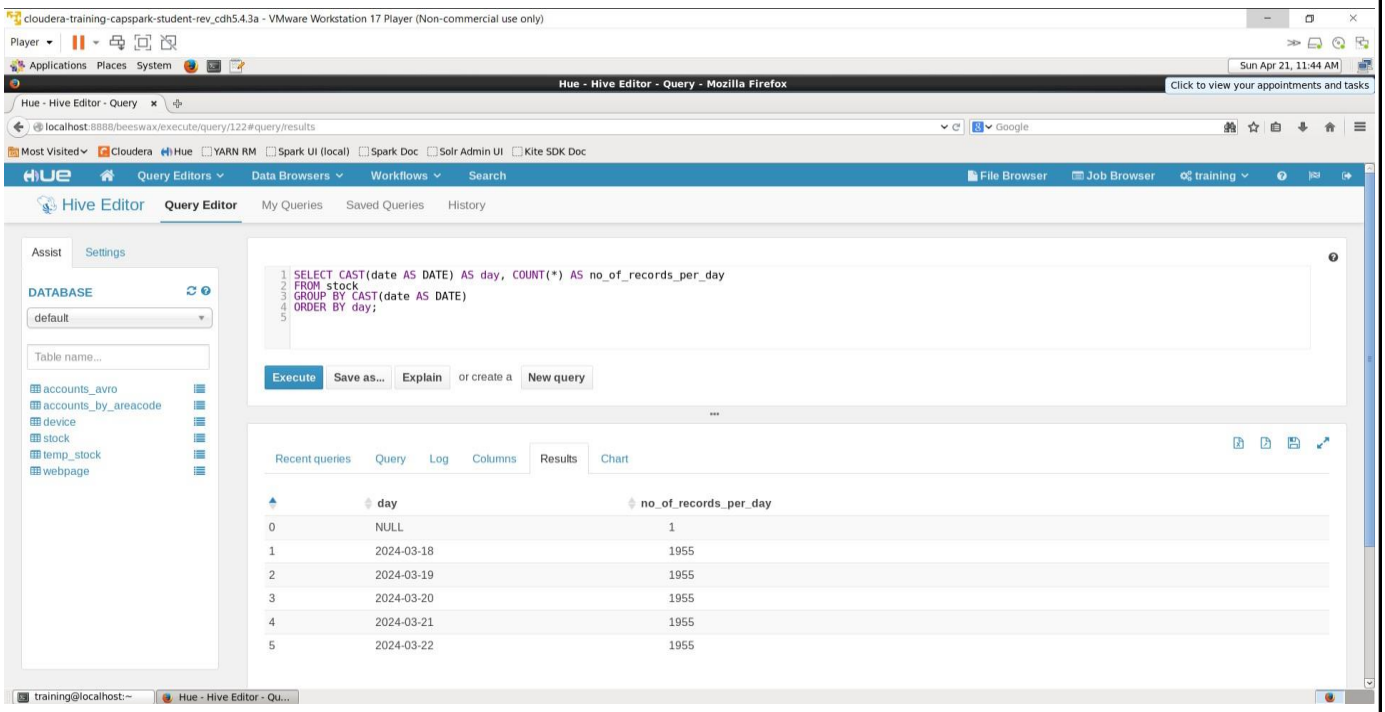
c. How many records per day are there on the table?

SELECT CAST(date AS DATE) AS day, COUNT(*) AS no_of_records_per_day

FROM stock

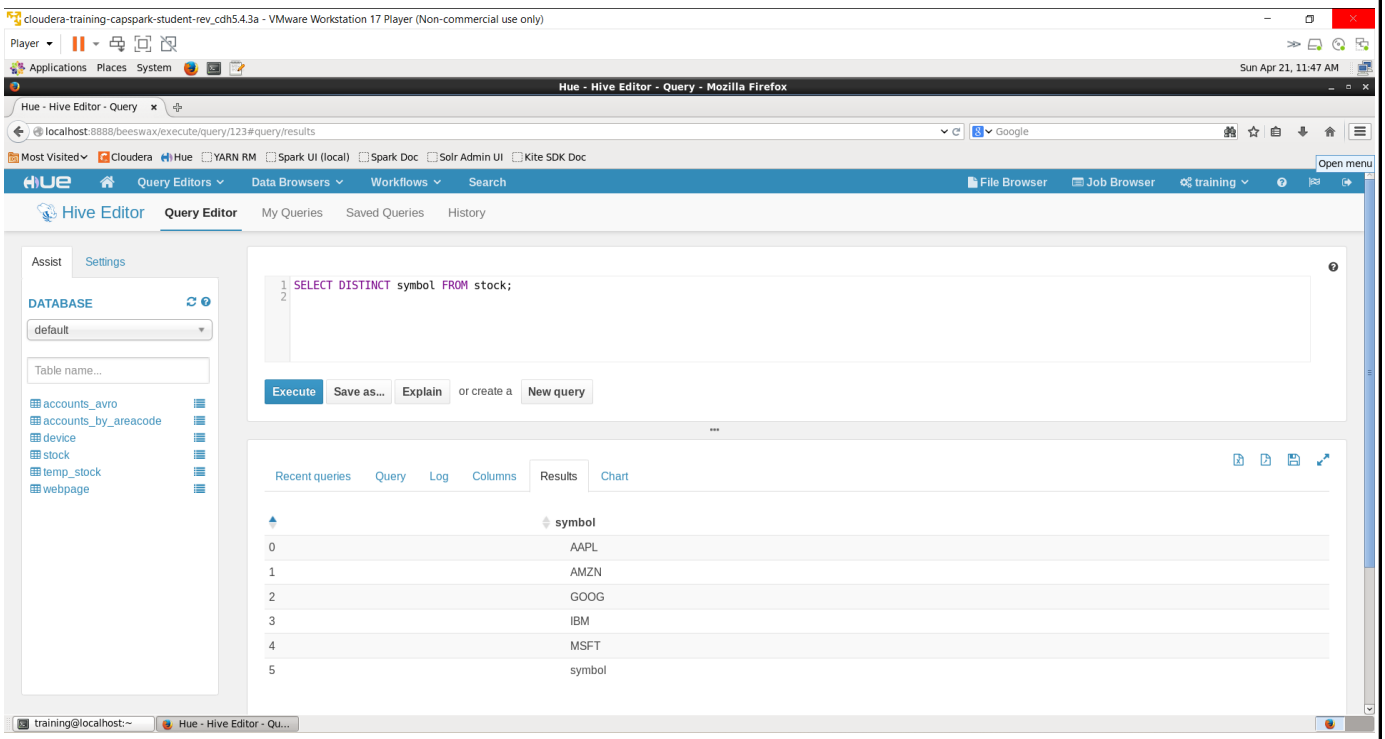
GROUP BY CAST(date AS DATE)

ORDER BY day;



d. What are the symbols on the table?

SELECT DISTINCT symbol FROM stock;

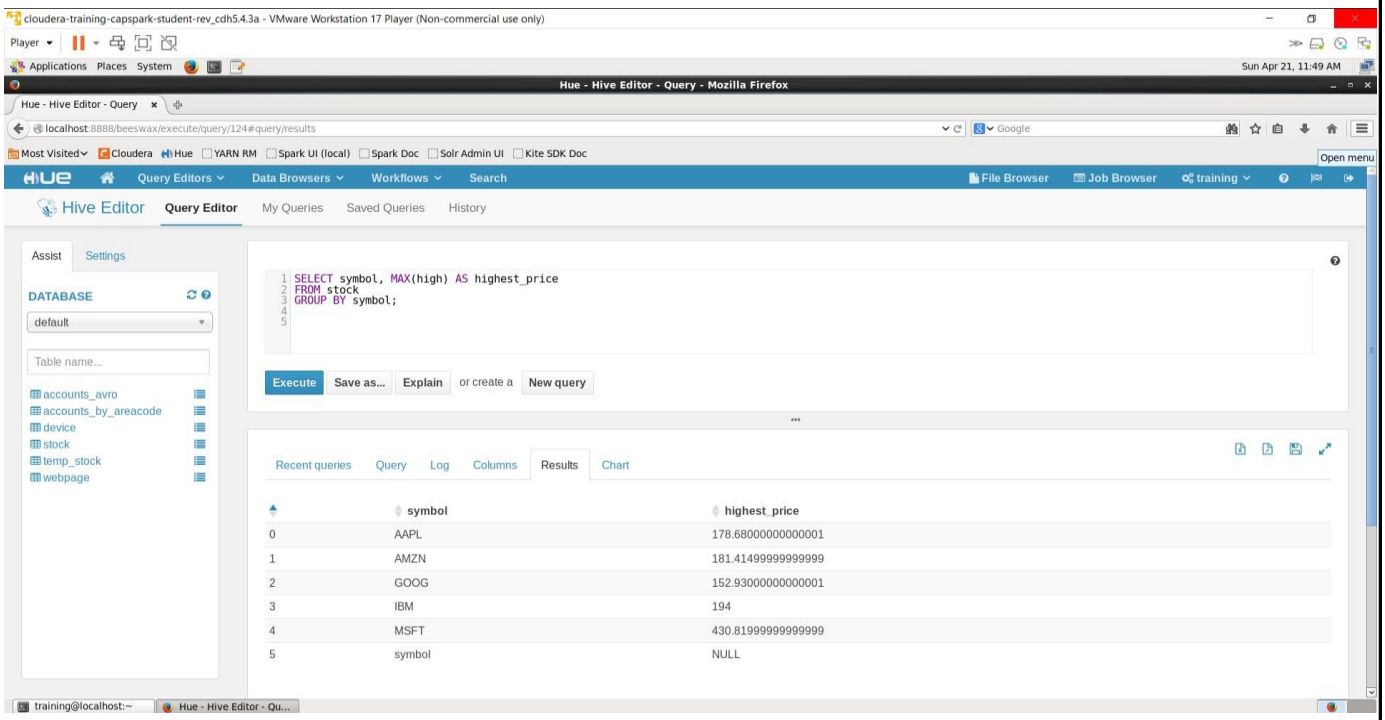


e. What is the highest price for each symbol?

SELECT symbol, MAX(high) AS highest_price

FROM stock

GROUP BY symbol;

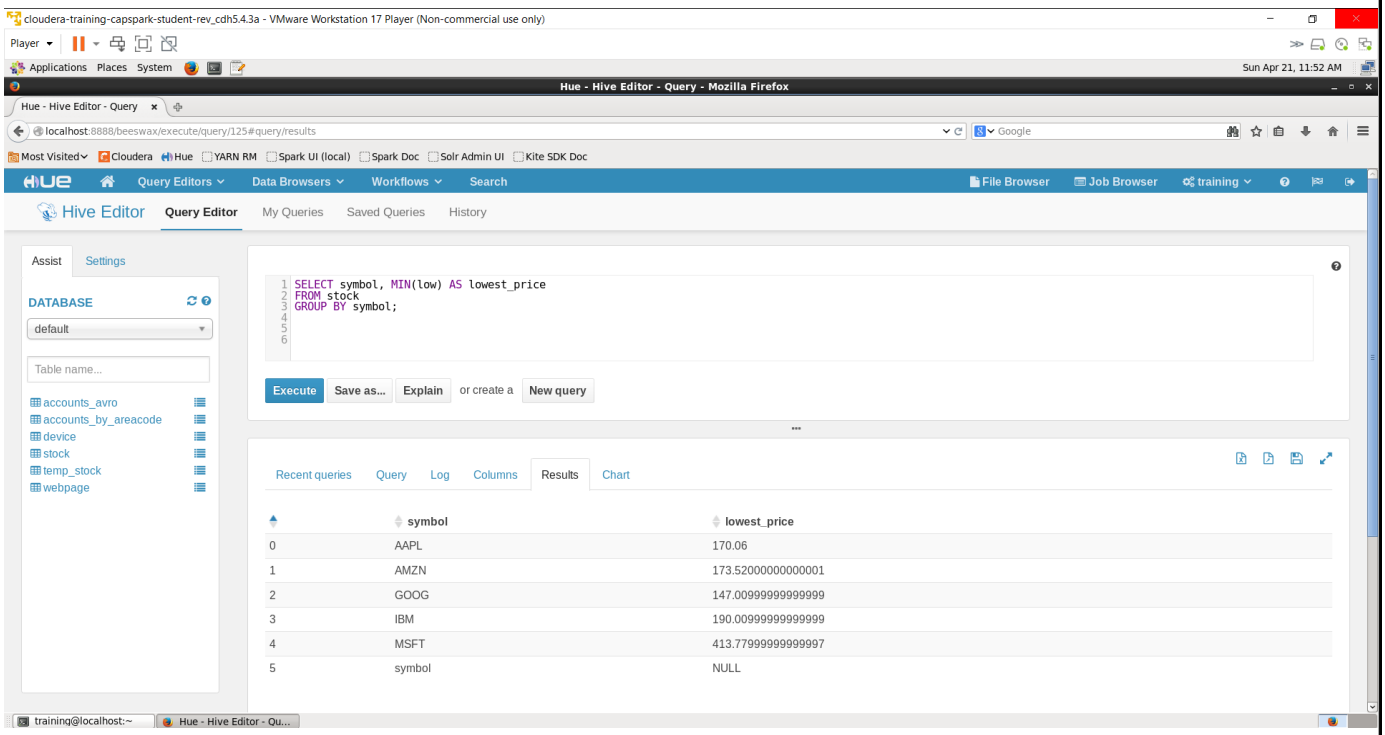


f. What is the lowest price for each symbol?

SELECT symbol, MIN(low) AS lowest_price

FROM stock

GROUP BY symbol;

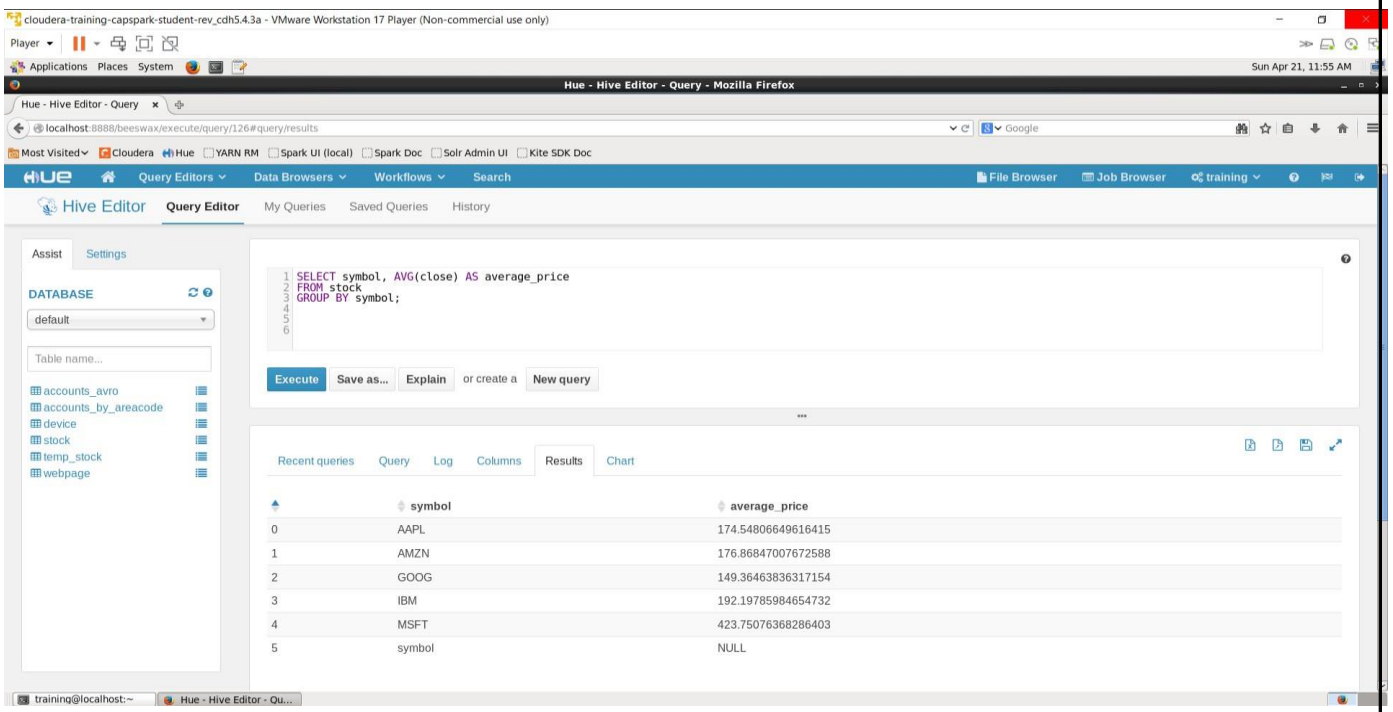


g. What is the average price for each symbol?

SELECT symbol, AVG(close) AS average_price

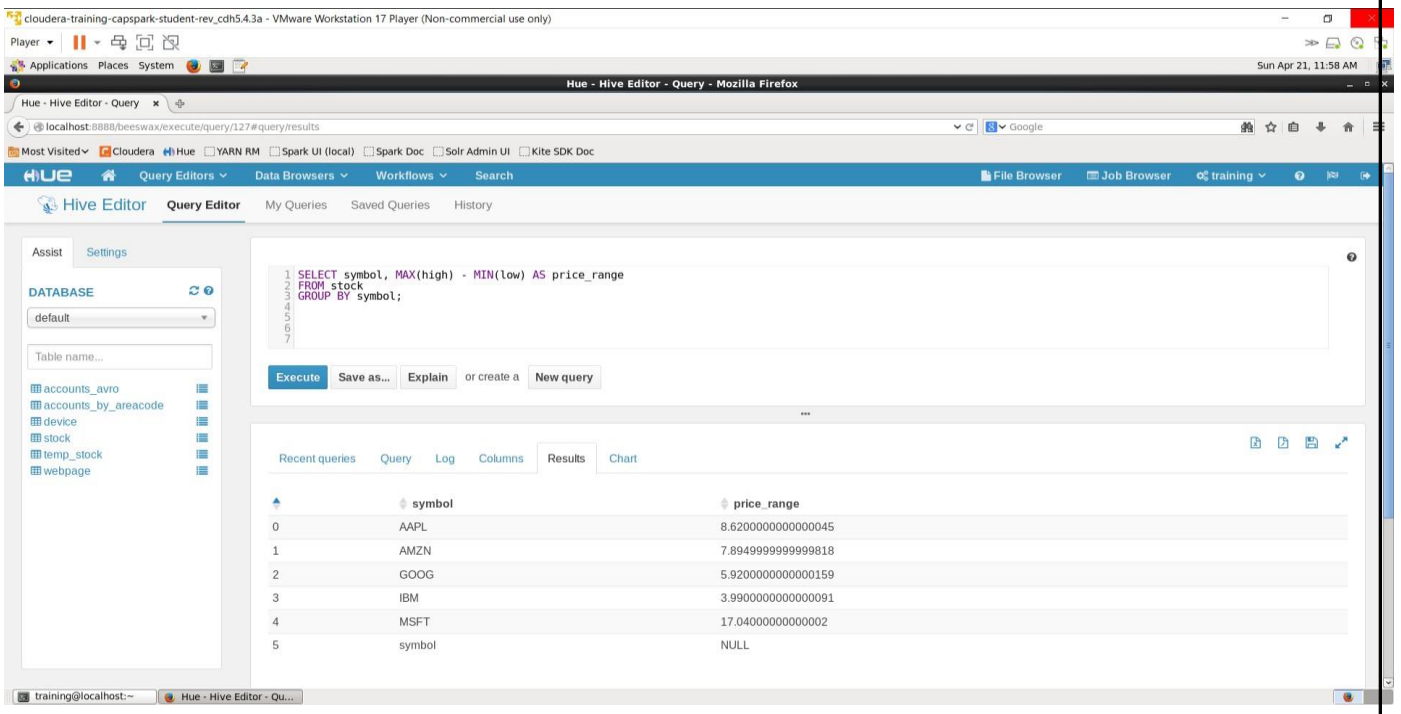
FROM stock

GROUP BY symbol;



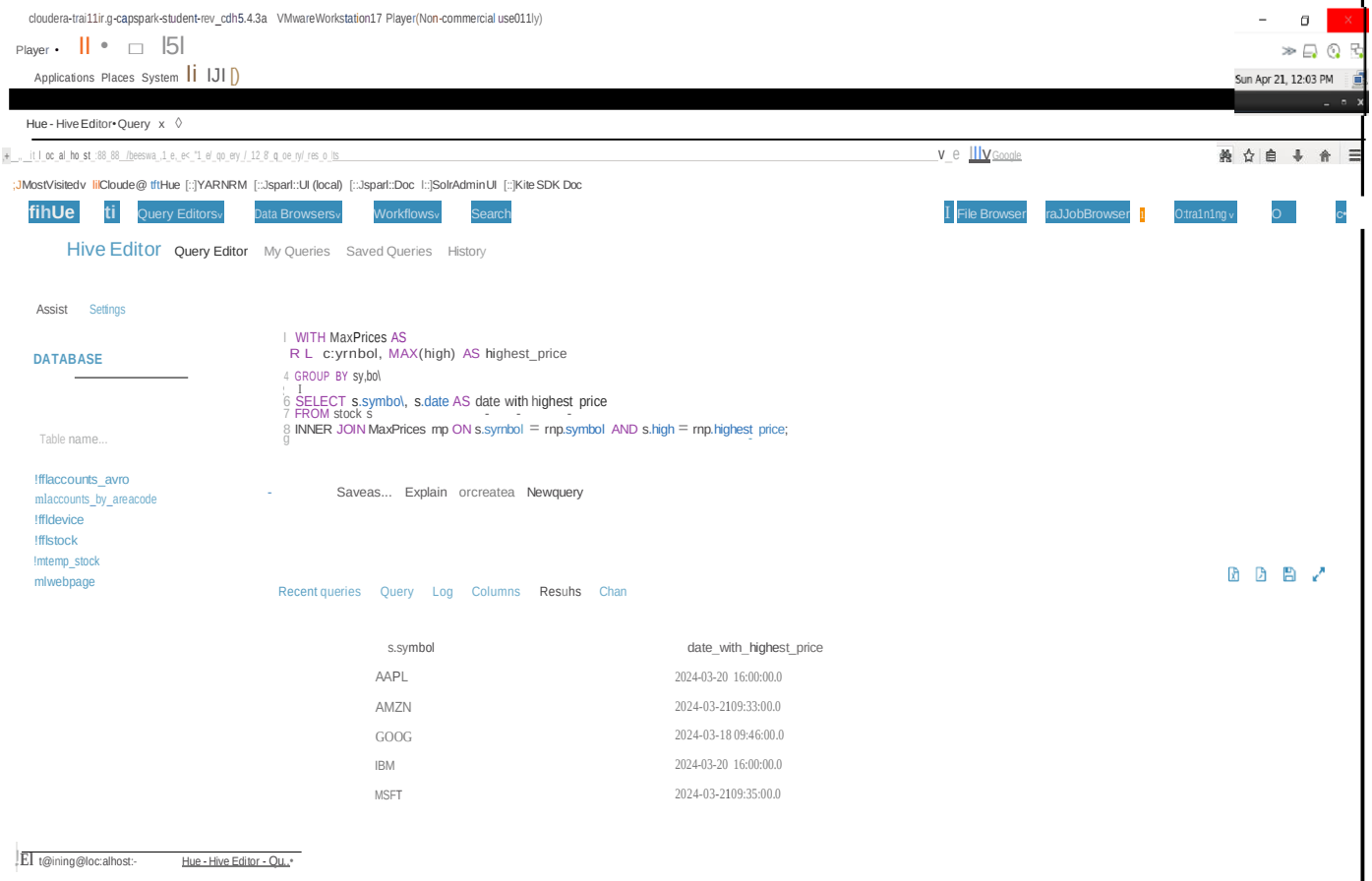
h. What is the range of price for each symbol?

```
SELECT symbol, MAX(high) - MIN(low) AS price_range
FROM stock
GROUP BY symbol;
```



I. What is the date on which each symbol experienced the highest price?

```
WITH MaxPrices AS
( SELECT symbol, MAX(high) AS highest_price
FROM stock
GROUP BY symbol
)
SELECT s.symbol, s.date AS date_with_highest_price
FROM stock s
INNER JOIN MaxPrices mp ON s.symbol = mp.symbol AND s.high = mp.highest_price;
```

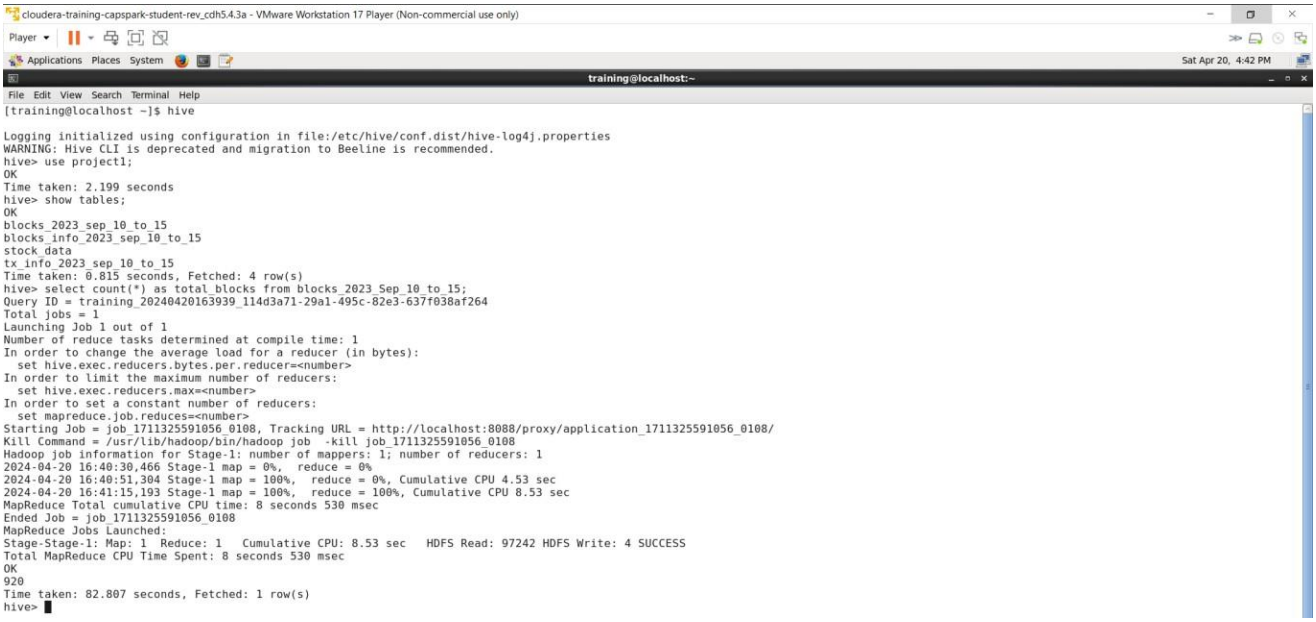

cloudera-trai11ir.g-capspark-student-rev_cdh5.4.3a VMwareWorkstation17 Player(Non-commercial use011ly)
Player
Applications Places System
Sun Apr 21, 12:03 PM
Hue - Hive Editor • Query x
Hue - Hive Editor
Query Editors
Data Browsers
Workflows
Search
File Browser
raJJobBrowser
Training
Query Editor
My Queries
Saved Queries
History
Assist
Settings
DATABASE
Table name...
!fflaccounts_avro
mlaccounts_by_areacode
!ffldevice
!fflstock
!mtemp_stock
mlwebpage
WITH MaxPrices AS
GROUP BY sy,bo\
SELECT s.symbo\, s.date AS date with highest price
FROM stock s
INNER JOIN MaxPrices mp ON s.symbol = mp.symbol AND s.high = mp.highest price;
Saveas... Explain orcreatea Newquery
Recent queries
Query
Log
Columns
Resuhs
Chan
s.symbol
date_with_highest_price
AAPL
2024-03-20 16:00:00.0
AMZN
2024-03-2109:33:00.0
GOOG
2024-03-18 09:46:00.0
IBM
2024-03-20 16:00:00.0
MSFT
2024-03-2109:35:00.0
t@ining@loc:alhost:-
Hue - Hive Editor - Qu..*

C. Data Analysis using Hive – Part 2

1. 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_2023_Sep_10_to_15;



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

SELECT MAX(block_index) as largest_block_height_ FROM blocks_2023_Sep_10_to_15;

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

```
SELECT b.time FROM blocks_2023_Sep_10_to_15 b
JOIN (
SELECT MAX(block_index) AS max_index
FROM blocks_2023_Sep_10_to_15
) maxblocksq
ON b.block_index = maxblocksq.max_index;
```

```
hive> SELECT b.time FROM blocks_2023_Sep_10_to_15 b
    > JOIN (
    > SELECT MAX(block_index) AS max_index
    > FROM blocks_2023_Sep_10_to_15
    > ) maxblocksq
    > ON b.block_index = maxblocksq.max_index;
Query ID = training_20240420164848_39b3121a-5e1e-4ac3-922f-17e37dfc1fc7
Total jobs = 2
Launching Job 1 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_0112, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0112/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1711325591056_0112
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-04-20 16:49:19,741 Stage-1 map = 0%,  reduce = 0%
2024-04-20 16:49:40,136 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.78 sec
2024-04-20 16:49:57,909 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.9 sec
MapReduce Total cumulative CPU time: 8 seconds 900 msec
Ended Job = job_1711325591056_0112
Execution log at: /tmp/training/training_20240420164848_39b3121a-5e1e-4ac3-922f-17e37dfc1fc7.log
2024-04-20 04:50:09     Starting to launch local task to process map join;      maximum memory = 1013645312
2024-04-20 04:50:14     Dump the side-table for tag: 0 with group count: 920 into file: file:/tmp/training/45d2901d-f8e3-40d1-8179-3779a9f05131/hive_2024-04-20_16-48-58_814_4023896937779369930-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile10--.hashtable
2024-04-20 04:50:14     Uploaded 1 File to: file:/tmp/training/45d2901d-f8e3-40d1-8179-3779a9f05131/hive_2024-04-20_16-48-58_814_4023896937779369930-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile10--.hashtable (40938 bytes)
2024-04-20 04:50:14     End of local task; Time Taken: 4.518 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
```

training@localhost:~

```
Launching Job 2 out of 2
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1711325591056_0113, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0113/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1711325591056_0113
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2024-04-20 16:50:33,390 Stage-4 map = 0%,  reduce = 0%
2024-04-20 16:50:49,022 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.56 sec
MapReduce Total cumulative CPU time: 2 seconds 560 msec
Ended Job = job_1711325591056_0113
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.9 sec   HDFS Read: 97036 HDFS Write: 117 SUCCESS
Stage-Stage-4: Map: 1   Cumulative CPU: 2.56 sec   HDFS Read: 5487 HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 460 msec
OK
2023-09-15 00:00:00.0
Time taken: 111.372 seconds, Fetched: 1 row(s)
hive>
```

training@localhost:~

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/
```

training@localhost:~

```
Starting Job = job_1711325591056_0115, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0115/
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 SUCCESS
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
00000000000000000002dbf9d0bc1c743ac17bdb60d5c6abc8cd94f2d253621d        7252
Time taken: 236.629 seconds, Fetched: 1 row(s)
hive>
```

training@localhost:~

Group14_Project3

```
# Sai Madhan Muthyam
# Karan Raturi
# Lakshmi Vinaya Yenumula
# Venkata Rama Anirudh Vikram Kesapragada
# Neelima Nandigam
# Amol Shadane
```

```
sc
```

```
<pyspark.context.SparkContext at 0x7fa6011a7a90>
```

Step 2: Data Analysis Using Spark Core

A. Blockchain Data Analysis – Part 1

```
import json

file_path = "hdfs://localhost/loudacre/blockchain_data.txt"
json_string = sc.textFile(file_path).reduce(lambda x, y: x + y)
data_blocks = json.loads(json_string)
```

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

```
total_blocks = len(data_blocks)
print("Total number of blocks:", total_blocks)
```

```
('Total number of blocks:', 1026)
```

2) What is the largest block height among the blocks in your dataset?

```
largest_block_height = max(block['height'] for block in data_blocks)
print("Largest block height:", largest_block_height)
```

```
('Largest block height:', 835509)
```

3) What is the date and time for that block?

```
block_with_largest_height = next((block for block in data_blocks if block.get('height') == largest_block_height), None)

if block_with_largest_height:
    time_largest_block = block_with_largest_height.get('time')
    print("Time for the largest block:", time_largest_block)
```

```
('Time for the largest block:', 1710924274)
```

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

```
# Initialize variables to store the highest number of transactions and corresponding block height
max_transactions = 0
highest_transactions_block_height = None

# Iterate over the data to find the block with the highest number of transactions
for block in data_blocks:
    num_transactions = block['n_tx']
    if num_transactions > max_transactions:
        max_transactions = num_transactions
        highest_transactions_block_height = block['height']

print("Block with the highest number of transactions:", highest_transactions_block_height)
print("Highest number of transactions in the blocks:", max_transactions)
```

```
('Block with the highest number of transactions:', 834900)
('Highest number of transactions in the blocks:', 5058)
```

B. Stock Market Data Analysis

```
hdfs_folder_path = "hdfs://localhost:/flume/data/"
data_stocks = sc.wholeTextFiles(hdfs_folder_path)
split = data_stocks.flatMap(lambda x: __import__('re').split(r'\r\n|\n', x[1]))
header = split.first()
RDD = split.filter(lambda line: line != header and line != "")
```

1) How many records are there in the table?

```
# Count the number of records
total_records = RDD.count()
print("Total number of records:", total_records)
```

```
('Total number of records:', 9775)
```

2) How many different days are there in the table?

```
unique_days = RDD.map(lambda x: x.split(',')[0]).distinct().count()
print("Number of different days in the table:",unique_days)
```

```
('Number of different days in the table:', 5)
```

3. How many records per each day are there in the table?

```
records_per_day = RDD.map(lambda x: (x.split(',')[0], 1)).reduceByKey(lambda x, y: x + y).collect()
records_per_day
```

```
[(u'GOOG', 1955),
 (u'AAPL', 1955),
 (u'AMZN', 1955),
 (u'MSFT', 1955),
 (u'IBM', 1955)]
```

4. What are the symbols in the table?

```
# Collect distinct symbols
symbols = RDD.map(lambda x: x.split(',')[0]).distinct().collect()
print("Symbols in the table:", symbols)
```

```
('Symbols in the table:', [u'GOOG', u'AAPL', u'AMZN', u'MSFT', u'IBM'])
```

5. What is the highest price for each symbol?

```
# Group data by symbol and find the highest price within each group
highest_price_per_symbol = RDD.map(lambda x: (x.split(',')[0], float(x.split(',')[2]))).\
    groupByKey().\
    mapValues(max).\
    collect()

# Print the result
print("Highest Price per Symbol:")
for symbol, highest_price in highest_price_per_symbol:
    print(symbol, highest_price)
```

```
Highest Price per Symbol:
(u'GOOG', 152.63)
(u'AAPL', 178.66)
(u'AMZN', 181.005)
(u'MSFT', 430.75)
(u'IBM', 193.95)
```

6. What is the lowest price for each symbol?

```
lowest_price_per_symbol = RDD.map(lambda x: (x.split(',')[0], float(x.split(',')[5]))).reduceByKey(min)
print("Lowest price for each symbol:")
for record in lowest_price_per_symbol.collect():
    print(record)
```

```
Lowest price for each symbol:
(u'GOOG', 147.045)
(u'AAPL', 170.075)
(u'AMZN', 173.915)
(u'MSFT', 415.03)
(u'IBM', 190.07)
```

7. What is the average price for each symbol?

```
average_price_per_symbol = RDD.map(lambda x: (x.split(',')[0], (float(x.split(',')[2]), 1))).\
    reduceByKey(lambda x, y: (x[0] + y[0], x[1] + y[1])).\
    map(lambda x: (x[0], x[1][0] / x[1][1])).collect()
average_price_per_symbol
```

```
[(u'GOOG', 149.36608895146657),
 (u'AAPL', 174.54984040920726),
 (u'AMZN', 176.86760358056276),
 (u'MSFT', 423.74892941176471),
 (u'IBM', 192.19820306905365)]
```

8. What is the range of price for each symbol?

```
price_range_per_symbol = RDD.map(lambda x: (x.split(',')[0], (float(x.split(',')[2]), float(x.split(',')[3])))).\
    reduceByKey(lambda x, y: (min(x[0], y[0]), max(x[1], y[1]))).collect()
price_range_per_symbol
```

```
[(u'GOOG', (147.05, 152.93)),
 (u'AAPL', (170.075, 178.68)),
 (u'AMZN', (173.87, 181.415)),
 (u'MSFT', (414.25, 430.82)),
 (u'IBM', (190.24, 194.0))]
```

9. What is the date on which each symbol experienced the highest price?

```
date_highest_price_per_symbol = RDD.map(lambda x: (x.split(',')[0], (float(x.split(',')[2]), x.split(',')[1]))).\
    reduceByKey(lambda x, y: x if x[0] > y[0] else y).collect()
date_highest_price_per_symbol
```

```
[(u'GOOG', (152.63, u'2024-03-18 09:46:00')),
 (u'AAPL', (178.66, u'2024-03-20 16:00:00')),
 (u'AMZN', (181.005, u'2024-03-21 09:33:00')),
 (u'MSFT', (430.75, u'2024-03-21 09:35:00')),
 (u'IBM', (193.95, u'2024-03-20 16:00:00'))]
```

C. Blockchain Data Analysis – Part 2

```
hdfs_path = "/user/hive/warehouse/project1.db"

blocks_rdd = sc.textFile(hdfs_path + "/blocks_2023_sep_10_to_15")
blocks_info_rdd = sc.textFile(hdfs_path + "/blocks_info_2023_sep_10_to_15")
tx_info_rdd = sc.textFile(hdfs_path + "/tx_info_2023_sep_10_to_15")
```

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

```
total_blocks = blocks_rdd.count()
print("Total number of blocks:", total_blocks)
```

```
('Total number of blocks:', 920)
```

2. What is the largest block height among the blocks in your blocks table?

```
largest_block_height = blocks_rdd.map(lambda x: int(x.split('\x01')[3])).max()

print("Largest block height:", largest_block_height)
```

```
('Largest block height:', 807751)
```

3. What is the date and time for that block?

```
largest_block_row = blocks_info_rdd.filter(lambda x: int(x.split('\x01')[9]) == largest_block_height).first()
largest_height_hash = largest_block_row.split('\x01')[1]
print("Largest_height_hash:", largest_height_hash)

matching_row = blocks_rdd.filter(lambda x: x.split('\x01')[1] == largest_height_hash).first()
timestamp = matching_row.split('\x01')[2]

print("Timestamp for the block with the largest height:", timestamp)
```

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

```
block_hashes = tx_info_rdd.map(lambda row: (row.split('\x01')[2], 1))

block_transaction_counts = block_hashes.groupByKey().mapValues(len)

largest_block = block_transaction_counts.top(1, key=lambda x: x[1])

print("Block with the largest number of transactions:")
print("Block Hash:", largest_block[0][0])
print("Number of Transactions:", largest_block[0][1])
```

```
Block with the largest number of transactions:
('Block Hash:', u'000000000000000000002dbf9d0bc1c743ac17bdb60d5c6abc8cd94f2d253621d')
('Number of Transactions:', 7252)
```

Step 3: Data Analysis Using Spark SQL

A. Blockchain Data Analysis – Part 1

```
from pyspark.sql import SQLContext

sqlContext = SQLContext(sc)

blocks_df = sqlContext.createDataFrame(data_blocks)

blocks_df.registerTempTable("blocks")

total_blocks = sqlContext.sql("SELECT COUNT(*) as total_blocks FROM blocks").first().total_blocks
print("Total Blocks: {}".format(total_blocks))
```

```
Total Blocks: 1026
/usr/lib/spark/python/pyspark/sql/context.py:156: UserWarning: Using RDD of dict to inferSchema is deprecated,please use pyspark.sql.Row instead
  warnings.warn("Using RDD of dict to inferSchema is deprecated,"
```

2. What is the largest block height among the blocks in your dataset?

```
max_block_height = sqlContext.sql("SELECT MAX(height) as max_block_height FROM blocks").first().max_block_height
print("Largest Block Height: {}".format(max_block_height))
```

```
Largest Block Height: 835509
```

3. What is the date and time for that block?

```
block_info = sqlContext.sql("SELECT time FROM blocks WHERE height = '{}'".format(max_block_height)).first()
timestamp_value = block_info.time
print("Time for the largest block: {}".format(timestamp_value))
```

```
Time for the largest block: 1710924274
```

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

```
max_transactions = sqlContext.sql("SELECT MAX(n_tx) as max_transactions FROM blocks").first().max_transactions
print("Highest Number of Transactions: {}".format(max_transactions))
```

```
Highest Number of Transactions: 5058
```

B. Stock Market Data Analysis

```
from pyspark.sql import Row
def parse_line(line):
    parts = line.split(',')
    return Row(symbol=parts[0], date=parts[1],open=float(parts[2]), high=float(parts[3]),
               low=float(parts[4]), close=float(parts[5]),
               volume=int(parts[6]))

# Convert RDD to DataFrame
df = sqlContext.createDataFrame(RDD.map(parse_line))

# Register the DataFrame as a temp table
df.registerTempTable("stock_data")

result = sqlContext.sql("SELECT * FROM stock_data")
result.show()
```

```
 close    date                 high    low     open    symbol volume
 172.39  2024-03-22 16:00:00 172.521 172.248 172.37   AAPL   17915930
 172.25  2024-03-22 15:59:00 172.48  172.23  172.479 AAPL   953094
 172.48  2024-03-22 15:58:00 172.49  172.39  172.4   AAPL   274813
 172.405 2024-03-22 15:57:00 172.42  172.33  172.34  AAPL   342157
 172.35  2024-03-22 15:56:00 172.43  172.33  172.37  AAPL   288187
 172.38  2024-03-22 15:55:00 172.47  172.28  172.29  AAPL   375835
 172.3   2024-03-22 15:54:00 172.535 172.26  172.485 AAPL   390478
 172.49  2024-03-22 15:53:00 172.51  172.41  172.42  AAPL   259557
 172.433 2024-03-22 15:52:00 172.47  172.37  172.39  AAPL   176532
 172.4   2024-03-22 15:51:00 172.51  172.3   172.46  AAPL   363401
 172.465 2024-03-22 15:50:00 172.63  172.415 172.58  AAPL   290657
 172.58  2024-03-22 15:49:00 172.638 172.56  172.635 AAPL   140335
 172.63  2024-03-22 15:48:00 172.68  172.56  172.591 AAPL   142526
 172.595 2024-03-22 15:47:00 172.68  172.59  172.67  AAPL   130568
 172.68  2024-03-22 15:46:00 172.68  172.615 172.635 AAPL   146608
 172.635 2024-03-22 15:45:00 172.84  172.6   172.84  AAPL   257051
 172.845 2024-03-22 15:44:00 172.93  172.75  172.9   AAPL   230256
 172.9   2024-03-22 15:43:00 173.02  172.87  172.985 AAPL   244025
 172.99  2024-03-22 15:42:00 173.02  172.94  173.015 AAPL   123976
 173.015 2024-03-22 15:41:00 173.02  172.95  172.96  AAPL   127515
```

1. How many records are there in the table?

```
record_count = sqlContext.sql("SELECT COUNT(*) as record_count FROM stock_data").first().record_count
print("Number of Records: {}".format(record_count))
```

```
Number of Records: 9775
```

2. How many different days are there in the table?

```
distinct_days = sqlContext.sql("SELECT COUNT(DISTINCT SUBSTRING(`date`, 1, 10)) as distinct_days FROM stock_data").first().distinct_days
print("Number of Different Days: {}".format(distinct_days))
```

```
Number of Different Days: 5
```

3. How many records per each day are there in the table?

```
# Execute the SQL query and collect the result
records_per_day = sqlContext.sql("SELECT SUBSTRING(`date`, 1, 10), COUNT(*) FROM stock_data GROUP BY SUBSTRING(`date`, 1, 10)")
records_per_day.show()
```

```
c0          c1
2024-03-18 1955
2024-03-19 1955
2024-03-20 1955
2024-03-21 1955
2024-03-22 1955
```

4. What are the symbols in the table?

```
symbols = sqlContext.sql("SELECT DISTINCT `symbol` as symbol FROM stock_data")
symbols.show()
```

```
symbol
AAPL
IBM
AMZN
MSFT
GOOG
```

5. What is the highest price for each symbol?

```
highest_price = sqlContext.sql("SELECT `symbol` as symbol, MAX(`high`) as highest_price FROM stock_data GROUP BY `symbol`")
highest_price.show()
```

```
symbol highest_price
AAPL  178.68
IBM   194.0
AMZN  181.415
MSFT  430.82
GOOG  152.93
```

6. What is the lowest price for each symbol?

```
lowest_price = sqlContext.sql("SELECT `symbol` as symbol, MIN(`low`) as lowest_price FROM stock_data GROUP BY `symbol`")
lowest_price.show()
```

```
symbol lowest_price
AAPL  170.06
IBM   190.01
AMZN  173.52
MSFT  413.78
GOOG  147.01
```

7. What is the average price for each symbol?

```
average_price = sqlContext.sql("SELECT `symbol` as symbol, AVG(`close`) as average_price FROM stock_data GROUP BY `symbol`")
average_price.show()
```

```
symbol average_price
AAPL  174.54886649616415
IBM   192.19785984654712
AMZN  176.86867007672645
MSFT  423.75076368286466
GOOG  149.36463836317134
```

8. What is the range of price for each symbol?

```
price_range = sqlContext.sql("SELECT `symbol` as symbol, MIN(`close`) as Minimum, MAX(`close`) as Maximum, MAX(`close`) - MIN(`close`) as price_range FROM stock_data GROUP BY `symb
price_range.show()
```

```
symbol Minimum Maximum price_range
AAPL  170.075 178.65  8.575000000000017
IBM   190.07  193.97  3.9000000000000057
AMZN  173.915 180.995 7.080000000000125
MSFT  415.03  430.71  15.680000000000007
GOOG  147.045 152.65  5.605000000000018
```

9. What is the date on which each symbol experienced the highest price?

```
date_highest_price = sqlContext.sql("""
SELECT s.`symbol` as symbol, s.`date` as transaction_date, s.`high` as highest_price
FROM stock_data s
JOIN (
    SELECT `symbol` as symbol, MAX(`high`) as max_price
    FROM stock_data
    GROUP BY `symbol`
) t
ON s.`symbol` = t.symbol AND s.`high` = t.max_price
""")
date_highest_price.show()
```

```
symbol transaction_date   highest_price
IBM   2024-03-20 16:00:00 194.0
AMZN  2024-03-21 09:33:00 181.415
MSFT  2024-03-21 09:35:00 430.82
GOOG  2024-03-18 09:46:00 152.93
AAPL  2024-03-20 16:00:00 178.68
```

C. Blockchain Data Analysis – Part 2

```
blocks = sqlCtx.load(source="jdbc",
                     url="jdbc:mysql://localhost/project1?user=training&password=training",
                     dbtable="blocks_2023_Sep_10_to_15")
```

```
blocks_info = sqlCtx.load(source="jdbc",
                     url="jdbc:mysql://localhost/project1?user=training&password=training",
                     dbtable="blocks_info_2023_Sep_10_to_15")

tx_info = sqlCtx.load(source="jdbc",
                     url="jdbc:mysql://localhost/project1?user=training&password=training",
                     dbtable="tx_info_2023_Sep_10_to_15")
```

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

```
total_blocks = blocks.count()
print("Total number of blocks: {}".format(total_blocks))
```

```
Total number of blocks: 920
```

2. What is the largest block height among the blocks in your blocks table?

```
largest_block_height = blocks.agg({"block_index": "max"}).collect()[0][0]
print("Largest block height:", largest_block_height)
```

```
('Largest block height:', 807751)
```

3. What is the date and time for that block?

```
largest_block_date_time = blocks.filter(blocks["block_index"] == largest_block_height).select("time").collect()[0][0]
formatted_time = largest_block_date_time.strftime("%Y-%m-%d %H:%M:%S")

print("Date and time for the largest block:", formatted_time)
```

```
('Date and time for the largest block:', '2023-09-15 00:00:00')
```

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

```
largest_num_transactions = tx_info.groupBy("block_index").count().agg({"count": "max"}).collect()[0][0]
print("Largest number of transactions in a block:", largest_num_transactions)
```

```
('Largest number of transactions in a block:', 7252)
```

```
```