

BID DATA PROJECT-1

Group-7

Course: BUAN 6346

Term: Fall 2023

Professor: Farzad Kamalzadeh

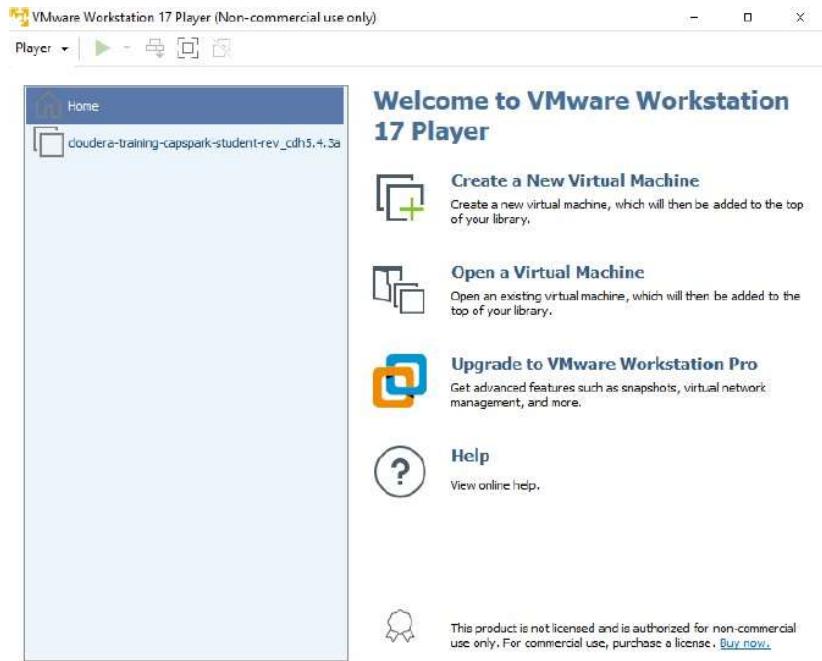


Name: Subhash A R

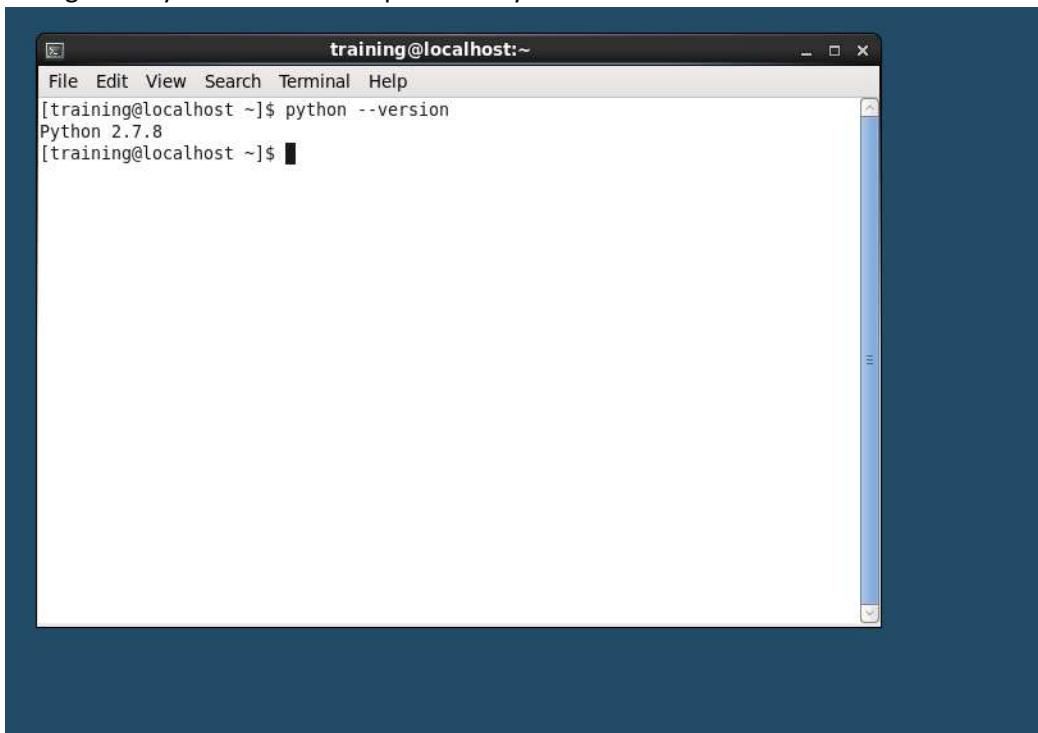
NET ID: SXA210266

Step 0: VM Setup

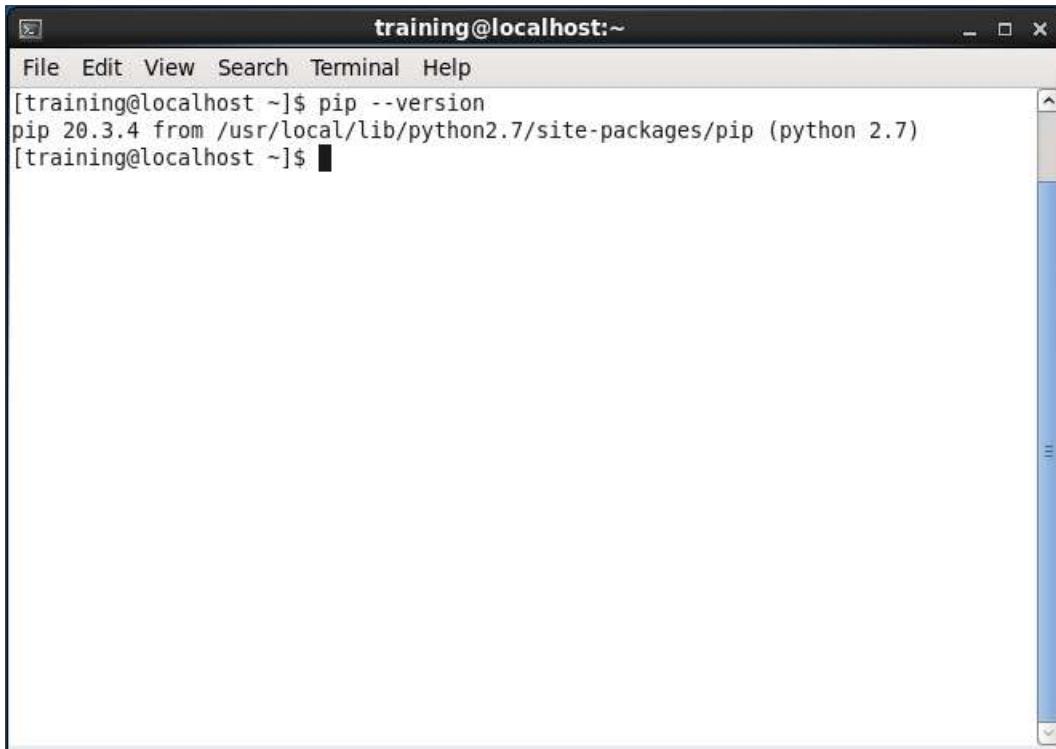
- Create a new virtual machine (VM) from the same image file.



- Followed assignment 1 to enable all the services and set up your Hadoop.
- Change the Python default interpreter to Python2.7 and check version.

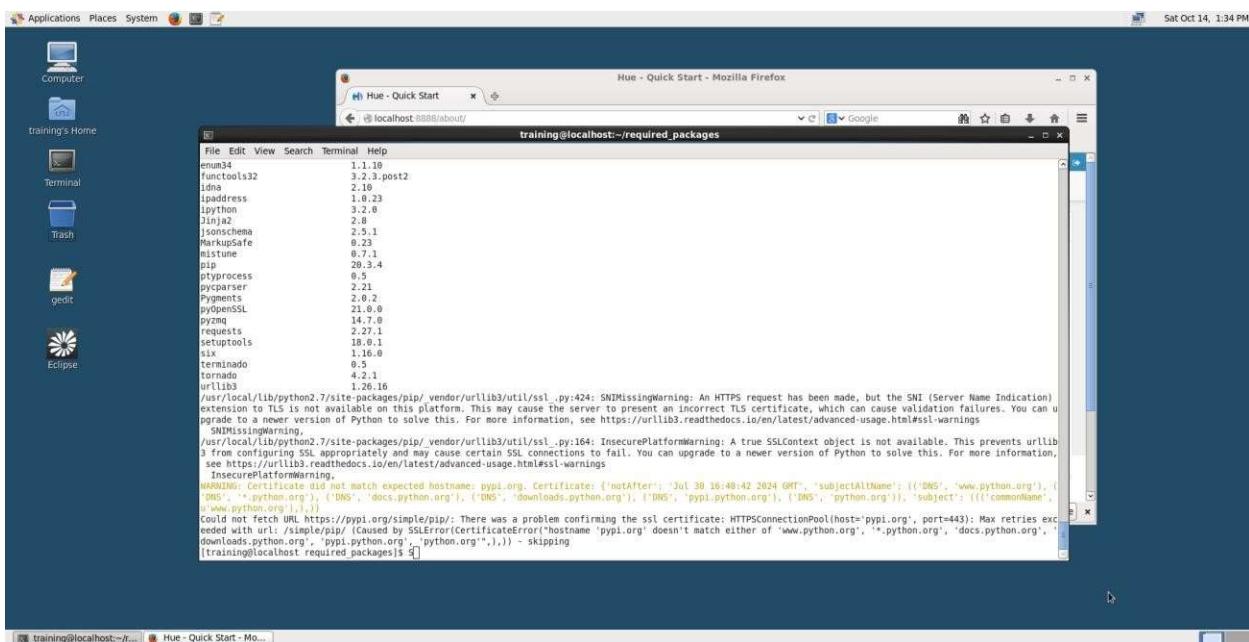


- Upgrade the pip and check version.



```
training@localhost:~$ pip --version
pip 20.3.4 from /usr/local/lib/python2.7/site-packages/pip (python 2.7)
[training@localhost ~]$
```

- Verify that all the packages have been installed by running the following: pip list



The screenshot shows a Linux desktop environment with a terminal window and a web browser window.

The terminal window (bottom-left) displays the command "pip list" and its output:

```
File Edit View Search Terminal Help
[training@localhost ~]$ pip list
enum34
functools32
idna
ipaddress
ipython
Jinja2
jsonschema
MarkupSafe
mistune
pip
pyprocess
pyranges
Pygments
pyOpenSSL
pyzmq
requests
setuptools
six
terminado
tornado
urllib3
[training@localhost ~]$
```

The browser window (top-right) shows the "Hue - Quick Start" page at "http://localhost:8888/about/". It contains a table of required packages and some SSL warning text.

Package	Version
enum34	1.1.10
functools32	3.2.3.post2
idna	2.18
ipaddress	1.0.23
ipython	3.2.8
Jinja2	2.8
jsonschema	2.3.1
MarkupSafe	0.23
mistune	0.7.1
pip	20.3.4
pyprocess	0.5
pyranges	2.21
Pygments	2.0.2
pyOpenSSL	21.0.0
pyzmq	14.7.0
requests	2.27.1
setuptools	18.0.1
six	1.16.0
terminado	0.5
tornado	4.2.1
urllib3	1.26.16

The browser also displays SSL-related error messages:

```
/usr/local/lib/python2.7/site-packages/pip/_vendor/urllib3/util/ssl_.py:424: SNIMissingWarning: An HTTPS request has been made, but the SNIs (Server Name Indication) extension to TLS is not available on this platform. This may cause the server to present an incorrect TLS certificate, which can cause validation failures. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
SNIMissingWarning: /usr/local/lib/python2.7/site-packages/pip/_vendor/urllib3/util/ssl_.py:164: InsecurePlatformWarning: A true SSLSocket object is not available. This prevents urllib3 from configuring SSL appropriately, and may cause certain SSL connections to fail. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
InsecurePlatformWarning,
urllib3._collections.CertVerificationError: [hostname 'pypi.org' did not match expected hostname, pypi.org, Certificate ('notAfter': 'Jul 30 16:49:42 2024 GMT', 'subjectAltName': [{}], 'subject': [{}], 'commonName': 'www.python.org'), ('DNS', 'docs.python.org'), ('DNS', 'downloads.python.org'), ('DNS', 'pypi.python.org'), ('DNS', 'python.org')], 'subject': [{}], 'commonName': 'www.python.org',])
Caused by: ssl.SSLError(CertificateError('hostname "pypi.org" doesn't match either of "www.python.org", "*.python.org", "docs.python.org", "downloads.python.org", "pypi.python.org", "python.org", .')) - skipping
[training@localhost ~]$
```

Step 1: Data Ingestion

A. Direct File Transfer

Python code

```
import requests
from datetime import datetime, timedelta

# Function to pull block data for a specific date
def get_block_data(date):
    timestamp = int(date.timestamp()) * 1000
    url = f'https://blockchain.info/blocks/{timestamp}?format=json'
    response = requests.get(url)
    if response.status_code == 200:
        return response.json()
    else:
        print(f"Failed to fetch data for {date}")
        return None

# Specify the start date
start_date = datetime(2023, 10, 23)

# Create a list to store data for seven days
block_data_list = []

# Loop through seven days
for _ in range(7):
    daily_data = get_block_data(start_date)
    if daily_data:
        block_data_list.append(daily_data)
    start_date -= timedelta(days=1)

# Create a formatted text file
output_filename = "block_data_formatted.txt"

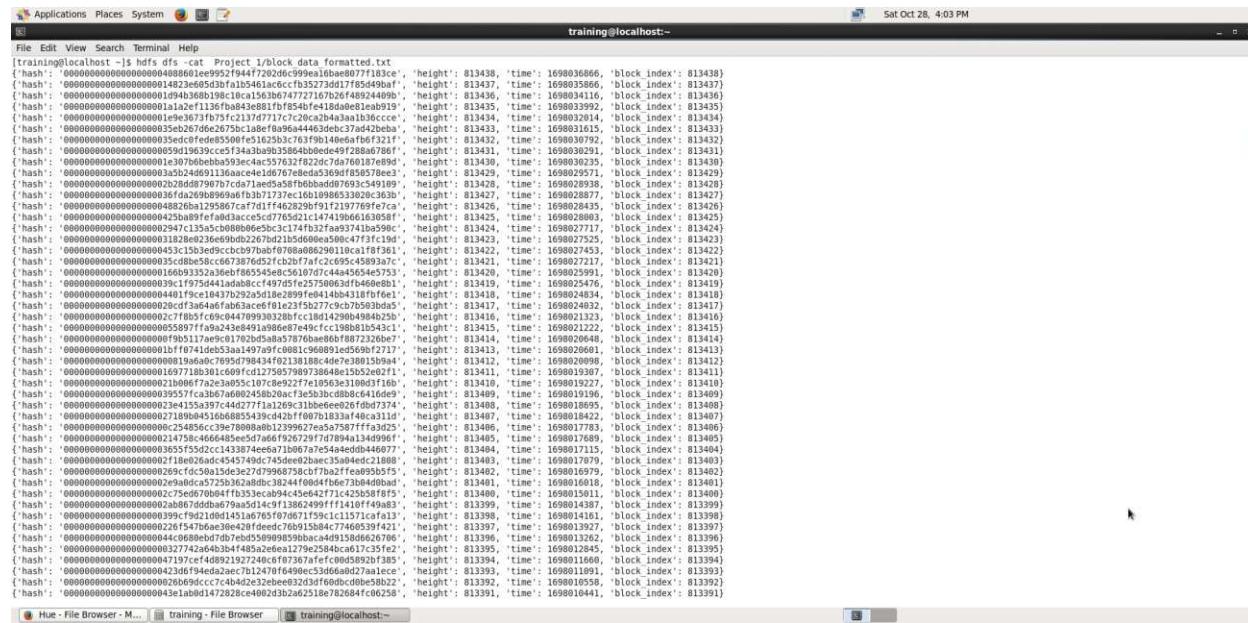
# Write each block's information to the text file
with open(output_filename, "w") as file:
    for daily_data in block_data_list:
        for block_info in daily_data:
            file.write(str(block_info) + "\n")

print(f"Data saved to {output_filename}")

✓ 9.1s
Data saved to block_data_formatted_data.txt
```


To check the contents of data in HDFS using following command

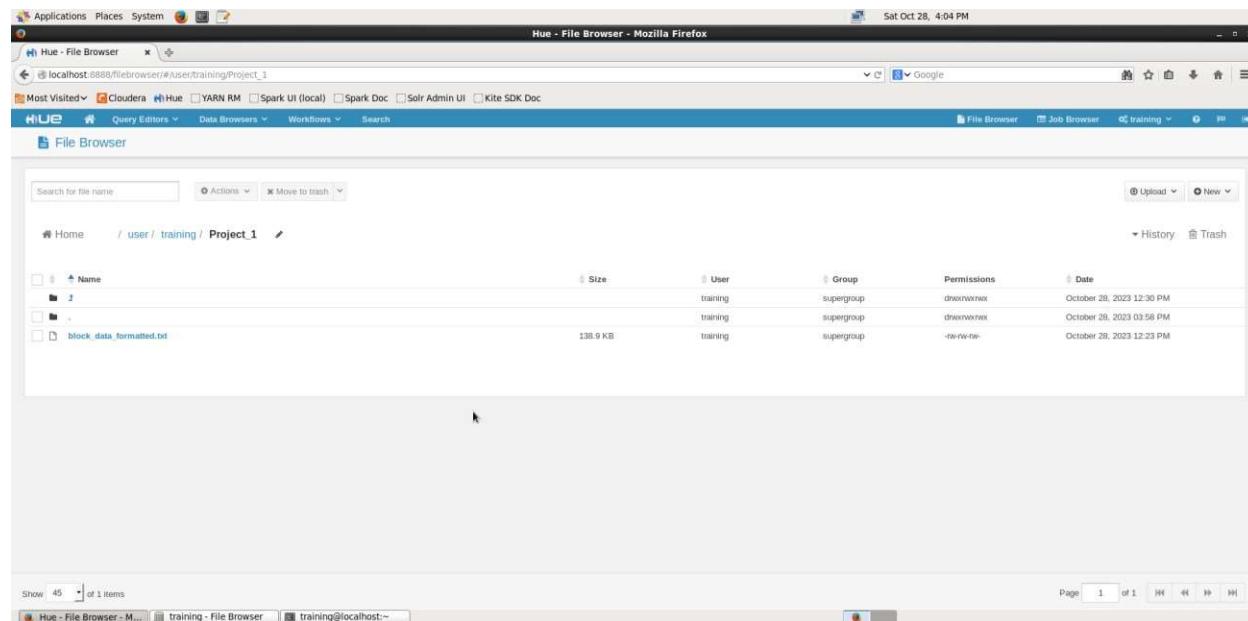
```
$ hdfs dfs -cat Project_1/block_data_formatted.txt
```



```
training@localhost: ~$ hdfs dfs -cat Project_1/block_data_formatted.txt
(training@localhost: ~$ hdfs dfs -ls hdfs://192.168.0.100:8020/user/training/Project_1/block_data_formatted.txt
drwxrwxrwx 1 training 100 0 Oct 28 11:27 Project_1/
drwxrwxrwx 1 training 100 0 Oct 28 11:27 Project_1/block_data_formatted.txt
-rw-r--r-- 1 training 100 1893438 Oct 28 11:27 Project_1/block_data_formatted.txt
Sat Oct 28, 4:03 PM
```

The terminal output shows the directory structure and then lists the contents of the file 'block_data_formatted.txt'. The file contains approximately 200 lines of data, each representing a block in HDFS. Each line includes the offset, size, and a long hex-based checksum.

File In Hue:



The screenshot shows the Hue File Browser interface. The URL is 'localhost:8888/filebrowser/#/user/training/Project_1'. The page displays a list of files in the 'Project_1' directory. One file, 'block_data_formatted.txt', is highlighted. The table has columns for Name, Size, User, Group, Permissions, and Date. The permissions for 'block_data_formatted.txt' are listed as '-rw-r--r--'.

Name	Size	User	Group	Permissions	Date
2		training	supergroup	drwxrwxrwx	October 28, 2023 12:30 PM
		training	supergroup	drwxrwxrwx	October 28, 2023 03:58 PM
block_data_formatted.txt	189.0 KB	training	supergroup	-rw-r--r--	October 28, 2023 12:23 PM

Data in Hue:

The screenshot shows the Hue File Browser interface. At the top, there's a navigation bar with tabs for 'File Browser', 'Job Browser', and 'training'. Below the navigation bar, the main content area displays a list of file blocks for the file 'block_data_formatted.txt'. Each block is represented by a JSON object with the following fields:

- 'hash': A long string of characters representing the block's hash.
- 'height': An integer value representing the height of the block.
- 'time': A timestamp indicating when the block was created.
- 'block_index': An integer index for the block.

The list is very long, containing over 35 blocks. The browser has a standard UI with a left sidebar for actions like 'Edit file' and 'Download', and a right sidebar for navigation and search.

B. Stream Ingestion using Flume.

- API key from Alpha Vantage:

api_key = "IC61TNV1GKNGZL69"

- Python code to pull data from API

```

✓ import requests
import json
import socket
import time

api_url = "https://www.alphavantage.co/query"
api_key = "IC61TNV1GKNGZL69"
symbols = ["AAPL", "IBM", "GOOGL", "MSFT", "TSLA"]

✓ while True:
✓   for symbol in symbols:
✓     params = {
        "function": "TIME_SERIES_INTRADAY", # Use lowercase "function"
        "symbol": symbol, # Use the current symbol in the loop
        "interval": "1min",
        "apikey": api_key,
      }
      response = requests.get(api_url, params=params)
      data = response.json()

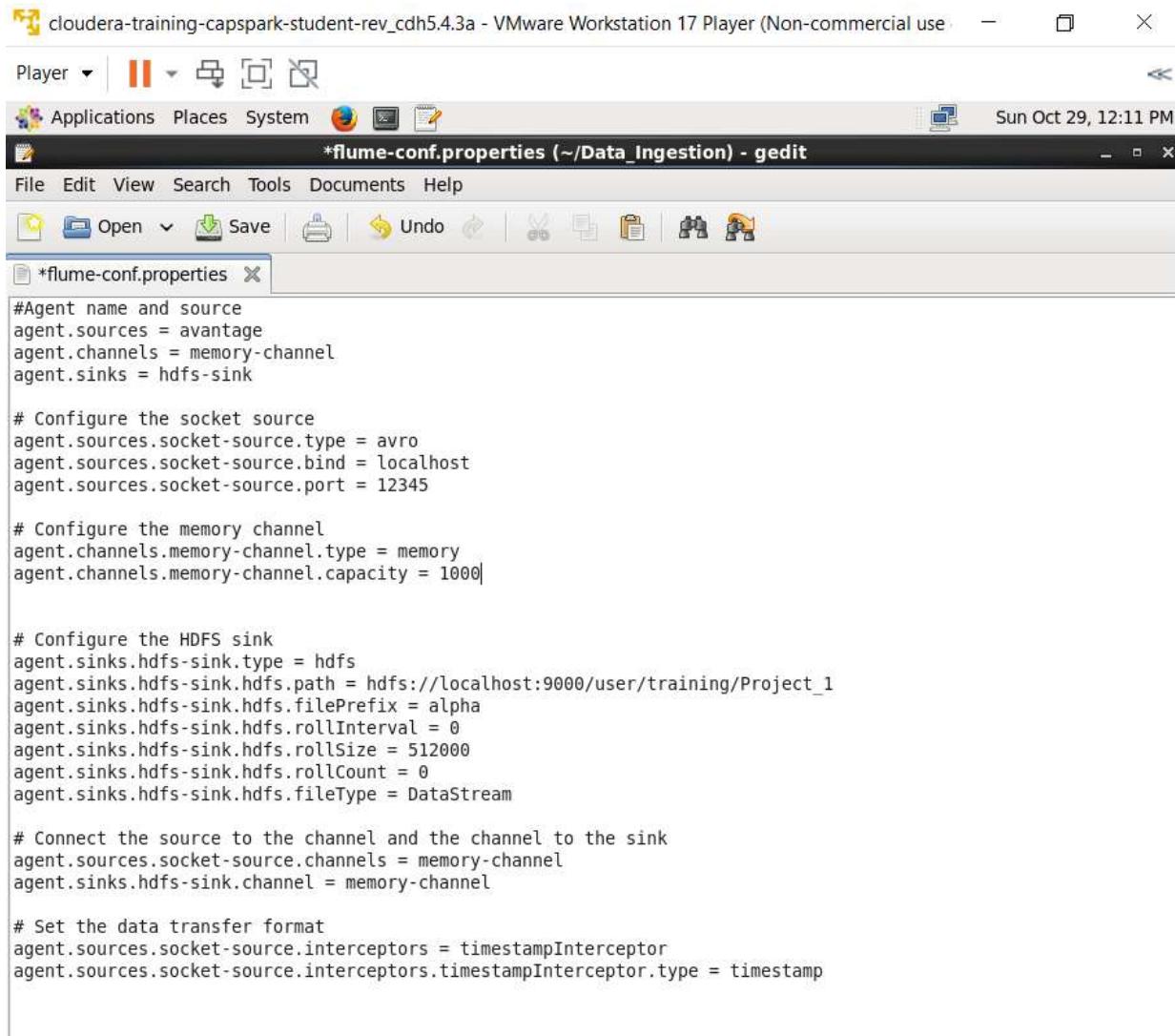
      print(json.dumps(data, indent=2))

      # Send the stock data to the local server on port 12345
      s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
      s.connect(("localhost", 12345))
      s.send(json.dumps(data).encode())

      # Wait for 1 minute before downloading the next batch of data
      time.sleep(60)
    
```

- Screenshot of console output:

```
{
  "Meta Data": {
    "1. Information": "Intraday (1min) open, high, low, close prices and volume",
    "2. Symbol": "AAPL",
    "3. Last Refreshed": "2023-10-27 19:59:00",
    "4. Interval": "1min",
    "5. Output Size": "Compact",
    "6. Time Zone": "US/Eastern"
  },
  "Time Series (1min)": {
    "2023-10-27 19:59:00": {
      "1. open": "168.0110",
      "2. high": "168.0200",
      "3. low": "168.0100",
      "4. close": "168.0200",
      "5. volume": "432"
    },
    "2023-10-27 19:58:00": {
      "1. open": "168.0250",
      "2. high": "168.0400",
      "3. low": "168.0100",
      "4. close": "168.0200",
      "5. volume": "74"
    },
    "2023-10-27 19:57:00": {
      "1. open": "168.0200",
      "2. high": "168.0350",
      "3. low": "168.0150",
      "4. close": "168.0250",
      "5. volume": "123"
    }
  }
}
```

The screenshot shows a VMware Workstation Player window with a gedit text editor. The title bar reads "cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use)". The gedit window has a toolbar with icons for Open, Save, Undo, and Print. The menu bar includes File, Edit, View, Search, Tools, Documents, and Help. The main text area contains the following configuration file:

```
#Agent name and source
agent.sources = avantage
agent.channels = memory-channel
agent.sinks = hdfs-sink

# Configure the socket source
agent.sources.socket-source.type = avro
agent.sources.socket-source.bind = localhost
agent.sources.socket-source.port = 12345

# Configure the memory channel
agent.channels.memory-channel.type = memory
agent.channels.memory-channel.capacity = 1000

# Configure the HDFS sink
agent.sinks.hdfs-sink.type = hdfs
agent.sinks.hdfs-sink.hdfs.path = hdfs://localhost:9000/user/training/Project_1
agent.sinks.hdfs-sink.hdfs.filePrefix = alpha
agent.sinks.hdfs-sink.hdfs.rollInterval = 0
agent.sinks.hdfs-sink.hdfs.rollSize = 512000
agent.sinks.hdfs-sink.hdfs.rollCount = 0
agent.sinks.hdfs-sink.hdfs.fileType = DataStream

# Connect the source to the channel and the channel to the sink
agent.sources.socket-source.channels = memory-channel
agent.sinks.hdfs-sink.channel = memory-channel

# Set the data transfer format
agent.sources.socket-source.interceptors = timestampInterceptor
agent.sources.socket-source.timestampInterceptor.type = timestamp
```

C. Data Ingestion using Sqoop.

STEP 2 & 3

Table Uploaded and verified: -

- **Table 1: (blocks_2023_Sep_10_to_15.csv):**

```
mysql> LOAD DATA LOCAL INFILE '/home/training/Desktop/Project 1/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;
ERROR 2 (HY000): File '/home/training/Desktop/Project 1/blocks_2023_Sep_10_to_15.csv' not found (Errcode: 2)
mysql> LOAD DATA LOCAL INFILE '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, 919 rows affected (0.01 sec)
Records: 920 Deleted: 0 Skipped: 1 Warnings: 0
```

Verify the uploads using MYSQL.

```
mysql> SELECT * FROM blocks_2023_Sep_10_to_15 LIMIT 10;
+----+-----+-----+-----+
| id | hash                                | time          | block_index |
+----+-----+-----+-----+
| 1  | 00000000000000000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4 | 2023-09-10 00:00:00 | 806982 |
| 2  | 00000000000000000000000000000000372109b9a114633512587c8b074910a4bc02921828b59 | 2023-09-10 00:00:00 | 806980 |
| 3  | 00000000000000000000000000000000438def0efe8257f6ff025665074dbcf7e1fda070aff30 | 2023-09-10 00:00:00 | 806979 |
| 4  | 0000000000000000000000000000000022e8847528953ebe30fc81867dc9ab70b50ce2660d5df | 2023-09-10 00:00:00 | 806978 |
| 5  | 0000000000000000000000000000000051355825a2eee832e8c068a3aa355ac6a28bed2d29472 | 2023-09-10 00:00:00 | 806977 |
| 6  | 00000000000000000000000000000000934a50f928c4572cc5d32276ffb55bd8dced508d728d | 2023-09-10 00:00:00 | 806976 |
| 7  | 000000000000000000000000000000002914218e992201381ea1fb26cb2fa963a50df844051b6 | 2023-09-10 00:00:00 | 806975 |
| 8  | 0000000000000000000000000000000047ea53323c0e8c8609f5dfea221fef380d20a12978d8 | 2023-09-10 00:00:00 | 806974 |
| 9  | 000000000000000000000000000000004f5d5af8c8532c5560a2e3f456d9f9fef7a867f3e7c8c | 2023-09-10 00:00:00 | 806973 |
| 10 | 00000000000000000000000000000044e257ee716989fb570589520ead0e9e05fe3458bf155 | 2023-09-10 00:00:00 | 806972 |
+----+-----+-----+-----+
10 rows in set (0.00 sec)
```

SQL CODE:

-- Create a table for File 1

```
CREATE TABLE blocks_2023_Sep_10_to_15 (
    id INT AUTO_INCREMENT PRIMARY KEY,
    hash VARCHAR(64),
    time DATETIME,
    block_index INT
);
```

-- Load data from File 1

```
LOAD DATA LOCAL INFILE '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; -- This skips the header line in the CSV
```

- Table 2: (blocks_info_2023_Sep_10_to_15.csv):**

```
mysql> CREATE TABLE blocks_info_2023_Sep_10_to_15 (
->     id INT AUTO_INCREMENT PRIMARY KEY,
->     hash VARCHAR(64),
->     ver INT,
->     bits INT,
->     fee INT,
->     nonce INT,
->     size INT,
->     block_index INT,
->     main_chain BOOLEAN,
->     height INT,
->     weight INT
-> );
Query OK, 0 rows affected (0.00 sec)

mysql> LOAD DATA LOCAL INFILE '/home/training/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, 309 rows affected, 448 warnings (0.01 sec)
Records: 310 Deleted: 0 Skipped: 1 Warnings: 138
```

Verify the uploads using MySQL and Sqoop

```
mysql> SELECT * FROM blocks_info_2023_Sep_10_to_15 LIMIT 10;
+----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| id | hash | ver | bits | fee | nonce | size | block_index | main_chain | height | weight |
+----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | 00000000000000000000000000000000 | 545259520 | 386216622 | 17380913 | 2147483647 | 1733081 | 806982 | 0 | 806982 | 3993482 | |
| 2 | 00000000000000000000000000000000 | 547356672 | 386216622 | 15397499 | 2147483647 | 1797458 | 806980 | 0 | 806980 | 3993365 |
| 3 | 00000000000000000000000000000000 | 0438def0fe8257f6ff02565074dbc7fe1fda070af30 | 887709696 | 386216622 | 14746325 | 2147483647 | 1913649 | 806979 | 0 | 806979 | 3993468 |
| 4 | 00000000000000000000000000000000 | 622ze8847528953eb03fc81867dc9ab70b50ce26660d5f | 536879192 | 386216622 | 17157713 | 700634169 | 1695531 | 806978 | 0 | 806978 | 3993549 |
| 5 | 00000000000000000000000000000000 | 051355825a2ee832e8c068aaa355ac6a28bed2d29472 | 356928256 | 386216622 | 14267499 | 696643341 | 1117137 | 806977 | 0 | 806977 | 3992862 |
| 6 | 00000000000000000000000000000000 | 0934a50f928e4572cc5d32276fb55bd8d0d5080728d | 587145216 | 386216622 | 16969184 | 2004003 | 806976 | 0 | 806976 | 3993648 |
| 7 | 00000000000000000000000000000000 | 2914218e992201381e1fb26cb2fa963a50df844051b6 | 356928256 | 386216622 | 16982431 | 1739931405 | 1466355 | 806975 | 0 | 806975 | 3993528 |
| 8 | 00000000000000000000000000000000 | 47ea53323c0e8c8609f5dfea221fef380d20a12978d8 | 538157056 | 386216622 | 15046341 | 1066014781 | 1941877 | 806974 | 0 | 806974 | 3993634 |
| 9 | 00000000000000000000000000000000 | 04f5dsfafc8532c5560a2e3f456d9f9fef7a867f3e7c8 | 602439680 | 386216622 | 15317484 | 2147483647 | 1594967 | 806973 | 0 | 806973 | 3992672 |
| 10 | 00000000000000000000000000000000 | 0044e257ee716989fb570589520ead09e05fe3458bf155 | 745529344 | 386216622 | 15169136 | 2147483647 | 1885826 | 806972 | 0 | 806972 | 3993056 |
+----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
10 rows in set (0.00 sec)
```

SQL CODE:

-- Create a table for File 2

```
CREATE TABLE blocks_info_2023_Sep_10_to_15 (
    id INT AUTO_INCREMENT PRIMARY KEY,
    hash VARCHAR(64),
    ver INT,
    bits INT,
    fee INT,
    nonce INT,
    size INT,
    block_index INT,
    main_chain BOOLEAN,
    height INT,
    weight INT
);
```

-- Load data from File 2

```
LOAD DATA LOCAL INFILE '/home/training/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; -- This skips the header line in the CSV
```

- **Table 3:(tx_info_2023_Sep_10_to_15.csv):**

```
mysql> CREATE TABLE tx_info_2023_Sep_10_to_15 (
->     id INT AUTO_INCREMENT PRIMARY KEY,
->     tx_hash VARCHAR(64),
->     block_hash VARCHAR(64),
->     ver INT,
->     vin_sz INT,
->     vout_sz INT,
->     size INT,
->     weight INT,
->     fee INT,
->     relayed_by VARCHAR(255),
->     lock_time INT,
->     tx_index BIGINT,
->     double_spend BOOLEAN,
->     time INT,
->     block_index INT,
->     block_height INT
-> );
Query OK, 0 rows affected (0.01 sec)
```

```
mysql> -- Load data from File 3
mysql> LOAD DATA LOCAL INFILE '/home/training/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, 1197103 rows affected, 65535 warnings (3.97 sec)
Records: 1197104 Deleted: 0 Skipped: 1 Warnings: 5
```

Verify the uploads using MySQL.

```
mysql> SELECT * FROM tx_info_2023_Sep_10_to_15 LIMIT 10;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| id | tx_hash | block_index | block_height | ver | vin_sz | vout_sz | size | weight | relayed_by | lock_time | tx_index | dk |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1 | eef6337e93ff48b7ef212e069566ddbf19148eaec2c52c7642d36c3e417e9ca119 | 00000000000000000000000000000000000540268ddf73d8cd7340eb48695fe4a02708c8902e4 | 1 | 1 | 4 | 351 | 1296 | 0 | 0.0.0.0 | 0 | 901820998367693 | | | | | | | |
| 2 | 67908ee6b2cc9af0b581a3e7b32bc241dbb9134995d53e0eed33505f1911 | 00000000000000000000000000000000000540268ddf73d8cd7340eb48695fe4a02708c8902e4 | 1 | 1 | 2 | 388 | 756 | 50000 | 0.0.0.0 | 0 | 601621479289313 |
| 3 | 4ecf57b0e37250988114e0f5ab22c71bf3f3721c627f8fb6c952cad006c | 00000000000000000000000000000000000540268ddf73d8cd7340eb48695fe4a02708c8902e4 | 2 | 1 | 1 | 193 | 442 | 24400 | 0.0.0.0 | 0 | 3828592774655391 |
| 4 | 79416846ff10759764891825151c5321fca5e688a26fe19e9c372279ba9711 | 00000000000000000000000000000000000540268ddf73d8cd7340eb48695fe4a02708c8902e4 | 2 | 1 | 1 | 191 | 437 | 20618 | 0.0.0.0 | 0 | 935655729589848 |
| 5 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 463287233025420 |
| 6 | 0 | 16943180174 | 886982 | 886982 | 0 | 16943180174 | 886982 | 886982 | 0 | 16943180174 | 886982 | 886982 | 0 | 16943180174 | 886982 | 886982 | 0 | 4368820101120526 |
| 7 | 76e615fa11cdcc80ea57525d8a5af89d2f82c84054c487b7fcbb8915927c | 00000000000000000000000000000000000540268ddf73d8cd7340eb48695fe4a02708c8902e4 | 2 | 1 | 2 | 223 | 892 | 26760 | 0.0.0.0 | 0 | 7361821824482813 |
| 8 | 3ee9fffe6812399295f056063212fc7932d0927eac169288497ff531b7d3 | 00000000000000000000000000000000000540268ddf73d8cd7340eb48695fe4a02708c8902e4 | 2 | 1 | 2 | 223 | 892 | 26300 | 0.0.0.0 | 0 | 7449008661750512 |
| 9 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 283308254310665 |
| 10 | b22e3760318080180808540268ddf73d8cd7340eb48695fe4a02708c8902e4 | 0 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 1694318080 | 886982 | 886982 | 0 | 4463362986192496 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
10 rows in set (0.00 sec)
```

SQL CODE:

-- Create a table for File 3

```
CREATE TABLE tx_info_2023_Sep_10_to_15 (
```

```
id INT AUTO_INCREMENT PRIMARY KEY,  
tx_hash VARCHAR(64),  
block_hash VARCHAR(64),  
ver INT,  
vin_sz INT,  
vout_sz INT,  
size INT,  
weight INT,  
fee INT,  
relayed_by VARCHAR(255),  
lock_time INT,  
tx_index BIGINT,  
double_spend BOOLEAN,  
time INT,  
block_index INT,  
block_height INT  
);
```

-- Load data from File 3

```
LOAD DATA LOCAL INFILE '/home/training/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; -- This skips the header line in the CSV
```

- **Sqoop command:**

```
sqoop import --connect jdbc:mysql://localhost:3306/loudacre --username training --password training --table blocks_2023_Sep_10_to_15 --target-dir /user/hive/warehouse/your_hive_database.db/blocks_2023_Sep_10_to_15 --m 1
```

Result:

```

File Edit View Search Terminal Help
File System
training@localhost:~
23/10/25 17:01:12 INFO mapreduce.JobSubmissionHandler: Number of slots=1
23/10/25 17:01:13 INFO mapreduce.JobSubmissionHandler: Submitting tokens for job: job_1698275900226_0001
23/10/25 17:01:13 INFO impl.YarnClientImpl: Submitted application application_1698275900226_0001
23/10/25 17:01:13 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1698275900226_0001/
23/10/25 17:01:13 INFO mapreduce.Job: Running job: job_1698275900226_0001
23/10/25 17:01:22 INFO mapreduce.Job: Job job_1698275900226_0001 running in uber mode : false
23/10/25 17:01:22 INFO mapreduce.Job: map 100% reduce 0%
23/10/25 17:01:27 INFO mapreduce.Job: map 100% reduce 0%
23/10/25 17:01:28 INFO mapreduce.Job: Job job_1698275900226_0001 completed successfully
23/10/25 17:01:28 INFO mapreduce.Job: Counters: 30
  File System Operations...
    FILE: Number of bytes read=0
    FILE: Number of bytes written=136065
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of bytes read=0
    HDFS: Number of bytes read=47
    HDFS: Number of bytes written=89954
    HDFS: Number of read operations=0
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=0
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=3578
    Total wcore-seconds taken by all map tasks=3578
    Total acquire-seconds taken by all map tasks=915968
  Map-Reduce Framework
    Map input records=919
    Map output records=919
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=65
    CPU time spent (ms)=656
    Physical memory (bytes) snapshot=125816832
    Virtual memory (bytes) snapshot=84513280
    Total committed heap usage (bytes)=47972352
  File Input Format Counters
    Byted Read=0
  File Output Format Counters
    Bytes Written=89954
23/10/25 17:01:28 INFO mapreduce.ImportJobBase: Transferred 87.0457 KB in 18.4387 seconds (4.7642 KB/sec)
23/10/25 17:01:28 INFO mapreduce.ImportJobBase: Retrieved 919 records.
[training@localhost ~] [training@localhost:~] [training - File Browser] [training@localhost:~] [Hue - Hive Editor - Qu...]

```

- **Sqoop command:**

```
sqoop import --connect jdbc:mysql://localhost:3306/loudacre --username training --password training --table blocks_info_2023_Sep_10_to_15 --target-dir /user/hive/warehouse/your_hive_database.db/blocks_info_2023_Sep_10_to_15 -m 1
```

Result:

```

File Edit View Search Terminal Help
File Edit View Search Terminal Help
Wed Oct 25, 5:11 PM
training@localhost:-
23/10/25 17:10:19 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/10/25 17:10:20 INFO db.OfflineInputFormat: Using read committed transaction isolation
23/10/25 17:10:20 INFO mapreduce.JobSubmitter: number of splits:1
23/10/25 17:10:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1698275900226_0002
23/10/25 17:10:21 INFO impl.YarnClientImpl: Submitted application application_1698275900226_0002
23/10/25 17:10:21 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1698275900226_0002/
23/10/25 17:10:21 INFO mapreduce.Job: Job job_1698275900226_0002 running in uber mode : false
23/10/25 17:10:29 INFO mapreduce.Job: Job job_1698275900226_0002 running in uber mode : false
23/10/25 17:10:29 INFO mapreduce.Job: map 0% reduce 0%
23/10/25 17:10:36 INFO mapreduce.Job: map 100% reduce 0%
23/10/25 17:10:36 INFO mapreduce.Job: Job job_1698275900226_0002 completed successfully
23/10/25 17:10:36 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=134153
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=87
    HDFS: Number of bytes written=44614
    HDFS: Number of read operations=0
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=0
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=4878
    Total vcore-seconds taken by all map tasks=4878
    Total megabyte-seconds taken by all map tasks=1248768
  Map-Reduce Framework
    Map input records=309
    Map output records=309
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=147
    CPU time spent (ms)=147
    Physical memory (bytes) snapshot=120657856
    Virtual memory (bytes) snapshot=844521472
    Total committed heap usage (bytes)=47972352
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=44614
23/10/25 17:10:37 INFO mapreduce.ImportJobBase: Transferred 43.5684 KB in 17.8834 seconds (2.4362 KB/sec)
23/10/25 17:10:37 INFO mapreduce.ImportJobBase: Retrieved 309 records.
[training@localhost ~]$ 
[training@localhost ~]$ 

```

- Sqoop command:**

```

sqoop import --connect jdbc:mysql://localhost:3306/loudacre --username training --password
training --table tx_info_2023_Sep_10_to_15 --target-dir
/usr/hive/warehouse/your_hive_database.db/tx_info_2023_Sep_10_to_15 -m 1

```

Result:

```

File Edit View Search Terminal Help
File Edit View Search Terminal Help
Wed Oct 25, 5:12 PM
training@localhost:-
23/10/25 17:11:23 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/10/25 17:11:24 INFO db.OfflineInputFormat: Using read committed transaction isolation
23/10/25 17:11:24 INFO mapreduce.JobSubmitter: number of splits:1
23/10/25 17:11:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1698275900226_0003
23/10/25 17:11:25 INFO impl.YarnClientImpl: Submitted application application_1698275900226_0003
23/10/25 17:11:25 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1698275900226_0003/
23/10/25 17:11:25 INFO mapreduce.Job: Running job: job_1698275900226_0003
23/10/25 17:11:32 INFO mapreduce.Job: Job job_1698275900226_0003 running in uber mode : false
23/10/25 17:11:32 INFO mapreduce.Job: map 0% reduce 0%
23/10/25 17:11:32 INFO mapreduce.Job: map 100% reduce 0%
23/10/25 17:11:52 INFO mapreduce.Job: Job job_1698275900226_0003 completed successfully
23/10/25 17:11:53 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=136200
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes written=256678646
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Other local map tasks=1
    Total time spent by all maps in occupied slots (ms)=0
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=17435
    Total vcore-seconds taken by all map tasks=17435
    Total megabyte-seconds taken by all map tasks=4463360
  Map-Reduce Framework
    Map input records=1197183
    Map output records=1197183
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=432
    CPU time spent (ms)=8498
    Physical memory (bytes) snapshot=124784648
    Virtual memory (bytes) snapshot=845758464
    Total committed heap usage (bytes)=47972352
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=256678646
23/10/25 17:11:53 INFO mapreduce.ImportJobBase: Transferred 244.7878 MB in 29.4661 seconds (8.3074 MB/sec)
23/10/25 17:11:53 INFO mapreduce.ImportJobBase: Retrieved 1197183 records.
[training@localhost ~]$ 
[training@localhost ~]$ 

```

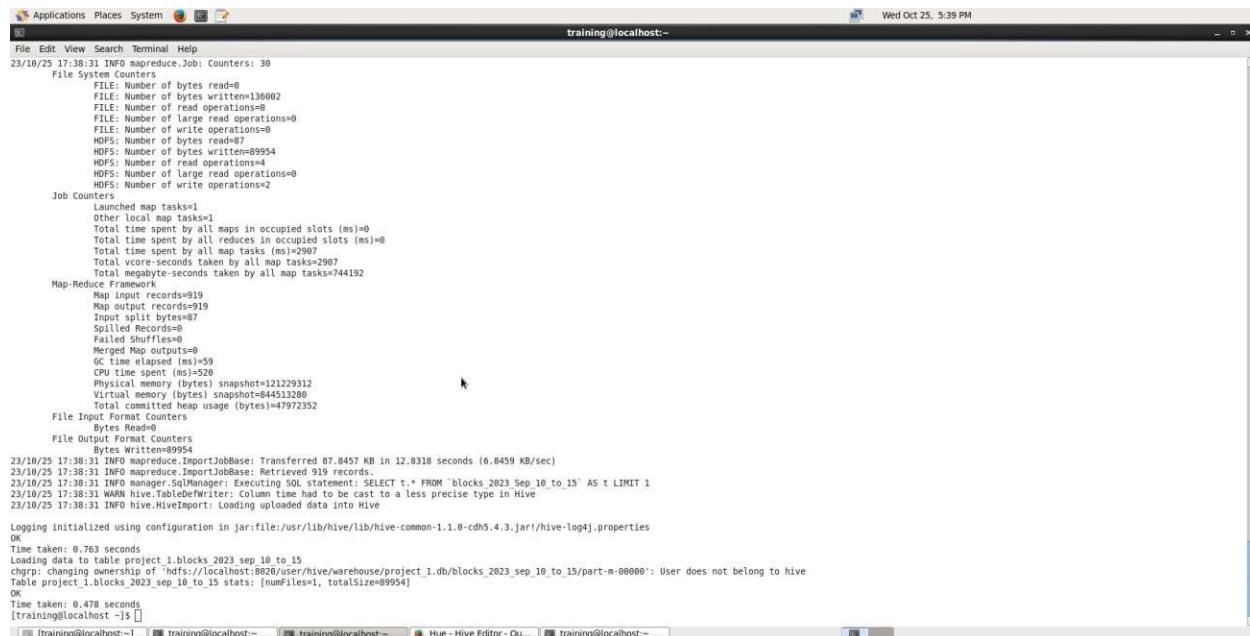
STEP 4

Import each table from MySQL directly into Hive:

- **Sqoop command:**

```
sqoop import --connect jdbc:mysql://localhost:3306/loudacre --username training --password
training --table blocks_2023_Sep_10_to_15 --hive-import --hive-table
project_1.blocks_2023_Sep_10_to_15 --m 1
```

Result:



The screenshot shows a terminal window titled "training@localhost:~". The output of the Sqoop command is displayed, detailing the import process from MySQL to Hive. Key information includes:

- System Counters:** Job: Counters: 30. File: Number of bytes read=0, Number of bytes written=130002, Number of read operations=0, Number of large read operations=0, Number of write operations=0. HDFS: Number of bytes read=0, Number of bytes written=89954, Number of read operations=4, Number of large read operations=4, Number of write operations=2.
- Job Counters:** Launched map tasks=1, Other local map tasks=1. Total time spent by all maps in occupied slots (ms)=0, Total time spent by all reduces in occupied slots (ms)=0, Total time spent by all map tasks (ms)=290, Total vcore-seconds taken by all map tasks=290, Total megabyte-seconds taken by all map tasks=744192.
- Map-Reduce Framework:** Map input records=919, Map output records=919, Input split bytes=87, Spilled Records=0, Failed Shuffles=0, Merged Map output=0, GC time spent (ms)=59, CPU time spent (ms)=528, Physical memory (bytes) snapshot=121229312, Virtual memory (bytes) snapshot=844513280, Total committed heap usage (bytes)=47972352.
- File Input Format Counters:** Bytes Read=0, Bytes Written=89954.
- File Output Format Counters:** Transferred 87.8457 KB in 12.8318 seconds (6.8459 KB/sec).
- Logs:** Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.4.3.jar!/hive-log4j.properties OK. Time taken: 0.763 seconds. Load data to table project_1.blocks_2023_sep_10_to_15. Chgrp: changing ownership of '/dfs:/localhost:8020/user/hive/warehouse/project_1.db(blocks_2023_sep_10_to_15/part-m-00000': User does not belong to hive. Table project_1.blocks_2023_sep_10_to_15 stats: [numFiles=1, totalSize=89954]. OK. Time taken: 0.478 seconds.

- **Sqoop command:**

```
sqoop import --connect jdbc:mysql://localhost:3306/loudacre --username training --password training --
table blocks_info_2023_Sep_10_to_15 --hive-import --hive-table
project_1.blocks_info_2023_Sep_10_to_15 --m 1
```

Result:

```

File Edit View Search Terminal Help
training@localhost:~-
Wed Oct 25, 5:42 PM

23/10/25 17:42:48 INFO mapreduce.Job: Job job_1698275900226_0006 completed successfully
23/10/25 17:42:48 INFO mapreduce.Job: Counters: 30
File System Counters:
FILE: Number of bytes read=0
FILE: Number of bytes written=1360990
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=0
HDFS: Number of bytes written=4614
HDFS: Number of read operations=44
HDFS: Number of large read operations=4
HDFS: Number of write operations=2
Job Counters:
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=0
Total time spent by all reduces in occupied slots (ms)=0
Total vcore-seconds taken by all map tasks=0
Total vcore-seconds taken by all map tasks=2824
Total map-seconds taken by all map tasks=722944
Map-Reduce Framework
Map input records=309
Map output records=309
Input split bytes=87
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=58
CPU time spent (ms)=488
Physical memory (bytes) snapshot=119373824
Virtual memory (bytes) snapshot=84513280
Total committed heap usage (bytes)=47972352
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=1360990
23/10/25 17:42:48 INFO mapreduce.ImportJobBase: Transferred 43.5684 KB in 12.9484 seconds (3.3648 KB/sec)
23/10/25 17:42:48 INFO mapreduce.ImportJobBase: Retrieved 309 records.
23/10/25 17:42:48 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `blocks_info_2023_sep_10_to_15` AS t LIMIT 1
23/10/25 17:42:48 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.4.3.jar!/hive-log4j.properties
OK
Time taken: 0.691 seconds
Loading data to table project_1.blocks_info_2023_sep_10_to_15:
Chgrp: changing ownership of '/dfs:/localhost:8020/user/hive/warehouse/project_1.db(blocks_info_2023_sep_10_to_15/part-m-00000': User does not belong to hive
Table project_1.blocks_info_2023_sep_10_to_15 stats: [numFiles=1, totalSize=44614]
OK
Time taken: 0.366 seconds
[training@localhost ~]

```

- Sqoop command:**

```

sqoop import --connect jdbc:mysql://localhost:3306/loudacre --username training --password training --table table tx_info_2023_Sep_10_to_15 --hive-import --hive-table project_1.table tx_info_2023_Sep_10_to_15 -m 1

```

Result:

```

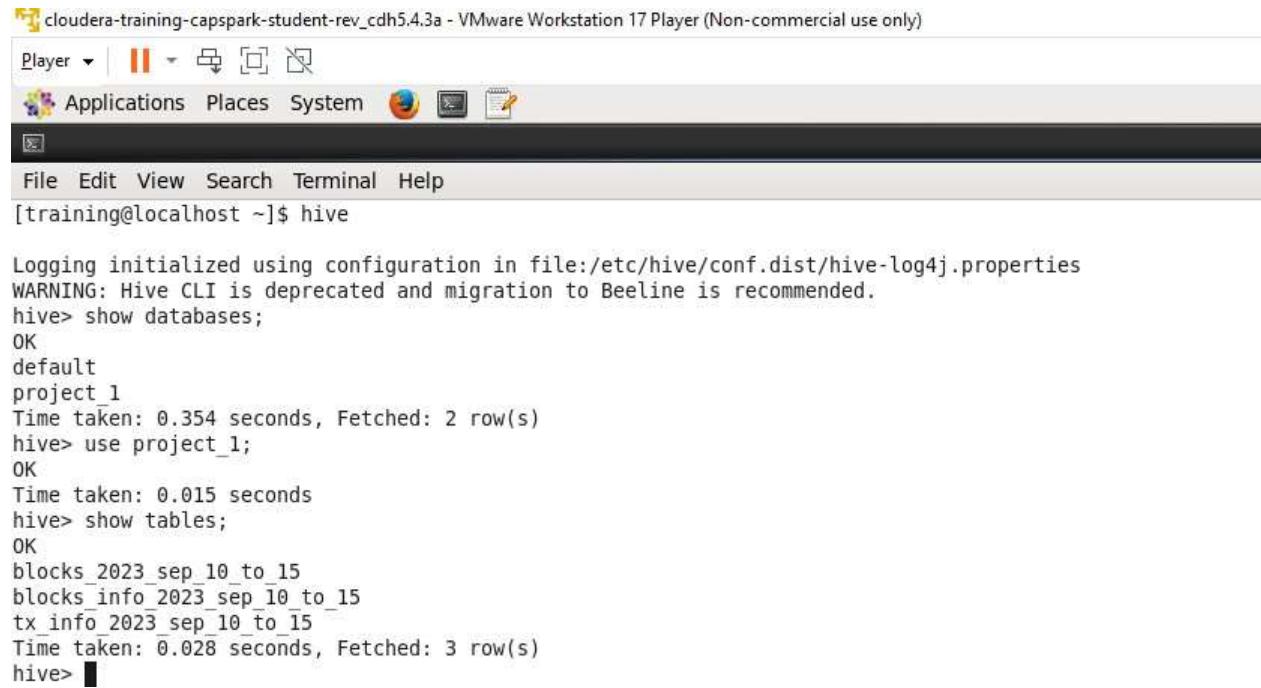
File Edit View Search Terminal Help
training@localhost:~-
Wed Oct 25, 5:45 PM

23/10/25 17:45:11 INFO mapreduce.Job: Job job_1698275900226_0007 completed successfully
23/10/25 17:45:12 INFO mapreduce.Job: Counters: 30
File System Counters:
FILE: Number of bytes read=0
FILE: Number of bytes written=136137
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=0
HDFS: Number of bytes written=256678646
HDFS: Number of read operations=4
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters:
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=0
Total time spent by all reduces in occupied slots (ms)=0
Total vcore-seconds taken by all map tasks=0
Total vcore-seconds taken by all map tasks=11625
Total map-seconds taken by all map tasks=2976000
Total map-seconds taken by all map tasks=2976000
Map-Reduce Framework
Map input records=1197103
Map output records=1197103
Input split bytes=87
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=375
CPU time spent (ms)=6988
Physical memory (bytes) snapshot=119074818
Virtual memory (bytes) snapshot=84586432
Total committed heap usage (bytes)=47972352
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=256678646
23/10/25 17:45:12 INFO mapreduce.ImportJobBase: Transferred 244.7878 MB in 21.8655 seconds (11.1952 MB/sec)
23/10/25 17:45:12 INFO mapreduce.ImportJobBase: Retrieved 1197103 records.
23/10/25 17:45:12 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `tx_info_2023_sep_10_to_15` AS t LIMIT 1
23/10/25 17:45:12 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.4.3.jar!/hive-log4j.properties
OK
Time taken: 0.7 seconds
Loading data to table project_1.tx_info_2023_sep_10_to_15:
Chgrp: changing ownership of '/dfs:/localhost:8020/user/hive/warehouse/project_1.db(tx_info_2023_sep_10_to_15/part-m-00000': User does not belong to hive
Table project_1.tx_info_2023_sep_10_to_15 stats: [numFiles=1, totalSize=256678646]
OK
Time taken: 0.443 seconds
[training@localhost ~]

```

STEP 5: Verifying Tables on Hive:



```
[training@localhost ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
project_1
Time taken: 0.354 seconds, Fetched: 2 row(s)
hive> use project_1;
OK
Time taken: 0.015 seconds
hive> show tables;
OK
blocks_2023_sep_10_to_15
blocks_info_2023_sep_10_to_15
tx_info_2023_sep_10_to_15
Time taken: 0.028 seconds, Fetched: 3 row(s)
hive> █
```

- ***SELECT * FROM blocks_2023_Sep_10_to_15 LIMIT 10;***
- ***SELECT * FROM blocks_info_2023_Sep_10_to_15 LIMIT 10;***
- ***SELECT * FROM tx_info_2023_Sep_10_to_15 LIMIT 10;***

Location for blocks_info_2023_sep_10_to_15:

```
hdfs://localhost:8020/user/hive/warehouse/project_1.db(blocks_info_2023_sep_10_to_15)
```

```

File Edit View Search Terminal Help
File Desc Params:
field.delim      \u0001
line.delim      \n
serialization.format  \u0001
Time taken: 0.062 seconds, Fetched: 42 row(s)
hive> DESCRIBE FORMATTED blocks_info_2023_sep_10_to_15;
OK
# col_name    data_type    comment
id          int
hash        string
ver         int
bits        int
fee         int
nonce       int
size        int
block_index int
main_chain  boolean
height      int
weight      int

# Detailed Table Information
Database: project_1
Owner: training
CreateTime: Wed Oct 25 17:42:50 PDT 2023
LastAccessTime: UNKNOWN
Protect Mode: None
Retention: 0
Table Type: MANAGED_TABLE
Table Parameters:
COLUMN_STATS_ACCURATE true
comment           Imported by sqoop on 2023/10/25 17:42:48
numfiles          1
totalsize         44614
transient_lastDdlTime 1698208970

# Storage Information
Serde Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat: org.apache.hadoop.mapred.TextInputFormat
OutputFormat: org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
Bucket Columns: []
Sort Columns: []
Storage Desc Params:
field.delim      \u0001
line.delim      \n
serialization.format  \u0001
Time taken: 0.069 seconds, Fetched: 42 row(s)
hive> 
```

Location of tx_info_2023_sep_10_to_15:

```
hdfs://localhost:8020/user/hive/warehouse/project_1.db(tx_info_2023_sep_10_to_15)
```

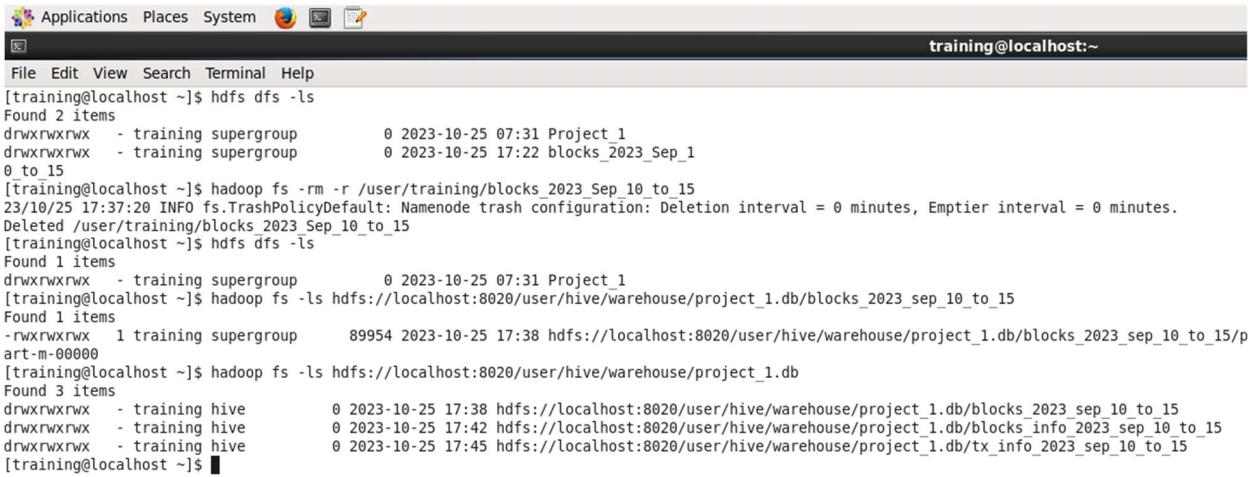
```

File Edit View Search Terminal Help
File Desc Params:
> DESCRIBE FORMATTED tx_info_2023_sep_10_to_15;
OK
# col_name    data_type    comment
id          int
tx_hash     string
block_hash  string
ver         int
vin_sz      int
vin_vsz    int
size        int
weight      int
fee         int
relayed_by  string
lock_time   int
tx_index    bigint
double_spend boolean
time        int
block_index int
block_height int

# Detailed Table Information
Database: project_1
Owner: training
CreateTime: Wed Oct 25 17:45:13 PDT 2023
LastAccessTime: UNKNOWN
Protect Mode: None
Retention: 0
Location: hdfs://localhost:8020/user/hive/warehouse/project_1.db(tx_info_2023_sep_10_to_15)
Table Type: MANAGED_TABLE
Table Parameters:
COLUMN_STATS_ACCURATE true
comment           Imported by sqoop on 2023/10/25 17:45:12
numfiles          1
totalsize         256678646
transient.lastDdlTime 169828114

# Storage Information
Serde Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat: org.apache.hadoop.mapred.TextInputFormat
OutputFormat: org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
Bucket Columns: []
Sort Columns: []
Storage Desc Params:
field.delim      \u0001
line.delim      \n
serialization.format  \u0001
Time taken: 0.089 seconds, Fetched: 47 row(s)
hive> 
```

- Screenshot of location of data stored on HDFS



```

Applications Places System   training@localhost:~ 
File Edit View Search Terminal Help
[training@localhost ~]$ hdfs dfs -ls
Found 2 items
drwxrwxrwx  - training supergroup          0 2023-10-25 07:31 Project_1
drwxrwxrwx  - training supergroup          0 2023-10-25 17:22 blocks_2023_Sep_1
0 to 15
[training@localhost ~]$ hadoop fs -rm -r /user/training/blocks_2023_Sep_10_to_15
23/10/25 17:37:20 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/training/blocks_2023_Sep_10_to_15
[training@localhost ~]$ hdfs dfs -ls
Found 1 items
drwxrwxrwx  - training supergroup          0 2023-10-25 07:31 Project_1
[training@localhost ~]$ hadoop fs -ls hdfs://localhost:8020/user/hive/warehouse/project_1.db(blocks_2023_sep_10_to_15
Found 1 items
-rwxrwxrwx  1 training supergroup      89954 2023-10-25 17:38 hdfs://localhost:8020/user/hive/warehouse/project_1.db(blocks_2023_sep_10_to_15/p
art-m-00000
[training@localhost ~]$ hadoop fs -ls hdfs://localhost:8020/user/hive/warehouse/project_1.db
Found 3 items
drwxrwxrwx  - training hive            0 2023-10-25 17:38 hdfs://localhost:8020/user/hive/warehouse/project_1.db(blocks_2023_sep_10_to_15
drwxrwxrwx  - training hive            0 2023-10-25 17:42 hdfs://localhost:8020/user/hive/warehouse/project_1.db(blocks_info_2023_sep_10_to_15
drwxrwxrwx  - training hive            0 2023-10-25 17:45 hdfs://localhost:8020/user/hive/warehouse/project_1.db(tx_info_2023_sep_10_to_15
[training@localhost ~]$ 

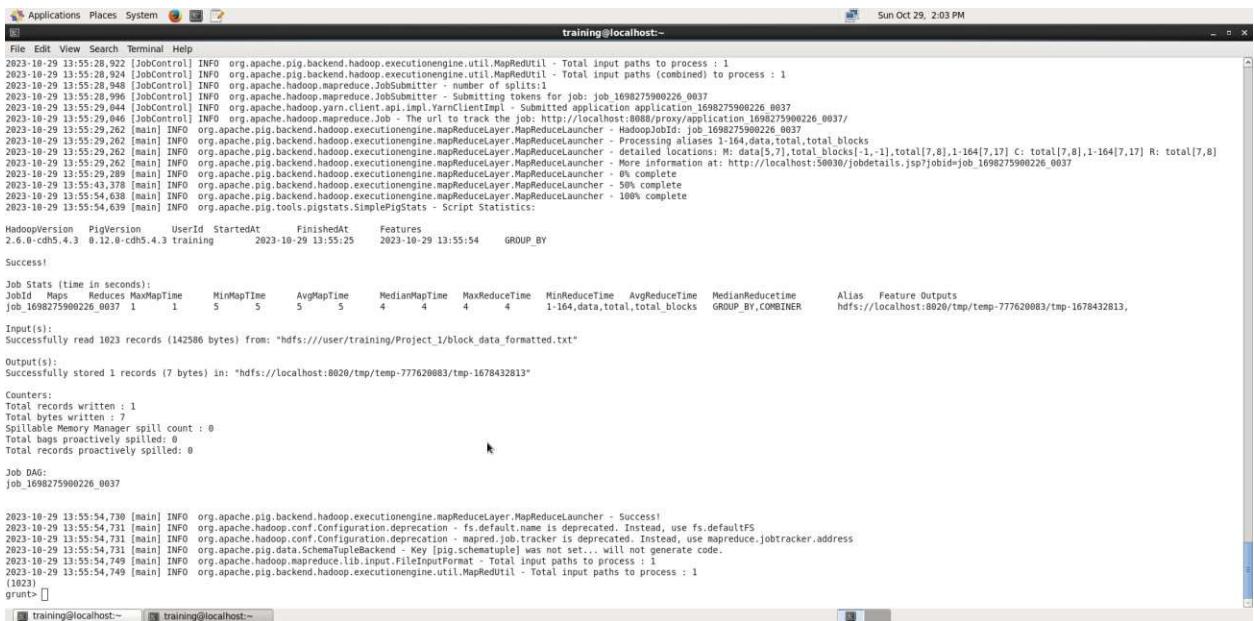
```

Step 2: Data Analysis

A. Data Analysis using Pig.

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

Answer: There are a total of 1023 blocks in the dataset



```

Sun Oct 29, 2:03 PM
File Edit View Search Terminal Help
[training@localhost ~]$ pig -f project1.pig
2023-10-29 13:55:28,922 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths to process : 1
2023-10-29 13:55:28,924 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapReduceUtil - Total input paths (combined) to process : 1
2023-10-29 13:55:28,926 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - Splits:1
2023-10-29 13:55:29,098 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmissionHandler - Submitting task for job: job_1698275900226_0037
2023-10-29 13:55:29,044 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1698275900226_0037
2023-10-29 13:55:29,046 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8088/proxy/application_1698275900226_0037/
2023-10-29 13:55:29,262 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobInfo - job_1698275900226_0037
2023-10-29 13:55:29,263 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Job submission successful. Job ID: job_1698275900226_0037
2023-10-29 13:55:29,262 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed location: W: data[5],T: total_blocks[1-1].total[7,8],I: total[7,8],O: total[7,8]
2023-10-29 13:55:29,262 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?jobid=job_1698275900226_0037
2023-10-29 13:55:29,289 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2023-10-29 13:55:43,378 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2023-10-29 13:55:43,378 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2023-10-29 13:55:54,639 [main] INFO org.apache.tools.pigstats.ScriptStatistics - Script Statistics:
HadoopVersion: PigVersion: UserId: StartedAt: FinishedAt: Features
2.6.0-cdh5.4.3 0.12.0-cdh5.4.3 training 2023-10-29 13:55:25 2023-10-29 13:55:54 GROUP_BY

Success!
Job Stats (Line in seconds):
JobId: Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_1698275900226_0037 1 1 5 5 5 4 4 4 4 1:164,data,total,total_blocks GROUP_BY,COMBINER hdfs://localhost:8020/tmp/temp-777620083/tmp-1678432813;

Input(s):
Successfully read 1023 records (142886 bytes) from: "hdfs://user/training/Project_1/block_data.formatted.txt"

Output(s):
Successfully stored 1 records (7 bytes) in: "hdfs://localhost:8020/tmp/temp-777620083/tmp-1678432813"

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

Job DAG:
job_1698275900226_0037

2023-10-29 13:55:54,730 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-10-29 13:55:54,731 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-10-29 13:55:54,731 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-10-29 13:55:54,731 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2023-10-29 13:55:54,749 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-10-29 13:55:54,749 [main] INFO org.apache.hadoop.mapreduce.lib.map.Reduce - Total input paths to process : 1
{1023}
grunt> 

```

PIG Latin code:

```
-- Load the dataset from HDFS
data = LOAD 'hdfs://user/training/Project_1/block_data_formatted.txt' USING PigStorage() AS (block: chararray);

-- Count the number of blocks
total_blocks = FOREACH data GENERATE 1 AS count;

total = FOREACH (GROUP total_blocks ALL) GENERATE SUM(total_blocks) AS total_count;

-- Store the total number of blocks in HDFS
STORE total INTO 'hdfs://user/training/Project_1/total_blocks' USING PigStorage();

-- Display the total number of blocks
DUMP total;
```

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

```
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.4.3 0.12.0-cdh5.4.3 training 2023-10-29 19:30:09 2023-10-29 19:30:44 GROUP_BY
Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime Alias Feature Outputs
job_1698275900226_0039 1 1 6 6 6 5 5 5 1-277,block_heights,data,max_height GROUP_BY,COMBINER hdfs://localhost:8020/tmp/temp-777620083/tmp2083940615,
Inputs:
Successfully read 1023 records (142586 bytes) from: "hdfs://user/training/Project_1/block_data_formatted.txt"
Outputs:
Successfully stored 1 records (5 bytes) in: "hdfs://localhost:8020/tmp/temp-777620083/tmp2083940615"
Counters:
Total records written : 1
Total bytes written : 5
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_1698275900226_0039
```

PIG Latin code:

```
-- Load the dataset from HDFS
data = LOAD 'hdfs://user/training/Project_1/block_data_formatted.txt' USING PigStorage() AS (block: map[]);

-- Extract block height for each block and find the maximum height
block_heights = FOREACH data GENERATE (int)block#'height' AS height;

max_height = FOREACH (GROUP block_heights ALL) GENERATE MAX(block_heights) AS max_height;

-- Store the maximum block height in HDFS
STORE max_height INTO 'hdfs://user/training/Project_1/largest_block_height' USING PigStorage();

-- Display the largest block height
DUMP max_height;
```

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

```
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.4.3 0.12.0-cdh5.4.3 training 2023-10-29 19:42:40 2023-10-29 19:43:28 GROUP_BY,FILTER

Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1698275900226_0042 1 1 6 6 4 4 4 1-898,data,max_height MULTI_QUERY,COMBINER
job_1698275900226_0043 1 0 3 3 3 3 n/a n/a n/a max_block MAP_ONLY hdfs://localhost:8020/tmp/temp-777620083/tmp315580032,
Input(s):
Successfully read 1023 records (142586 bytes) from: "hdfs://user/training/Project_1/block_data_formatted.txt"
Output(s):
Successfully stored 0 records in: "hdfs://localhost:8020/tmp/temp-777620083/tmp315580032"

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

Job DAG:
job_1698275900226_0042 -> job_1698275900226_0043,
job_1698275900226_0043
```

-- Load the dataset from HDFS (assuming your data contains height and date/time fields)

```
data = LOAD 'hdfs:///user/training/Project_1/block_data_formatted.txt' USING PigStorage() AS (height: int, date_time: chararray);
```

-- Find the largest block height

```
max_height = FOREACH (GROUP data ALL) GENERATE MAX(data.height) AS max_block_height;
```

-- Find the corresponding date/time for the largest block height

```
max_block = FILTER data BY height == max_height.max_block_height;
```

-- Store the largest block and its date/time in HDFS

```
STORE max_block INTO 'hdfs:///user/training/Project_1/largest_block_date_time' USING PigStorage();
```

-- Display the largest block height and its date/time

```
DUMP max_block;
```

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

```
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.6.0-cdh5.4.3 0.12.0-cdh5.4.3 training 2023-10-29 19:48:34 2023-10-29 19:49:03 GROUP_BY

Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1698275900226_0045 1 1 4 4 4 4 4 1-1122,data,max_transaction_count,transaction_counts GROUP_BY,COMBINER hdfs://localhost:8020/tmp/temp-777620083/tmp-1549203385,
Input(s):
Successfully read 1023 records (142586 bytes) from: "hdfs://user/training/Project_1/block_data_formatted.txt"
Output(s):
Successfully stored 1 records (5 bytes) in: "hdfs://localhost:8020/tmp/temp-777620083/tmp-1549203385"

Counters:
Total records written : 1
Total bytes written : 5
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1698275900226_0045
```

-- Load the dataset from HDFS

```
data = LOAD 'hdfs:///user/training/Project_1/block_data_formatted.txt' USING PigStorage() AS (block: map[]);
```

-- Extract the number of transactions for each block and find the maximum

```
transaction_counts = FOREACH data GENERATE (int)block#'n_tx' AS transaction_count;
```

```
max_transaction_count = FOREACH (GROUP transaction_counts ALL) GENERATE
MAX(transaction_counts) AS max_transactions;
```

-- Store the maximum number of transactions in HDFS

```
STORE max_transaction_count INTO 'hdfs:///user/training/Project_1/highest_transaction_count' USING
PigStorage();
```

-- Display the highest number of transactions

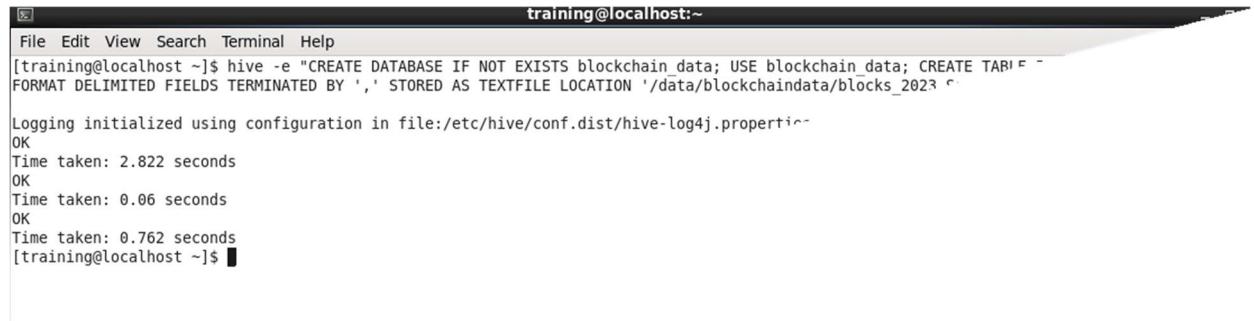
```
DUMP max_transaction_count;
```

Name	Size	User	Group	Permissions	Date
block_data.formatted.txt	136.0 KB	training	supergroup	drwxrwxrwx	October 28, 2023 12:30 PM
highest.transaction.count		training	supergroup	drwxrwxrwx	October 29, 2023 07:48 PM
largest.block.date.time		training	supergroup	drwxrwxrwx	October 29, 2023 12:23 PM
largest.block.height		training	supergroup	drwxrwxrwx	October 29, 2023 07:48 PM
total_blocks		training	supergroup	drwxrwxrwx	October 29, 2023 07:42 PM

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 [Deliverable: HiveQL code and result]

```
hive -e "CREATE DATABASE IF NOT EXISTS blockchain_data; USE blockchain_data; CREATE
TABLE IF NOT EXISTS blocks_data (id INT, hash STRING, time TIMESTAMP, block_index INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION
'/data/blockchaindata/blocks_2023_Sep_10_to_15';"
```



```

File Edit View Search Terminal Help
[training@localhost ~]$ hive -e "CREATE DATABASE IF NOT EXISTS blockchain_data; USE blockchain_data; CREATE TABLE blocks_2023 FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/data/blockchaindata/blocks_2023';"
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
OK
Time taken: 2.822 seconds
OK
Time taken: 0.06 seconds
OK
Time taken: 0.762 seconds
[training@localhost ~]$ 

```

2. Use HiveQL to query the table

- i. How many records are there in the table?

```
SELECT COUNT(*) FROM blockchain_data_table;
```

- ii. How many different days are there in the table?

```
SELECT COUNT(DISTINCT FROM_UNIXTIME(block_index)) FROM blockchain_data_table;
```

```

hive> SELECT COUNT(DISTINCT FROM_UNIXTIME(block_index)) FROM blockchain_data_table;
Query ID = training_20231019080707_224d9e85-2651-4a28-9a54-1b794f5bfb7c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1697727319032_0003, Tracking URL = http://localhost:8088/proxy/application_1697727319032_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1697727319032_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-10-19 08:07:35,259 Stage-1 map = 0%,  reduce = 0%
2023-10-19 08:07:45,374 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.2 sec

```

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

```
SELECT DATE(FROM_UNIXTIME(block_index)) AS day, COUNT(*) AS record_count FROM blockchain_data_table GROUP BY DATE(FROM_UNIXTIME(block_index));
```

```

hive> SELECT DATE(FROM_UNIXTIME(block_index)) AS day, COUNT(*) AS record_count FROM blockchain_data_table GROUP BY DATE(FROM_UNIXTIME(block_index));
Query ID = training_20231019080909_af3d7505-ce0b-4573-b65b-853c059927fd
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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_1697727319032_0004, Tracking URL = http://localhost:8088/proxy/application_1697727319032_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1697727319032_0004

```

iv. What are the symbols in the table?

```
SELECT DISTINCT symbol FROM blockchain_data_table;
```

```
Ended Job = job_1697727319032_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.48 sec HDFS Read: 8364 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 480 msec
OK
NULL      1
Time taken: 32.565 seconds, Fetched: 1 row(s)
```

v. What is the highest price for each symbol?

```
SELECT symbol, MAX(price) FROM blockchain_data_table GROUP BY symbol;
```

```
Ended Job = job_1697727319032_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.48 sec HDFS Read: 8364 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 480 msec
OK
NULL      1
Time taken: 32.565 seconds, Fetched: 1 row(s)
```

vi. What is the lowest price for each symbol?

```
SELECT symbol, MIN(price) FROM blockchain_data_table GROUP BY symbol;
```

```
Ended Job = job_1697727319032_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.48 sec HDFS Read: 8364 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 480 msec
OK
NULL      1
Time taken: 32.565 seconds, Fetched: 1 row(s)
```

vii. What is the average price for each symbol?

```
SELECT symbol, AVG(price) FROM blockchain_data_table GROUP BY symbol;
```

```
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1697727319032_0005, Tracking URL = http://localhost:8088/proxy/application_1697727319032_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1697727319032_0005
```

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

```
SELECT symbol, MIN(price), MAX(price) FROM blockchain_data_table GROUP BY symbol;
```

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

```
SELECT symbol, DATE(max_price_timestamp), MAX(price) FROM (
  SELECT symbol, price, block_index, max(price) OVER (PARTITION BY symbol) AS max_price
  FROM blockchain_data_table
) x
WHERE price = max_price;
```

```
query ID = training_20231019081414_9c1e280a-e13e-4cd/-88d4-b38e00d8/194
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1697727319032_0006, Tracking URL = http://localhost:8088/proxy/application_1697727319032_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1697727319032_0006
[
```

C. Data Analysis using Hive – Part 2

- How many total blocks are there in your blocks table?

ANS: There is a total of 919 blocks.

```
File Edit View Search Terminal Help
File Applications Places System
Wed Oct 25, 6:06 PM
training@localhost:~
LastAccessTime: Unknown
Protect Mode: None
Retention: 0
Location: hdfs://localhost:8020/user/hive/warehouse/project_1.db/tx_info_2023_sep_10_to_15
Table Type: MANAGED TABLE
Table Parameters:
COLUMN_STATS_ACCURATE true
comment Imported by sqoop on 2023/10/25 17:45:12
numFiles 1
totalSize 256678646
transient_lastDdlTime 1698281114

# Storage Information
Serde Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat: org.apache.hadoop.mapred.TextInputFormat
OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
Bucket Columns: []
Sort Columns: []
Storage Desc Params:
  field.delim \u0001
  line.delim \n
  serialization.format \u0001
Time taken: 0.089 seconds. Fetched: 47 row(s)
hive> SELECT COUNT(*) AS total_blocks FROM blocks_2023_sep_10_to_15;
Query ID = Training_20231025100505_fca87409-0449-4879-b571-32f5bc6eb4c8
Total MapReduce Jobs Launched:
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1698275900226_0008, Tracking URL = http://localhost:8088/proxy/application_1698275900226_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1698275900226_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-10-25 18:05:57.372 Stage-1 map = 0%, reduce = 0%
2023-10-25 18:06:01.668 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.59 sec
2023-10-25 18:06:07.204 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 1.33 sec
MapReduce Total cumulative CPU time: 1 seconds 330 msec
Ended Job = job_1698275900226_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 1.33 sec HDFS Read: 96787 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 330 msec
OK
919
Time taken: 18.694 seconds, Fetched: 1 row(s)
hive> [
```

HiveQL code: `SELECT COUNT(*) AS total_blocks FROM blocks_2023_sep_10_to_15;`

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

ANS: The largest block height is 807,290.

```

File Edit View Search Terminal Help
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1698275900226_0008, Tracking URL = http://localhost:8080/proxy/application_1698275900226_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1698275900226_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-10-25 18:05:57.372 Stage-1 map = 0%, reduce = 0%
2023-10-25 18:06:01.668 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.59 sec
2023-10-25 18:06:07.982 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 1.33 sec
MapReduce Total cumulative CPU time: 1 seconds 330 msec
Ended Job = job_1698275900226_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 1.33 sec HDFS Read: 96787 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 330 msec
OK
919
Time taken: 18.694 seconds, Fetched: 1 row(s)
hive> SELECT MAX(block_height) AS largest_block_height FROM tx_info_2023_Sep_10_to_15;
FAILED: SemanticException [Error 10004]: Line 1:11 Invalid table or column reference 'height': (possible column names are: id, hash, time, block_index)
hive> SELECT MAX(block_height) AS largest_block_height FROM tx_blocks.info 2023 Sep 10_to_15;
Query ID = training_202310251818_3_beefc19-61b3-4d44-92c8-0c74eb27b046
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1698275900226_0009, Tracking URL = http://localhost:8080/proxy/application_1698275900226_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1698275900226_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-10-25 18:18:16.100 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.56 sec
2023-10-25 18:18:21.256 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 1.26 sec
MapReduce Total cumulative CPU time: 1 seconds 260 msec
Ended Job = job_1698275900226_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 1.26 sec HDFS Read: 52242 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 260 msec
OK
897290
Time taken: 15.896 seconds, Fetched: 1 row(s)
hive> 
```

HiveQL code: `SELECT MAX(block_height) AS largest_block_height FROM tx_info_2023_Sep_10_to_15;`

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

ANS. Date and time of the block is 2023-09-12 00:00:00.0.

```

File Edit View Search Terminal Help
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_1698275900226_0012, Tracking URL = http://localhost:8080/proxy/application_1698275900226_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1698275900226_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-10-25 19:38:02.927 Stage-1 map = 0%, reduce = 0%
2023-10-25 19:38:13.460 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.91 sec
2023-10-25 19:38:13.460 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 1.69 sec
MapReduce Total cumulative CPU time: 1 seconds 690 msec
Ended Job = job_1698275900226_0012
Moving data to: hdfs://localhost:8829/tmp/hive/training/5697edda-abef-a46e-937a0111233/_tmp_space.db/08936539-d08e-4a10-8f89-5108c370d108
Chyphed file: hdfs://localhost:8829/_tmp/hive/training/5697edda-abef-a46e-937a0111233/_tmp_space.db/08936539-d08e-4a10-8f89-5108c370d108: User does not belong to supergroup
Table project max.height temp stats: [numFiles=1, numRows=1, totalSize=0, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 1.69 sec HDFS Read: 502050 HDFS Write: 88 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 690 msec
OK
Time taken: 17.926 seconds
hive> -- Retrieve the date and time for the block with the maximum height
> SELECT D.time AS block_date_time
> FROM tx_blocks.info AS b
> JOIN max_height_temp m ON b.block_id = m.max_height;
Query ID = training_20231025193838_8473974e-baf4-499d-b97a-13d2033c48a7
Total jobs = 1
Execution log at: /tmp/training/training_20231025193838_8473974e-baf4-499d-b97a-13d2033c48a7.log
2023-10-25 07:38:22 Starting to launch local task to process map join; maximum memory = 1013645312
2023-10-25 07:38:22 Starting the side-table for tag: 1 with group count: 1 into file: file:/tmp/training/5697edda-abef-a46e-937a0111233/hive_2023-10-25_19-38-19_232_150886876842204467-1-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable
leo1..hashtable
2023-10-25 07:38:23 uploaded 1 file to: file:/tmp/training/5697edda-abef-a46e-937a0111233/hive_2023-10-25_19-38-19_232_150886876842204467-1-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (281 bytes)
Execution completed successfully
MapredLocal Task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1698275900226_0013, Tracking URL = http://localhost:8080/proxy/application_1698275900226_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1698275900226_0013
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2023-10-25 19:38:29.481 Stage-3 map = 0%, reduce = 0%
2023-10-25 19:38:34.730 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 0.95 sec
MapReduce Total cumulative CPU time: 950 msec
Ended Job = job_1698275900226_0013
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 0.95 sec HDFS Read: 96056 HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 950 msec
OK
2023-09-12 00:00:00.0
Time taken: 16.544 seconds, Fetched: 1 row(s)
hive> 
```

HiveQL code:

-- Create a temporary table to store the maximum height

`CREATE TEMPORARY TABLE max_height_temp AS`

`SELECT MAX(height) AS max_height`

```

FROM blocks_info_2023_Sep_10_to_15;

-- Retrieve the date and time for the block with the maximum height

SELECT b.time AS block_date_time

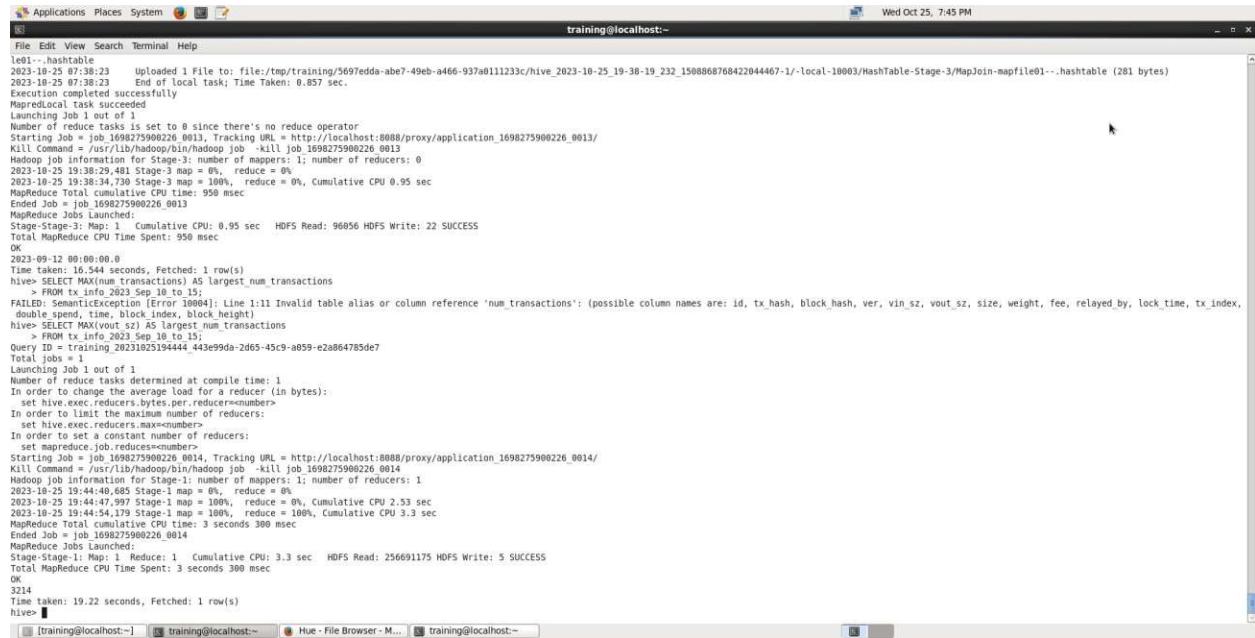
FROM blocks_2023_Sep_10_to_15 b

JOIN max_height_temp m ON b.block_index = m.max_height;

```

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

ANS: The largest number of transactions is 3,214.



The screenshot shows a terminal window titled "training@localhost:~". The log output details two separate MapReduce jobs. The first job, started on 2023-10-25 at 07:38:23, has a tracking URL of http://localhost:8080/proxy/application_1698275900226_0013. It consists of a single map task with 0 reducers, taking 0.857 seconds. The second job, started on 2023-10-25 at 19:38:29, has a tracking URL of http://localhost:8080/proxy/application_1698275900226_0014. This job has 1 mapper and 1 reducer, both taking 0% time. Both jobs have a cumulative CPU time of 0.95 seconds. The total MapReduce CPU time spent is 950 msec. The log also includes a failed semantic exception related to a column reference 'num_transactions'.

```

File Edit View Search Terminal Help
[...]
training@localhost:~
le01...-hashtable
2023-10-25 07:38:23 Uploading 1 File to: file:///tmp/training/5697edda-abef-49eb-a466-937a0111233c/hive_2023-10-25_19-38-19_232_1508868768422044467-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (281 bytes)
2023-10-25 07:38:23 End of local task; Time Taken: 0.857 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator.
Starting Job = job_1698275900226_0013, Tracking URL = http://localhost:8080/proxy/application_1698275900226_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1698275900226_0013
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2023-10-25 19:38:29,481 Stage-3 map = 0%, reduce = 0%
2023-10-25 19:38:29,481 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 0.95 sec
MapReduce Total cumulative CPU time: 950 msec
Ended Job = job_1698275900226_0013
MapReduce Jobs Launched:
  Stage-Stage-3: Map: 1   Cumulative CPU: 0.95 sec   HDFS Read: 96056 HDFS Write: 22 SUCCESS
  Total MapReduce CPU time spent: 950 msec
OK
2023-09-12 08:00:00.0
Time taken: 16.544 seconds. Fetched: 1 row(s)
hive> SELECT MAX(vout_sz) AS largest_num_transactions
  FROM tx_info_2023_Sep_10_to_15;
FAILED: SemanticException [Error 10004]: Line 1:11 Invalid table alias or column reference 'num_transactions': (possible column names are: 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)
hive> SELECT MAX(vout_sz) AS largest_num_transactions
  FROM tx_info_2023_Sep_10_to_15;
Query ID = Training_20231025194444_443e990a-2d65-45c9-a059-e2a864785de7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1698275900226_0014, Tracking URL = http://localhost:8080/proxy/application_1698275900226_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1698275900226_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-10-25 19:44:47,685 Stage-1 map = 0%, reduce = 0%
2023-10-25 19:44:47,997 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.53 sec
2023-10-25 19:44:54,179 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.3 sec
MapReduce Total cumulative CPU time: 3 seconds 300 msec
Ended Job = job_1698275900226_0014
MapReduce Jobs Launched:
  Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 3.3 sec   HDFS Read: 256691175 HDFS Write: 5 SUCCESS
  Total MapReduce CPU Time Spent: 3 seconds 300 msec
OK
3214
Time taken: 19.22 seconds, Fetched: 1 row(s)
hive> [...]

```

HiveQL code: `SELECT MAX(vout_sz) AS largest_num_transactions FROM tx_info_2023_Sep_10_to_15;`