

Big Data Project

BUAN 6346.504

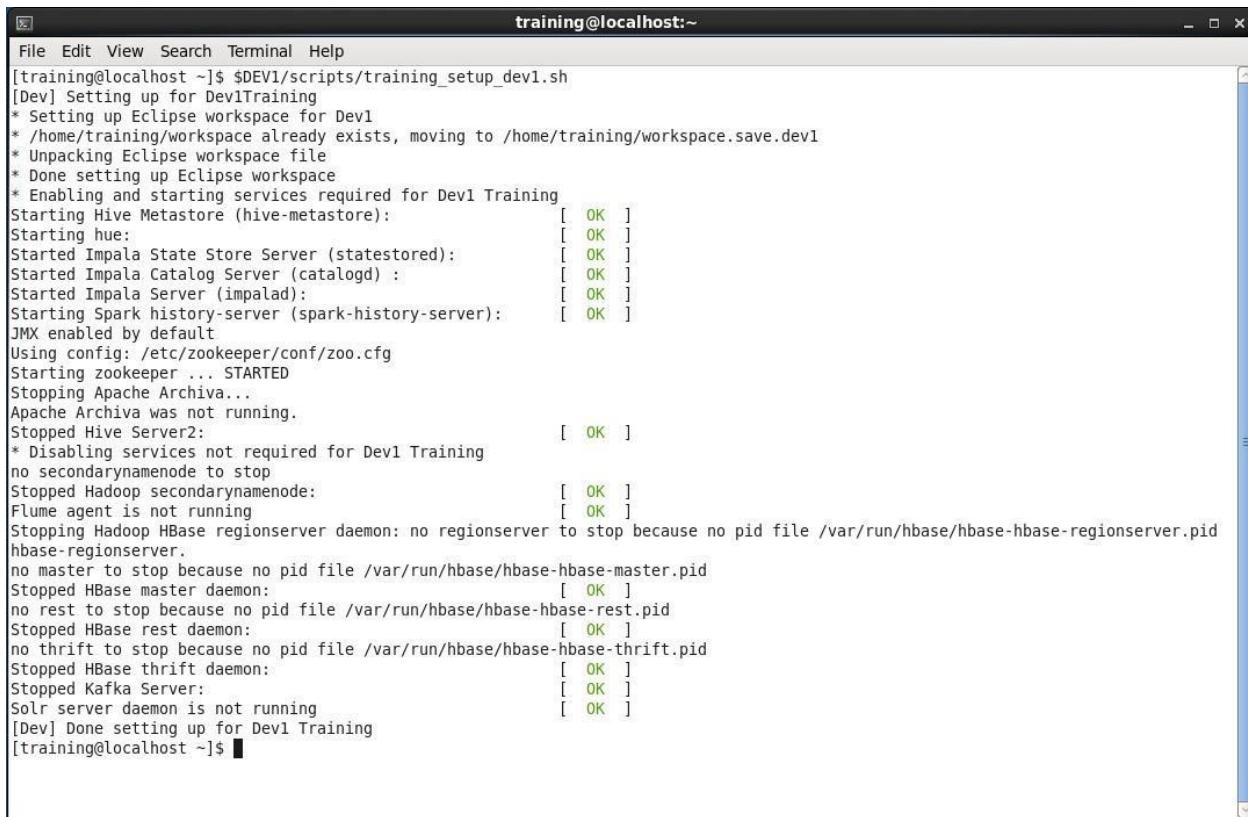
Group 14

Group Members:

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

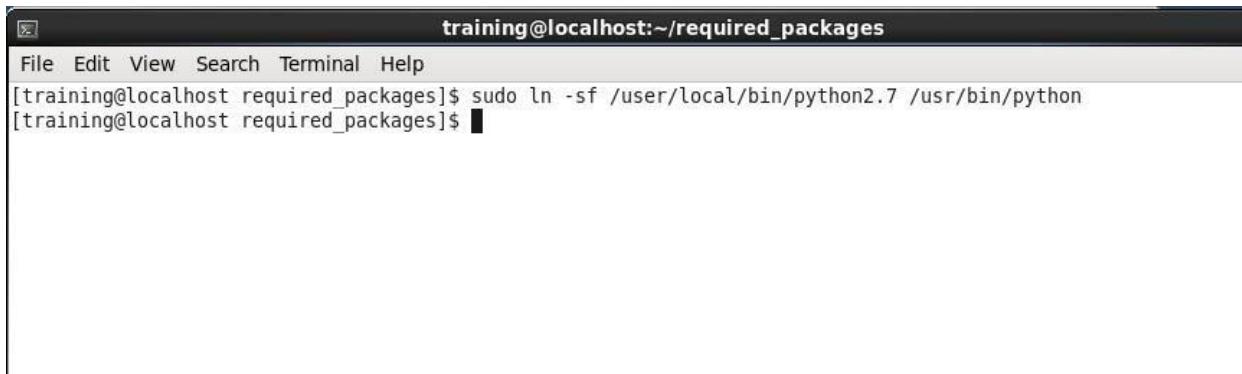
Step 0: VM Setup

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



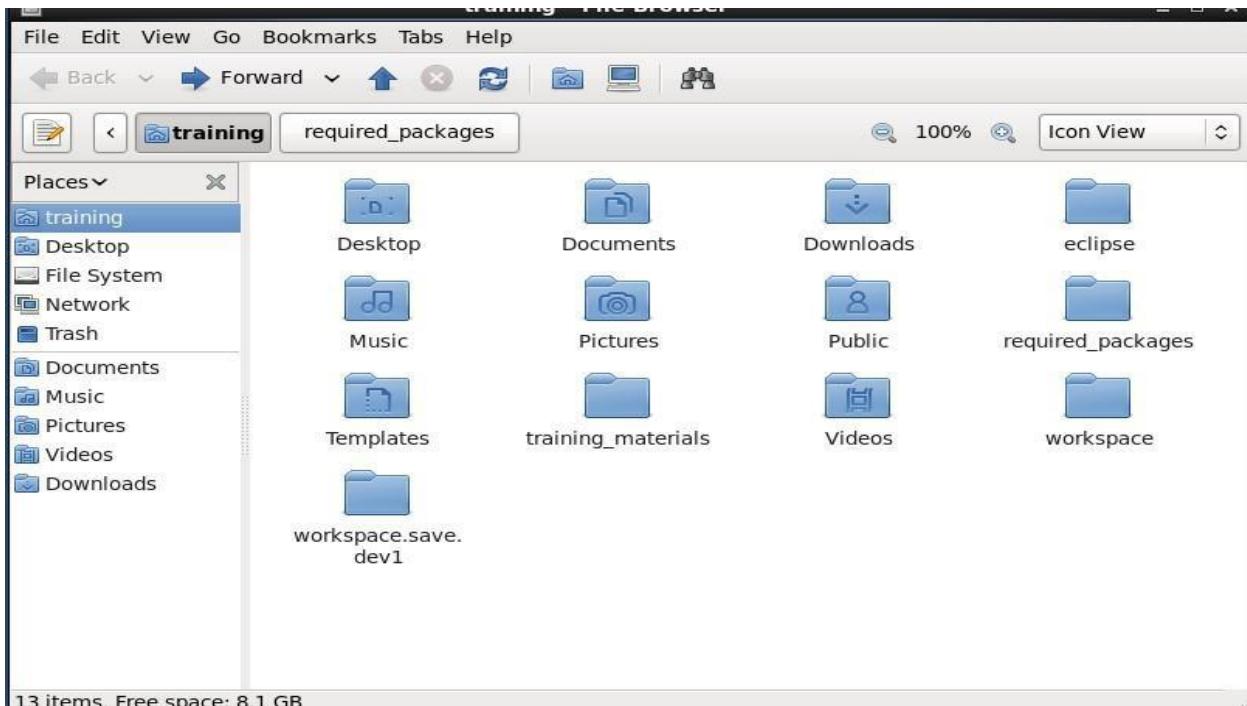
```
File Edit View Search Terminal Help
[training@localhost ~]$ $DEV1/scripts/training_setup_dev1.sh
[Dev] Setting up for DevlTraining
* Setting up Eclipse workspace for Devl
* /home/training/workspace already exists, moving to /home/training/workspace.save.devl
* Unpacking Eclipse workspace file
* Done setting up Eclipse workspace
* Enabling and starting services required for Devl Training
Starting Hive Metastore (hive-metastore): [ OK ]
Starting hue: [ OK ]
Started Impala State Store Server (statestored): [ OK ]
Started Impala Catalog Server (catalogd) : [ OK ]
Started Impala Server (impalad): [ OK ]
Starting Spark history-server (spark-history-server): [ OK ]
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... STARTED
Stopping Apache Archiva...
Apache Archiva was not running.
Stopped Hive Server2: [ OK ]
* Disabling services not required for Devl Training
no secondarynamenode to stop
Stopped Hadoop secondarynamenode: [ OK ]
Flume agent is not running [ OK ]
Stopping Hadoop HBase regionserver daemon: no regionserver to stop because no pid file /var/run/hbase/hbase-hbase-regionserver.pid
hbase-regionserver.
no master to stop because no pid file /var/run/hbase/hbase-hbase-master.pid
Stopped HBase master daemon: [ OK ]
no rest to stop because no pid file /var/run/hbase/hbase-hbase-rest.pid
Stopped HBase rest daemon: [ OK ]
no thrift to stop because no pid file /var/run/hbase/hbase-hbase-thrift.pid
Stopped HBase thrift daemon: [ OK ]
Stopped Kafka Server: [ OK ]
Solr server daemon is not running [ OK ]
[Dev] Done setting up for Devl Training
[training@localhost ~]$
```

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

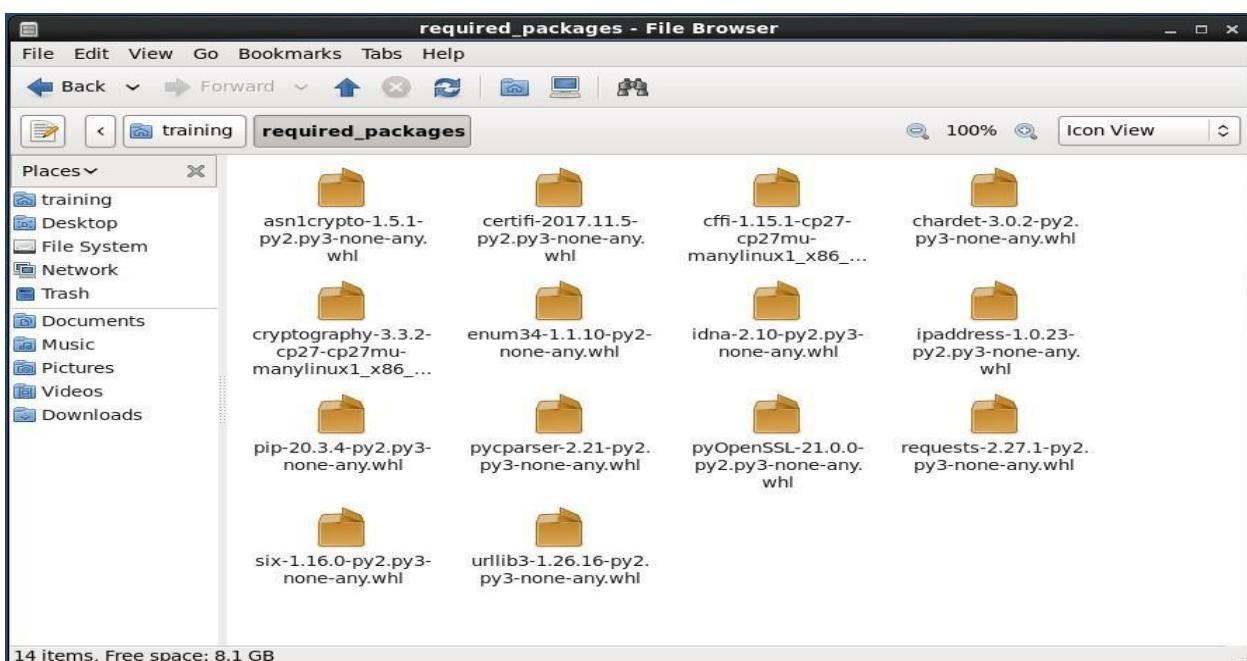


```
File Edit View Search Terminal Help
[training@localhost ~/required_packages]$ sudo ln -sf /user/local/bin/python2.7 /usr/bin/python
[training@localhost required_packages]$
```

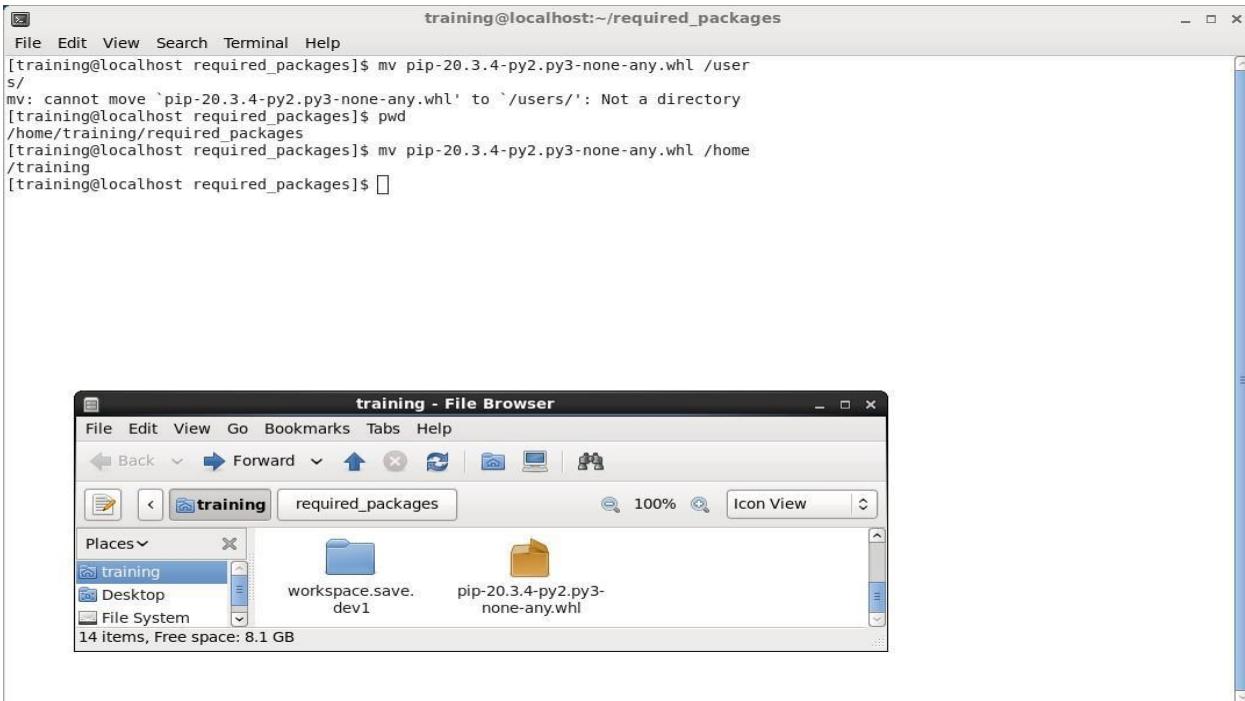
5. Create a directory in the user directory of your VM and call it required_packages:
/home/training/required_packages



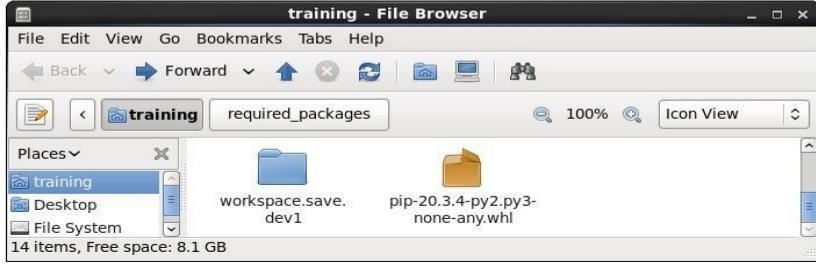
6. Copy the contents of the Required Packages on eLearning to your VM to the directory you created in the previous step (required_packages)



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

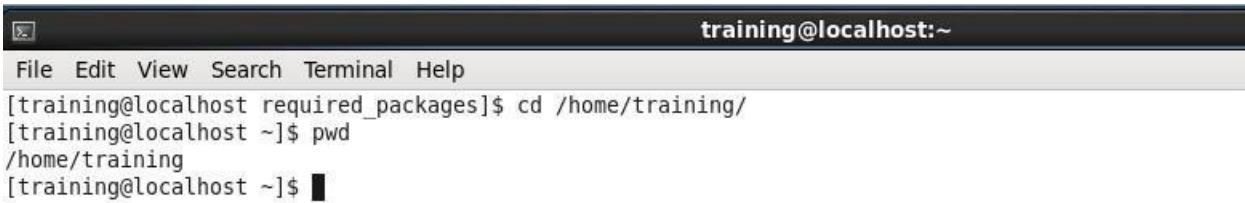


```
training@localhost:~/required_packages
File Edit View Search Terminal Help
[training@localhost required_packages]$ mv pip-20.3.4-py2.py3-none-any.whl /user
mv: cannot move `pip-20.3.4-py2.py3-none-any.whl' to `/users/': Not a directory
[training@localhost required_packages]$ pwd
/home/training/required_packages
[training@localhost required_packages]$ mv pip-20.3.4-py2.py3-none-any.whl /home/training
[training@localhost required_packages]$
```

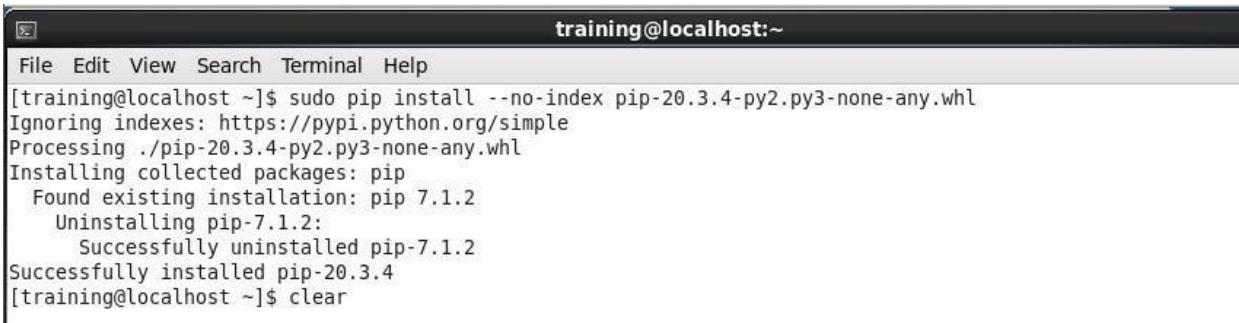
The file browser window shows the file "pip-20.3.4-py2.py3-none-any.whl" located in the "required_packages" folder under the "training" directory.

8. Change your working directory to /home/training



```
training@localhost:~
File Edit View Search Terminal Help
[training@localhost required_packages]$ cd /home/training/
[training@localhost ~]$ pwd
/home/training
[training@localhost ~]$
```

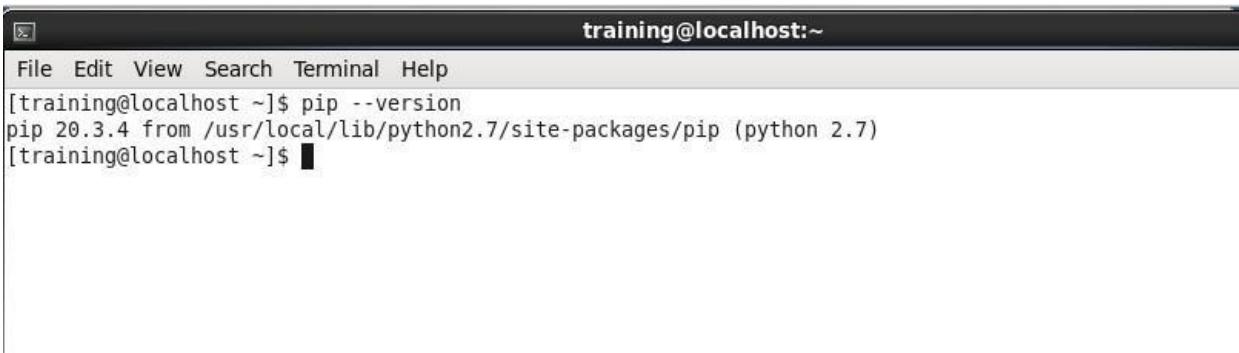
9. Upgrade the pip on your machine by executing the following command in Terminal:



A screenshot of a terminal window titled "training@localhost:~". The window shows the command "sudo pip install --no-index pip-20.3.4-py2.py3-none-any.whl" being run. The output indicates that pip 7.1.2 is being uninstalled and replaced with pip 20.3.4. The process is completed successfully.

```
File Edit View Search Terminal Help
[training@localhost ~]$ sudo pip install --no-index pip-20.3.4-py2.py3-none-any.whl
Ignoring indexes: https://pypi.python.org/simple
Processing ./pip-20.3.4-py2.py3-none-any.whl
Installing collected packages: pip
  Found existing installation: pip 7.1.2
    Uninstalling pip-7.1.2:
      Successfully uninstalled pip-7.1.2
Successfully installed pip-20.3.4
[training@localhost ~]$ clear
```

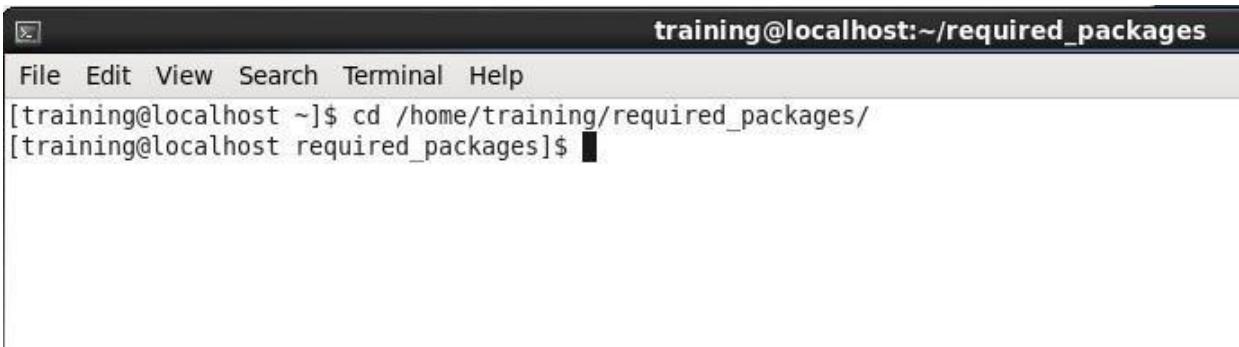
10. Verify that the pip has been upgraded by executing the following:



A screenshot of a terminal window titled "training@localhost:~". The command "pip --version" is run, displaying the output "pip 20.3.4 from /usr/local/lib/python2.7/site-packages/pip (python 2.7)".

```
File Edit View Search Terminal Help
[training@localhost ~]$ pip --version
pip 20.3.4 from /usr/local/lib/python2.7/site-packages/pip (python 2.7)
[training@localhost ~]$
```

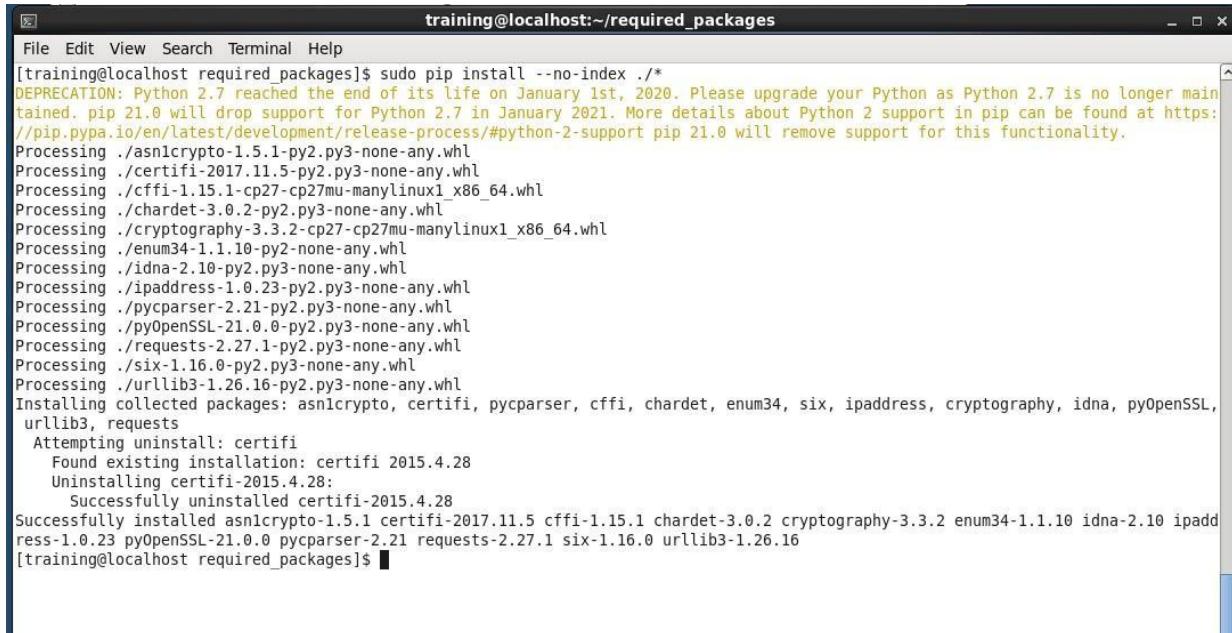
11. Change your directory to /home/training/required_packages



A screenshot of a terminal window titled "training@localhost:~/required_packages". The command "cd /home/training/required_packages/" is run, changing the current working directory to the specified path.

```
File Edit View Search Terminal Help
[training@localhost ~]$ cd /home/training/required_packages/
[training@localhost required_packages]$
```

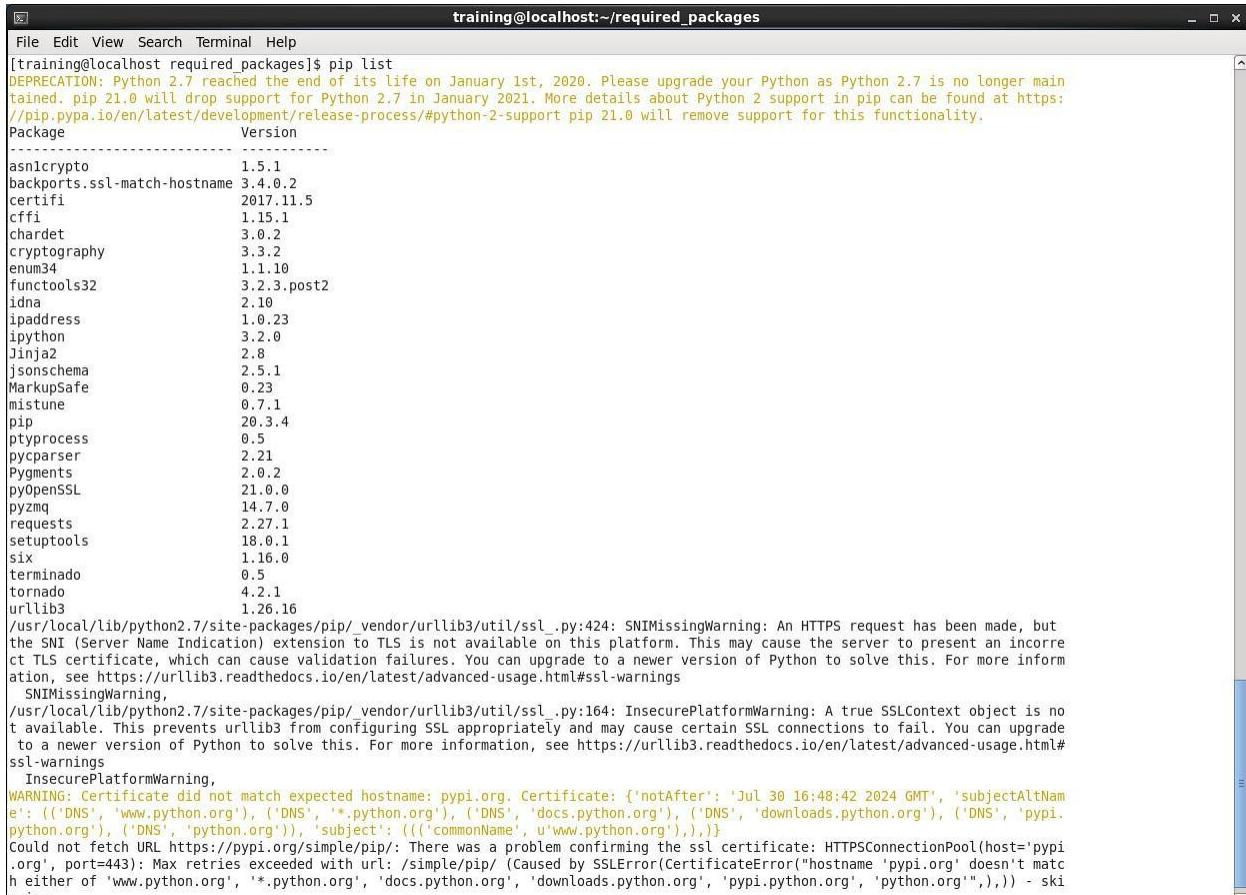
12. Install all the packages in the directory by executing the following in Terminal:



A screenshot of a terminal window titled "training@localhost:~/required_packages". The window shows the command "sudo pip install --no-index ./*" being run. The output indicates that Python 2.7 is reaching the end of its life and advises upgrading to Python 3. It lists numerous packages being processed and installed, including asn1crypto, certifi, pycparser, cffi, chardet, enum34, six, ipaddress, cryptography, idna, pyOpenSSL, and urllib3. It also attempts to uninstall certifi and successfully removes it. Finally, it lists all the packages that were successfully installed.

```
[training@localhost required_packages]$ sudo pip install --no-index ./*
DEPRECATION: Python 2.7 reached the end of its life on January 1st, 2020. Please upgrade your Python as Python 2.7 is no longer maintained. pip 21.0 will drop support for Python 2.7 in January 2021. More details about Python 2 support in pip can be found at https://pip.pypa.io/en/latest/development/release-process/#python-2-support pip 21.0 will remove support for this functionality.
Processing ./asn1crypto-1.5.1-py2.py3-none-any.whl
Processing ./certifi-2017.11.5-py2.py3-none-any.whl
Processing ./cffi-1.15.1-cp27mu-manylinux1_x86_64.whl
Processing ./chardet-3.0.2-py2.py3-none-any.whl
Processing ./cryptography-3.3.2-cp27-cp27mu-manylinux1_x86_64.whl
Processing ./enum34-1.1.10-py2-none-any.whl
Processing ./idna-2.10-py2.py3-none-any.whl
Processing ./ipaddress-1.0.23-py2.py3-none-any.whl
Processing ./pycparser-2.21-py2.py3-none-any.whl
Processing ./pyOpenSSL-21.0.0-py2.py3-none-any.whl
Processing ./requests-2.27.1-py2.py3-none-any.whl
Processing ./six-1.16.0-py2.py3-none-any.whl
Processing ./urllib3-1.26.16-py2.py3-none-any.whl
Installing collected packages: asn1crypto, certifi, pycparser, cffi, chardet, enum34, six, ipaddress, cryptography, idna, pyOpenSSL, urllib3, requests
Attempting uninstall: certifi
  Found existing installation: certifi 2015.4.28
  Uninstalling certifi-2015.4.28:
    Successfully uninstalled certifi-2015.4.28
Successfully installed asn1crypto-1.5.1 certifi-2017.11.5 cffi-1.15.1 chardet-3.0.2 cryptography-3.3.2 enum34-1.1.10 idna-2.10 ipaddress-1.0.23 pyOpenSSL-21.0.0 pycparser-2.21 requests-2.27.1 six-1.16.0 urllib3-1.26.16
[training@localhost required_packages]$
```

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



```
File Edit View Search Terminal Help
[training@localhost required_packages]$ pip list
DEPRECATION: Python 2.7 reached the end of its life on January 1st, 2020. Please upgrade your Python as Python 2.7 is no longer maintained. pip 21.0 will drop support for Python 2.7 in January 2021. More details about Python 2 support in pip can be found at https://pip.pypa.io/en/latest/development/release-process/#python-2-support pip 21.0 will remove support for this functionality.
Package           Version
-----
asnincrypto      1.5.1
backports.ssl-match-hostname 3.4.0.2
certifi          2017.11.5
cffi              1.15.1
chardet          3.0.2
cryptography     3.3.2
enum34           1.1.10
functools32      3.2.3.post2
idna              2.10
ipaddress        1.0.23
ipython           3.2.0
Jinja2            2.8
jsonschema       2.5.1
MarkupSafe        0.23
mistune          0.7.1
pip               20.3.4
ptyprocess        0.5
pycparser         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
/usr/local/lib/python2.7/site-packages/pip/_vendor/urllib3/util/ssl_.py:424: SNIMissingWarning: An HTTPS request has been made, but the SNI (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 SSLContext 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,
WARNING: Certificate did not match expected hostname: pypi.org. Certificate: {'notAfter': 'Jul 30 16:48:42 2024 GMT', 'subjectAltName': (('DNS', 'www.python.org'), ('DNS', '*.python.org'), ('DNS', 'docs.python.org'), ('DNS', 'downloads.python.org'), ('DNS', 'pypi.python.org'), ('DNS', 'python.org')), 'subject': ((('commonName', u'www.python.org'),),)}
```

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.

```
# Directory containing your JSON files\n",
"downloads_directory = r\"content\"\n",
"\n",
"\n# Create an empty list to store all the block data\n",
"all_block_data = [ ]\n",
"\n",
"\n# Function to fetch block information by hash\n",
"def fetch_block_info(hash_value):\n",
"    api_url = f\"https://blockchain.info/rawBlock/{hash_value}\"\n",
"    try:\n",
"        response = requests.get(api_url)\n",
"        if response.status_code == 200:\n",
"            data = response.json()\n",
"            # Remove the 'tx' key to exclude transaction data\n",
"            data.pop('tx', None)\n",
"            return data\n",
"        else:\n",
"            print(f\"Failed to fetch data for hash {hash_value}. Status code: {response.status_code}\")\n",
"    except Exception as e:\n",
"        print(f\"Error fetching data for hash {hash_value}: {e}\")\n",
"    return None\n",
"\n",
"\n# Define the common part of the file name\n",
"common_file_name = \"blockchain_data_day_\"\n",
"\n",
"\n# Iterate over the days\n",
"for day in range(0, 21):\n",
"    filename = f\"{common_file_name}{day}.json\"\n",
"    json_path = os.path.join(downloads_directory, filename)\n",
"\n",
"    if os.path.exists(json_path):\n",
"        with open(json_path, \"r\") as file:\n",
"            block_data = json.load(file)\n",
"\n",
"            # Extract the block hash from the list of block data and fetch additional info\n",
"            for block in block_data:\n",
"                block['hash'] = block['get('hash')']\n",
"                if hash_value:\n",
"                    fetched_data = fetch_block_info(hash_value)\n",
"                    if fetched_data:\n",
"                        all_block_data.append(fetched_data)\n",
"\n",
"\n# Save all the block data into a single output file\n",
"output_path = os.path.join(downloads_directory, \"blockchain_data.json\")\n",
```

- 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]

The screenshot shows the Hue File Browser interface in Mozilla Firefox, displaying a list of files in the 'Project1' directory. The table below provides details about the transferred file.

Name	Size	User	Group	Permissions	Date
blockchain_data.txt	683.9 KB	hdbs	supergroup	drwxr-xr-x	March 24, 2024 05:15 PM
training		training	supergroup	drwxrwxr-x	March 24, 2024 05:29 PM
		training	supergroup	-rwxrwxr--	March 24, 2024 05:29 PM

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

```
    row = [key] + value.values()
    csv_data.append(row)

return csv_data

# Function to fetch stock data and send it to HDFS
def fetch_stock_data():

    api.key = 'EY1NWVXT000005K8'

    for symbol in config["symbols"]:

        url = "http://128.118.194.104/mars/function-TIME_CERTIFIC_INTRADAY?symbol={symbol}&interval={interval}&count={count}&symbol={symbol}&api_key={api.key}").format(symbol=symbol, interval=interval, count=count, api_key=api.key)
```

3. Setup and configure your 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]

```
# cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player | ||| & Applications Places System File Edit View Search Tools Documents Help
File flume.conf (~/training_materials/dev1/exercises/flume) - gedit
flume.conf (~/training_materials/dev1/exercises/flume) - gedit
# Flume.conf file
# Define the agent
agent1.sources = netcatSource
agent1.channels = memoryChannel
agent1.sinks = hdfsSink loggerSink

# Configure Source
agent1.sources.netcatSource.type = netcat
agent1.sources.netcatSource.bind = localhost
agent1.sources.netcatSource.port = 12345
agent1.sources.netcatSource.max-line-length = 5000

# Configure Channel
agent1.channels.memoryChannel.type = memory
agent1.channels.memoryChannel.capacity = 1000
agent1.channels.memoryChannel.transactionCapacity = 100

# Configure Sink to write data to HDFS
agent1.sinks.hdfsSink.type = hdfs
agent1.sinks.hdfsSink.hdfs.path = hdfs://localhost:8020/flume/Project1_new/
agent1.sinks.hdfsSink.hdfs.fileType = DataStream
agent1.sinks.hdfsSink.hdfs.writeFormat = Text
agent1.sinks.hdfsSink.hdfs.rollSize = 524288
agent1.sinks.hdfsSink.hdfs.rollCount = 0

# Configure Logger Sink for Logging
agent1.sinks.loggerSink.type = logger
agent1.sinks.loggerSink.channel = memoryChannel

# Connect Source, Channel, and Sink
agent1.sources.netcatSource.channels = memoryChannel
agent1.sinks.hdfsSink.channel = memoryChannel
```

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

The screenshot shows a Mozilla Firefox browser window titled "Hue - File Browser - Mozilla Firefox". The address bar displays "localhost:8888/filebrowser/#flume/Project1_new". The main content area is a file browser interface with a sidebar on the left and a detailed table view on the right.

Table Headers:

Name	Size	User	Group	Permissions	Date
------	------	------	-------	-------------	------

Data Rows:

FlumeData.1711283003779	50.0 KB	training	supergroup	rwxrwxrwx	March 23, 2024 07:37 PM
FlumeData.1711283162979	513.5 KB	training	supergroup	rwxrwxrwx	March 24, 2024 05:25 AM
FlumeData.1711283162980	389.5 KB	training	supergroup	rwxrwxrwx	March 24, 2024 05:26 AM
FlumeData.1711283251004	513.5 KB	training	supergroup	rwxrwxrwx	March 24, 2024 05:27 AM
FlumeData.1711283251005	398.5 KB	training	supergroup	rwxrwxrwx	March 24, 2024 05:28 AM
FlumeData.1711283337788	513.5 KB	training	supergroup	rwxrwxrwx	March 24, 2024 05:29 AM
FlumeData.1711283337889	49.0 KB	training	supergroup	rwxrwxrwx	March 24, 2024 05:29 AM

At the bottom of the browser window, there are several tabs: "flume - File Browser", "training@localhost:~\$", "Hue - File Browser - M...", and "flume.conf (~/training...)".

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	00000000000000000000000000000000540268ddfc73d8cd7348eb48695fe4a602708c89b2e4	2023-09-10 00:00:00	806982
1	000000000000000000000000000000003f3a19f16e20cc5f50f108ed7dabd5957f2e61e90868d	2023-09-10 00:00:00	806981
2	00000000000000000000000000000000372109b9a114633512587c8b074910a4bc02921828b59	2023-09-10 00:00:00	806980
3	00000000000000000000000000000000438def0efe8257f6ff025665074dbc7e1fda070aff30	2023-09-10 00:00:00	806979
4	0000000000000000000000000000000022e8847528953ebe30fc81867dc9ab70b50ce2660d5df	2023-09-10 00:00:00	806978

```
select * from blocks_info_2023_Sep_10_to_15 LIMIT 5;
```

```
mysql> select * from blocks_info_2023_sep_10_to_15 limit 5;
```

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;
+----+
```

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 --connection jdbc:mysql://localhost/project1 --username training --password training --query "SELECT * FROM blocks_2023_Sep_10_to_15 LIMIT 5"
[2/8/18 15:25:02] INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
[2/8/18 15:25:02] WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
[2/8/18 15:25:02] INFO tool.BaseSqoopTool: Using default connector MySQL
[2/8/18 15:25:02] INFO manager.SqlManager: Drop table if exists blocks_2023_Sep_10_to_15
[2/8/18 15:25:02] INFO manager.SqlManager: Create table blocks_2023_Sep_10_to_15
```

```
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: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/18 15:26:33 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	00000000000000000000000000000000540268ddfc73d8cd7340eb48695fe4a602708c89b2e4	545259520	386216622	17380913	3862147940	1733081	806982	0	806982	3993482	
1	00000000000000000000000000000000313a19f10e20c5f50f108e07da0d5957f2e61e90868d	536879104	386216622	17655615	822115101	1747360	806981	0	806981	3993154	
2	00000000000000000000000000000000372109099114633512587cb0749104ac0c2921828b59	54735672	386216622	15397499	2453646967	1797458	806980	0	806980	3993365	
3	00000000000000000000000000000000438def0e8e257f6ff025665074dcfc7e1fda070aff30	887709696	386216622	1474632	2618089852	1913649	806979	0	806979	3993468	
4	0000000000000000000000000000000022ze847528953eb30fc1867dc9ab7050ce2660df	53687912	386216622	17157713	70083109	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: 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
0	ee63374e93ff48b7ef212e00696d6d019140ea23c32c7642d36c3d4179c5119	00000000000000000000000000000000540268ddfc73d8cd7340eb48695fe4a602708c89b2e4	1	1	4	351	1296	0				
1	991820990367693	0	1694313239	806982	806982							
1	cec2973eacc7db4c4ff4108e4caf11a1abe718040b7580af6b4efc2c8e6	00000000000000000000000000000000540268ddfc73d8cd7340eb48695fe4a602708c89b2e4	1	1	1	192	438	33148				
2	6798aae6b2cc9af0b581a3e7b32b241ddbb9134a99d5533e0e0e34505f1911	00000000000000000000000000000000806982	806982	806982	806982	380	758	50000				
3	4cef8574937250098114efc6ab22c220745708227fbfb525d0c7340eb48695fe4a602708c89b2e4	0	1694319236	806982	806982	1	1	193	442	24406		
4	79416846f1f07597964891825151c5321fc5ae688a2b6f16e9c372279bc971a	00000000000000000000000000000000540268ddfc73d8cd7340eb48695fe4a602708c89b2e4	2	1	1	191	437	20618				
0	936557279589848	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
```

```
doodera-training-G1psarl-student-rev_cd_4.3a -VMware Workstation 17P1, type[Non-commercial use only]
" " • II • : 15!
ii Applications Places System • II J
File Edit View search Terminal Help
Try--helpforusageinstructions.
[training@localhost ~]$sqoop import --connect jdbc:mysql://localhost/project1
--username training
--password training
--table bloct2823Sep18to15
hive-import
--hive-table project1.blocks2823Sep18to15
--hive-overwrite
--m-mappers 4
24/8/18 15:56:58 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-edh$4.3
24/8/18 15:56:58 WARN tool.BiSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/8/18 15:56:58 INFO tool.BiSqoopTool: Using Hive-specific delimiters for output. You can override
24/8/18 15:56:58 INFO tool.BiSqoopTool: delimiters with --fields-separated-by, etc.
24/8/18 15:56:51 INFO manager.HiveManager: Preparing to use a MySQL streaming resultset.
24/8/18 15:56:51 INFO OTool.CodeGenTool:Beginningcodegen
24/8/18 15:56:52 INFO manager.HiveManager: Executing SQL statement: SELECT t FROM `blocks` 2823 Sep 18 to 15 AS t UHIT 1
24/8/18 15:56:52 INFO manager.HiveManager: Executing SQL statement: SELECT t FROM `blocks` 2823-Sep-18-to-15 AS t UHIT 1
24/8/18 15:56:52 INFO manager.HiveManager: HADOOP_HOME is /usr/lib/hadoop-apache- - -
Note: hive-import-training:java:7@34afe52b12579b38a5fdc2159474/blocks_2923 Sep 19 to 15 uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/8/18 15:56:57 INFO org.apache.hadoop.mapred.lib.SqoopMapper: Writing jar file: /tmp/sqoop/training/jar/_job_17d34afe52b12579b38a5fdc2159474/blocks_2923 Sep 19 to 15.jar
24/8/18 15:56:57 WARN manager.HiveManager: It looks like you are importing from MySQL.
24/8/18 15:56:57 WARN manager.HiveManager: This transfer can be faster! Use the --direct
24/8/18 15:56:57 WARN manager.HiveManager: option to exercise a MySQL-specific fast path.
24/8/18 15:56:57 INFO apreduce.LibOrJobElse: Setting zero DATETIME behavior to convertToNull inmysql
24/8/18 15:56:57 INFO apreduce.LibOrJobElse: Beginning import of blocks_2923 Sep 18 to 15
24/8/18 15:56:57 INFO Configuration.deprecation: libpred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/8/18 15:56:57 INFO configuration.deprecation: apred.jar is deprecated. Instead, use mapreduce.job.apbs
24/8/18 15:57:09 INFO client.MrProxy: connecting to esourceManager at 9.8.8.8:8832
24/8/18 15:57:10 INFO client.MrProxy: Using readCommitter=transacted, isolation=S ECTCHI||| id"1, M.L.X("id"1 F Rrfl "blocks_2823Sep18 to15"
24/8/18 15:57:11 INFO apreduce.JOOSubmitter: number of splits:4
24/8/18 15:57:12 INFO apreduce.JOOSubmitter: Submitting tokens for job: job_1787673354943_9685
24/8/18 15:57:12 INFO apreduce.JOOSubmitter: Submitted application application_178767BS4943_98e6
24/8/18 15:57:14 INFO apreduce.Job: The url to track the job: http://localhost:88118/pro/y/application/1797673354943_98e6/
24/8/18 15:57:14 INFO apreduce.Job: Running job: job_1787673354943_9896
24/8/18 15:57:45 INFO apreduce.Job: Job job_1787673354943_88e6 running in uber mode : false
24/8/18 15:57:45 INFO apreduce.JOO libp 8%reduce 9%
24/8/18 15:58:55 INFO apreduce.JOO libp 15% reduce 8%
24/8/18 15:58:22 INFO apreduce.JOO libp 58% reduce 8%
24/8/18 15:58:49 INFO apreduce.Job: libp 751; reduce 8%
24/8/18 15:58:50 INFO apreduce.Job: ap 188; reduce 8%
24/8/18 15:58:51 INFO apreduce.Job: Job job_1787613354943_8886llalleted successfully
24/8/18 15:58:51 INFO apreduce.Job counters: 38
```

File System Counter
FILE: Number of bytes read=8
FILE: Number of bytes written=c53968

training@localhost: ~ (p_datasets-File ... & Hue-File Brow - M...

OOT

```
layer: 11 • : 1/1
• Applications Places System I Ii [i
File Edit View search Terminal Help
FILE: Number of large read operations=8
FILE: Number of write operations=6
HDFS: Mapped bytes re B7
11DFS: Number of bytes written=9058
11DFS: Number of read operations<16
11DFS: Number of large read operations<9
11DFS: Number of write operations.
Job counters
Unfinished app tasks=4
Other failed tasks=4
Total bytes in occupied slots(ms):cc9
Total time spent by all map tasks (ms)=55338
Total vcore-second taken by all map tasks=55338
Total byte-seconds taken by all map tasks=1415448E
Map-Reduce Framework
Killed pure tasks=28
Killed pure tasks=928
Input records=x&J
Spilled records=B
Failed Shuffles=B
Killed Kap outputs=&
GC time elapsed (ms)=1671
CPU time spent (ms)=1671
Physical memory (bytes) snapshot=5BBB72448
Virtual memory (bytes) snapshot=3377397768
Total allocated heap usage(bytes)=198894B
File Input Format Counters
Bytes Read=8
File output Format counters
Bytes Written=98959
24/8/18 15:58:51 INFO mapreduce.ImportJobBase: Transferred 87.9395 KB in 118.311 seconds (916.3275 bytes/sec)
24/8/18 15:58:51 INFO mapreduce.ImportJobBase: Retrieved 929 records.
24/8/18 15:58:51 INFO manager.SqoopManager: Executing SQL statement: SELECT t FROM `blocks` 2923 Sep 18 to 15 AS t UNIT 1
24/8/18 15:58:51 W/H hive.TableDefWriter: Cohan time had to be cast to a less precise type in Hive:- -
24/8/18 15:58:51 INFO hive.HiveImport: Loading uploaded data into Hive
Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-correctly-1.1.8-cdh5.4.3.jar!/hive-log4j.properties
OK
Time taken: 9.5 S (sec)
Loading data to table project1.blocks 2823 sep 19 to 15
chgrp: changing ownership of 'hdfs://localhost:8922/user/hive/warehouse/project1.rdb(blocks_2823_sep_19_to_15/part-00000' User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8922/user/hive/warehouse/project1.rdb(blocks_2823_sep_19_to_15/part-11-88881' User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8922/user/hive/warehouse/project1.rdb(blocks_2823_sep_19_to_15/part-11-88881' User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8922/user/hive/warehouse/project1.rdb(blocks_2823_sep_19_to_15/part-11-88881' User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8922/user/hive/warehouse/project1.rdb(blocks_2823_sep_19_to_15/part-11-88881' User does not belong to hive
Table project1.blocks 2823 sep 19 to 15 stats: [numfiles=4, maxRows=9, totalSize=9e859, rawDataSize=1]
Time taken: 3.935 second
[training@localhost ~]
```

training@localhost: ~ (sqoop_datasets - File ... & Hue - File Brow - M...)

f Mon-Edu9, 4:03PM

Q

```

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

```

```

[training@localhost ~]$ 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
24/03/18 16:07:49 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.3
24/03/18 16:07:49 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/03/18 16:07:49 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
24/03/18 16:07:49 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
24/03/18 16:07:51 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/03/18 16:07:51 INFO tool.CodeGenTool: Beginning code generation
24/03/18 16:07:52 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `blocks_info_2023 Sep 10 to 15` AS t LIMIT 1
24/03/18 16:07:52 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `blocks_info_2023 Sep 10 to 15` AS t LIMIT 1
24/03/18 16:07:52 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/2af9f47c4585dcadd2d2847ac6056c5f2/blocks_info_2023_Sep_10_to_15.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/03/18 16:08:01 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/compile/2af9f47c4585dcadd2d2847ac6056c5f2/blocks_info_2023_Sep_10_to_15.jar
24/03/18 16:08:01 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/03/18 16:08:01 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/03/18 16:08:01 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/03/18 16:08:01 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/03/18 16:08:01 INFO mapreduce.ImportJobBase: Beginning import of blocks_info_2023 Sep_10_to_15
24/03/18 16:08:02 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/03/18 16:08:02 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/03/18 16:08:11 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/03/18 16:08:11 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/03/18 16:08:23 INFO db.DBInputFormat: Using read committed transaction isolation
24/03/18 16:08:23 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN('id'), MAX('id') FROM `blocks_info_2023 Sep_10_to_15`
24/03/18 16:08:23 INFO mapreduce.JobSubmitter: number of splits:4
24/03/18 16:08:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1707673354943_0007
24/03/18 16:08:24 INFO impl.YarnClientImpl: Submitted application application_1707673354943_0007
24/03/18 16:08:24 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1707673354943_0007/
24/03/18 16:08:24 INFO mapreduce.Job: Running job: job_1707673354943_0007
24/03/18 16:08:51 INFO mapreduce.Job: Job job_1707673354943_0007 running in uber mode : false
24/03/18 16:08:51 INFO mapreduce.Job: map 0% reduce 0%
24/03/18 16:09:22 INFO mapreduce.Job: map 25% reduce 0%
24/03/18 16:09:37 INFO mapreduce.Job: map 50% reduce 0%
24/03/18 16:09:53 INFO mapreduce.Job: map 75% reduce 0%
24/03/18 16:10:05 INFO mapreduce.Job: map 100% reduce 0%
24/03/18 16:10:08 INFO mapreduce.Job: Job job_1707673354943_0007 completed successfully
24/03/18 16:10:08 INFO mapreduce.Job: Counters: 30
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=544312

```

```

cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player | || ▾ ▷ 🔍
Applications Places System 🌐 📁
File Edit View Search Terminal Help
training@localhost:~ Mon Mar 18, 4:11 PM
File Input Format Counters
  Bytes Read=0
  File Output Format Counters
  Bytes Written=44756
24/03/18 16:18:07 INFO mapreduce.ImportJobBase: Transferred 43.707 KB in 116.5168 seconds (384.1163 bytes/sec)
24/03/18 16:18:07 INFO mapreduce.ImportJobBase: Retrieved 310 records.
24/03/18 16:18:07 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `blocks_info_2023_Sep_10_to_15` AS t LIMIT 1
24/03/18 16:18:07 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: 6.16 seconds
Loading data to table project1.blocks_info 2023 sep 10 to 15
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/project1.db(blocks_info 2023 sep 10 to 15/part-m-00000)': User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/project1.db(blocks_info 2023 sep 10 to 15/part-m-00001)': User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/project1.db(blocks_info 2023 sep 10 to 15/part-m-00002)': User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8020/user/hive/warehouse/project1.db(blocks_info 2023 sep 10 to 15/part-m-00003)': User does not belong to hive
Table project1.blocks_info 2023 sep 10 to 15 stats: [numFiles=4, numRows=0, totalSize=44756, rawDataSize=0]
OK
Time taken: 3.263 seconds
[training@localhost ~]$ 

```

```

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-training11g-capspark-student-rev_cdh5.4.3a VMware Workstation 17 Player (Non-commercial use only)
Elayer • || * □ I"il_
Applications Places System • || ij
File Edit View Search Terminal Help
[training@localhost ~]$ clear

[training@localhost ~]$ sqoop import
--connect jdbc:mysql://localhost/project
--username training
--password training
--table txinfo2B23SepHito5
--hive-import \
  hive-table=txinfo2B23SepHito5
--hive-overwrite \
--num-mappers 4
24/03/18 16:16:22 INFO sqoop: Running sqoop version: 1.4.5-cdh5.4.3
24/03/18 16:16:22 WARN tool.BaseSqoopTool: Setting your password in the config/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
24/03/18 16:16:22 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by.
24/03/18 16:16:23 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/03/18 16:16:23 INFO tool.CodeGenTool: Beginning code generation.
24/03/18 16:16:25 INFO manager.SqlManager: Executing SQL statement: SELECT t_ FROM tx_info 2023 Sep 18 to 15' AS t LIMIT 1
24/03/18 16:16:25 INFO org.apache.hadoop.mapred.JobClient$JobTracker: HADOOP_NAMENODE_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-training/compile/d285413f568359619ab5dd3849ff8753/tx_info 2023 Sep 18 to 15.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/03/18 16:16:25 INFO manager.MySQLManager: Executing SQL statement: SELECT t_ FROM tx_info 2023 Sep 18 to 15' AS t LIMIT 1
24/03/18 16:16:34 INFO manager.MySQLManager: It looks like you are importing from mysql.
24/03/18 16:16:34 WARN manager.MySQLManager: This transfer may be faster! Use the --direct
24/03/18 16:16:34 INFO manager.MySQLManager: Option to exercise a MySQL specific fast import.
24/03/18 16:16:34 INFO manager.MySQLManager: Using zero DATETIME behavior by command-line (mysql)
24/03/18 16:16:34 INFO mapreduce.ImportJobBase: Beginning import of tx_info 2023 Sep 18 to 15
24/03/18 16:16:34 INFO configuration.deprecation: mapred.job.tracker is deprecated; instead, use mapreduce.jobtracker.address
24/03/18 16:16:35 INFO configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/03/18 16:16:39 INFO configuration.deprecation: mapred.map.tasks is deprecated; instead, use mapreduce.job.maps
24/03/18 16:16:49 INFO db.DBInputFormat: Using read committed transaction isolation
24/03/18 16:16:49 INFO mapreduce.JobSubmitter: DataDrivenJobSubmitter: number of splits:4
24/03/18 16:16:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1787673354943_8808
24/03/18 16:16:52 INFO mapreduce.Job: Job job_1787673354943_8808
24/03/18 16:17:28 INFO mapreduce.Job: Job job_1787673354943_8808 running in uber mode : false
24/03/18 16:18:13 INFO mapreduce.Job: map B=reduce E=%
24/03/18 16:18:13 INFO mapreduce.Job: map 25 reduce 8
24/03/18 16:19:42 INFO mapreduce.Job: map SB=reduce B
24/03/18 16:19:55 INFO mapreduce.Job: map 75 reduce 8
24/03/18 16:21:22 INFO mapreduce.Job: map 188 reduce %
24/03/18 16:21:24 INFO mapreduce.Job: Job job_1787673354943_8808 completed successfully
24/03/18 16:21:26 INFO mapreduce.Job: counters: 3Ei
File system counters
FILE: Numberofbytesread=9

```

```

claudera-training11g-capspark-student-rev_cdh5.4.3a VMware Workstation 17 Player (Non-commercial use only)
Elayer • || * □ I"il_
'-5 Applications Places System li iii D
File Edit View search Terminal Help
[training@localhost ~]$
File Edit View search Terminal Help
FILE: Number of read operations=8
FILE: Number of large read operations=8
FILE: Number of write operations=8
HDFS: Number of bytes read=429
HDFS: Number of bytes written=256678856
HDFS: Number of read operations=6
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)=8
Total time spent by all reduces in occupied slots (ms)=8
Total time spent by all map tasks (ms)=234737
Total vcore-seconds taken by all map tasks,234737
Total megabyte-seconds taken by all map tasks=60992672
Map-Reduce Framework
  Map Input records=97184
  Map Output records=97184
  Inputsplitbytes=429
  SpilledRecords=0
  FailedShuffles=0
  Merged Map outputs=8
  GC time elapsed (ms)=3421
  CPUtilspent(ms)=9811B
  Physical memory (bytes) snapshot=517734498
  Virtual memory (bytes) snapshot=3382149128
  Total committed heap usage (bytes)=91889488
File Input Format Counters
  BytesRead=8
File Output Format counters
  Bytes Written=256678856
24/03/18 16:21:26 INFO mapreduce.ImportJobBase: Transferred 244.788 MB in 287.1819 seconds (873.EL799 KB/sec)
24/03/18 16:21:26 INFO mapreduce.ImportJobBase: Retrieved 1197104 records.
24/03/18 16:21:26 INFO manager.SqlManager: Executing SQL statement: SELECT t_ FROM tx_info 2023 Sep 18 to 15' AS t LIMIT 1
24/03/18 16:21:27 INFO hive.HiveImport: Loading uploaded data into Hive
Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-co!Mlon-IJ.B-cdh5.4.3.jar!/hive-log4j.properties
OK
Time taken: 18.818 seconds
Loading data into table project.txinfo2B23sepHito5 took 15
chgrp: changing ownership of 'hdfs://localhost:8820/user/hive/warehouse/project._db/tx_info 2023 sep 18 to 15/part-m-0EIBBB' User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8820/user/hive/warehouse/project._db/tx_info 2023 sep 18 to 15/part-m-0EIBB1' User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8820/user/hive/warehouse/project._db/tx_info 2023 sep 18 to 15/part-m-0EIBB2' User does not belong to hive
chgrp: changing ownership of 'hdfs://localhost:8820/user/hive/warehouse/project._db/tx_info 2023 sep 18 to 15/part-m-0EIBB3' User does not belong to hive
Table project.tx_info 2023 sep 18 to 15 stats: [numFiles:4, numRows: 81, totalSize:256678856, rawDataSize:81]
OK
Time taken: 5.896 seconds
[training@localhost ~]$ clear

```

5. Verify the import by browsing Hive

```
SELECT * from blocks_2023_Sep_10_to_15 LIMIT 5;
```

The screenshot shows the Hue interface with the 'Query Editor' tab selected. A query has been run:

```
1. SELECT * from blocks_2023_Sep_10_to_15 LIMIT 5;
```

The results table has five columns:

	blocks_2023_sep_10_to_15.id	blocks_2023_sep_10_to_15.hash	blocks_2023_sep_10_to_15.time	blocks_2023_sep_10_to_15.block_index
0	0	0000000000000000000540268dcf738cd7348eb48695fe4a02708c892e4	2023-09-10 00:00:00.0	806982
1	1	0000000000000000000313a1916e20c5f50f08ed7dab59572e61e9096bd	2023-09-10 00:00:00.0	806981
2	2	000000000000000000037210989a114633812587c8b074910a4bc02921828e59	2023-09-10 00:00:00.0	806980
3	3	0000000000000000000438de0fe625716f025665674dbcf7e16a070af30	2023-09-10 00:00:00.0	806979
4	4	00000000000000000002e8847528953ebe30f81867dc9ab7050ce266005df	2023-09-10 00:00:00.0	806978

```
SELECT * from blocks_info_2023_Sep_10_to_15 LIMIT 5;
```

The screenshot shows the Hue interface with the 'Query Editor' tab selected. A query has been run:

```
1. SELECT * from blocks_info_2023_Sep_10_to_15 LIMIT 5;
```

The results table has eight columns:

	blocks_info_2023_sep_10_to_15.id	blocks_info_2023_sep_10_to_15.hash	blocks_info_2023_sep_10_to_15.ver	blocks_info_2023_sep_10_to_15.bits	blocks_info_2023_sep_10_to_15.fee	blocks_info_2023_sep_10_to_15.block_index	blocks_info_2023_sep_10_to_15.time
0	0	0000000000000000000540268dcf738cd7348eb48695fe4a02708c892e4	545259520	386216622	17380913	386216622	2023-09-10 00:00:00.0
1	1	0000000000000000000313a1916e20c5f50f08ed7dab59572e61e9096bd	536979104	386216622	17655615	822115	2023-09-10 00:00:00.0
2	2	000000000000000000037210989a114633812587c8b074910a4bc02921828e59	547556672	386216622	15397499	245364	2023-09-10 00:00:00.0
3	3	0000000000000000000438de0fe625716f025665674dbcf7e16a070af30	887709696	386216622	14746325	261808	2023-09-10 00:00:00.0
4	4	00000000000000000002e8847528953ebe30f81867dc9ab7050ce266005df	536870912	386216622	17157713	70003	2023-09-10 00:00:00.0

`SELECT * from tx_info_2023_Sep_10_to_15 LIMIT 5;`

The screenshot shows the Hue interface for running Hive queries. The query `SELECT * from tx_info_2023_Sep_10_to_15 LIMIT 5;` has been entered into the query editor. The results pane displays five rows of data from the `tx_info_2023_Sep_10_to_15` table. The columns shown are `tx_info_2023.sep.10_to.15.id`, `tx_info_2023.sep.10_to.15.tx.hash`, `tx_info_2023.sep.10_to.15.block.hash`, and `tx_info_2023.sep.10_to.15.ver`. The data includes various transaction IDs and their corresponding hashes.

6. Show where the actual data is stored on HDFS

The screenshot shows the Hue File Browser interface. The current path is `/user/hive/warehouse/project1.db`. The browser lists several tables and their partitions: `blocks_2023_sep_10_to_15`, `blocks.info_2023_sep_10_to_15`, and `tx_info_2023_sep_10_to_15`. Each entry shows its size, owner, group, permissions, and last modified date. The table `tx_info_2023_sep_10_to_15` has a size of 45 bytes and was last modified on March 18, 2024, at 04:21 PM by user `training`.

A. Data Analysis using Pig

Hue - File Browser - Mozilla Firefox

localhost:8888/filebrowser/view/user/training

File Browser

Name	Size	User	Group	Permissions	Date
output.csv	287.3 KB	training	supergroup	-rw-rw-rw-	April 19, 2024 11:15 AM

Show 45 of 1 items

Page 1 of 1

Firefox automatically sends some data to Mozilla so that we can improve your experience.

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

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

-- Loading the dataset

```
data = LOAD '/user/training/output.csv' USING PigStorage(',') AS (
    hash:chararray,
    ver:int,
    prev_block:chararray,
    mrkl_root:chararray,
    time:long,
    bits:int,
    fee:long,
    nonce:long,
    n_tx:int,
    size:long,
    block_index:int,
    main_chain:boolean,
    height:int,
    weight:long
);
```

-- 1) Count the total number of blocks

```
total_blocks = FOREACH (GROUP data ALL) GENERATE COUNT(data);
```

-- Output the result

```
DUMP total_blocks;
```

```
[>7] 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, LoadTypeCast inserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCast inserter}, 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 optimization? 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.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
```

```
[cl]                                     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  AvgReduceTim
e       MedianReducetime   Alias  Feature outputs
job 1709496148644 0002 1      10        10        10        10        8         8         8         8         1-10,data,total blocks  GROU
P_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 \'la
rning FIBLO 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, u
se 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
1
(1027)
grunt> clear
```

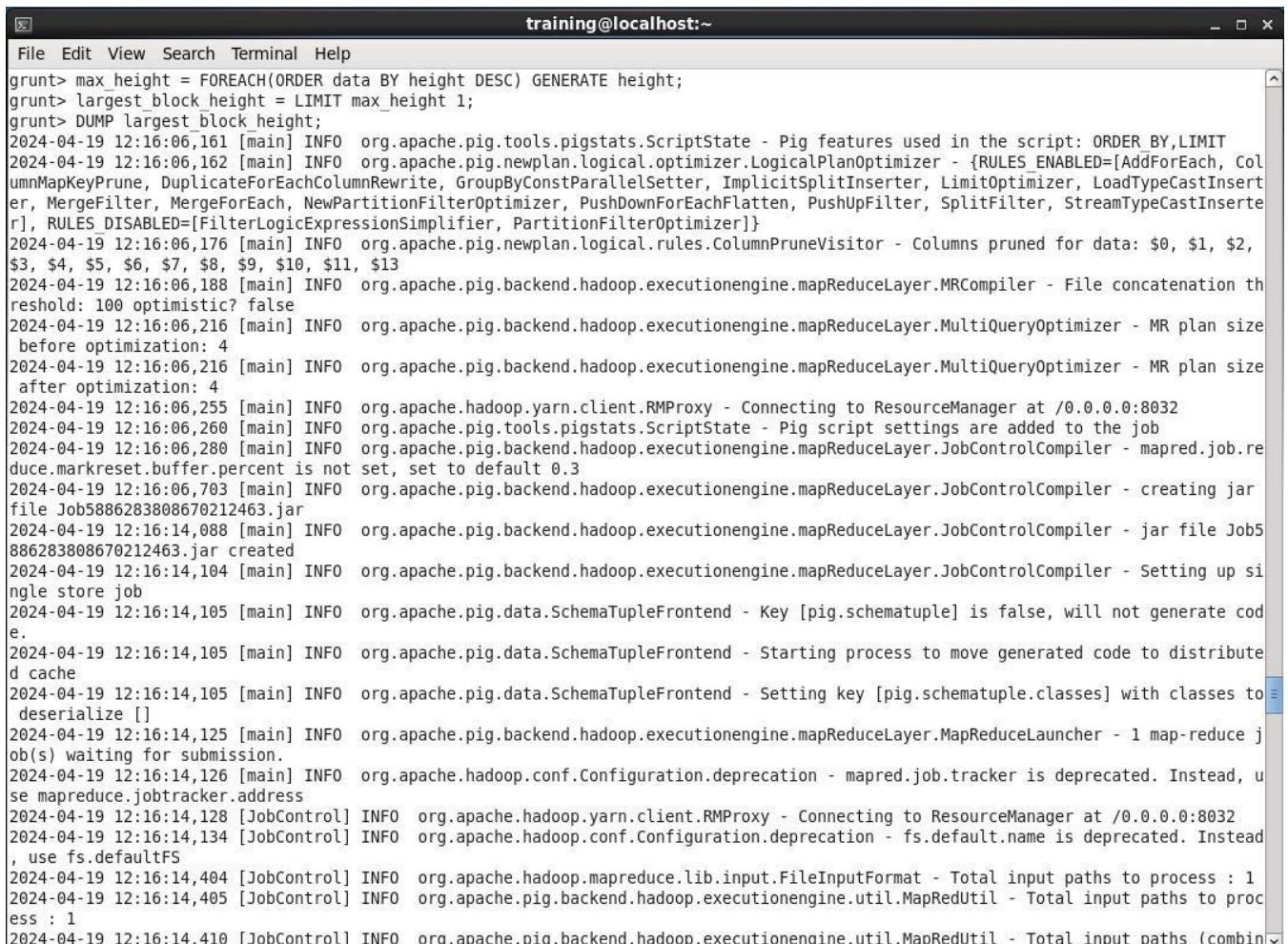
-- 2) Find the largest block height

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

```
largest_block_height = LIMIT max_height 1;
```

-- Output the result

```
DUMP largest_block_height;
```



The screenshot shows a terminal window titled "training@localhost:~". The window displays the execution of a Pig Latin script. The script starts with defining variables for finding the maximum height and the largest block height. It then prints the largest block height and finally dumps it. The terminal output shows various system logs and information from the Apache Pig and Hadoop frameworks, including INFO messages about script settings, job control, and map-reduce launcher. The logs indicate the script is running on a local host at port 8032.

```
File Edit View Search Terminal Help
training@localhost:~
File Edit View Search Terminal Help
grunt> max_height = FOREACH(ORDER data BY height DESC) GENERATE height;
grunt> largest_block_height = LIMIT max_height 1;
grunt> DUMP largest_block_height;
2024-04-19 12:16:06,161 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,LIMIT
2024-04-19 12:16:06,162 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInsert, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInsert], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-04-19 12:16:06,176 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for data: $0, $1, $2, $3, $4, $5, $6, $7, $8, $9, $10, $11, $13
2024-04-19 12:16:06,188 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-04-19 12:16:06,216 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2024-04-19 12:16:06,216 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 4
2024-04-19 12:16:06,255 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2024-04-19 12:16:06,260 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-04-19 12:16:06,280 [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 12:16:06,703 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job5886283808670212463.jar
2024-04-19 12:16:14,088 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job5886283808670212463.jar created
2024-04-19 12:16:14,104 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2024-04-19 12:16:14,105 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-04-19 12:16:14,105 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2024-04-19 12:16:14,105 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2024-04-19 12:16:14,125 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2024-04-19 12:16:14,126 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-04-19 12:16:14,128 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2024-04-19 12:16:14,134 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-04-19 12:16:14,404 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-04-19 12:16:14,405 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2024-04-19 12:16:14,410 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined)
```

C:\

training@acalhost:-

[] X

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 AvgReduceTime
e MedianReduceTime Alias Feature Outputs
job 1709496148644 0003 1 0 8 8 8 n/a n/a n/a data MAP ONLY
job-1709496148644-0004 1 1 7 7 7 8 8 8 1-26 SAMPLER
job-1709496148644-0005 1 1 8 8 8 9 9 9 1-26 ORDER BY.COMBINER
job-1709496148644-0006 1 1 7 7 7 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 bytes) 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 -> Job 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 tried 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 tried 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 tried 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. FinalApplicationStatus=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 tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)

```

C:\

training@acalhost:-

[] X

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 tried 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. FinalApplicationStatus=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 tried 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 tried 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 tried 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. FinalApplicationStatus=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 tried 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 tried 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 tried 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.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=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 tried 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 tried 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 tried 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. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:19:07,305 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD DISCARDED TYPE CONVERSION FAILED 1 time(s).
2024-04-19 12:19:07,305 [main] 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, use 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> I

```

-- 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

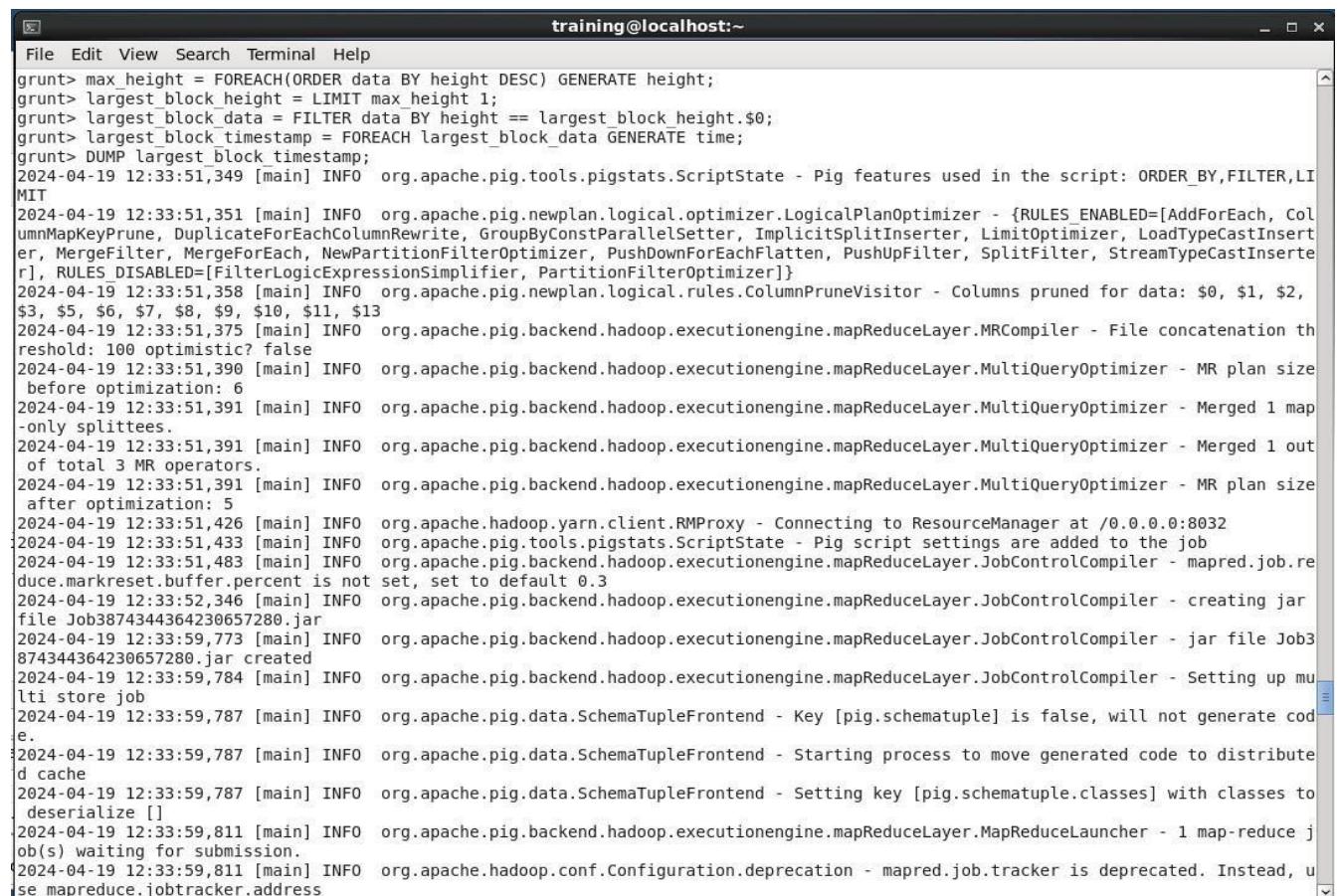
```
largest_block_data = FILTER data BY height == largest_block_height.$0;
```

-- Extract date and time for the largest block

```
largest_block_timestamp = FOREACH largest_block_data GENERATE time;
```

-- Output the result

```
DUMP largest_block_timestamp;
```



```
File Edit View Search Terminal Help
grunt> max_height = FOREACH(ORDER data BY height DESC) GENERATE height;
grunt> largest_block_height = LIMIT max_height 1;
grunt> largest_block_data = FILTER data BY height == largest_block_height.$0;
grunt> largest_block_timestamp = FOREACH largest_block_data GENERATE time;
grunt> DUMP largest_block_timestamp;
2024-04-19 12:33:51,349 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER,LIMIT
2024-04-19 12:33:51,351 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInsert, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInsert], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2024-04-19 12:33:51,358 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for data: $0, $1, $2, $3, $5, $6, $7, $8, $9, $10, $11, $13
2024-04-19 12:33:51,375 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-04-19 12:33:51,390 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 6
2024-04-19 12:33:51,391 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 1 map-only splittees.
2024-04-19 12:33:51,391 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged 1 out of total 3 MR operators.
2024-04-19 12:33:51,391 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 5
2024-04-19 12:33:51,426 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2024-04-19 12:33:51,433 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2024-04-19 12:33:51,483 [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 12:33:52,346 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job3874344364230657280.jar
2024-04-19 12:33:59,773 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job3874344364230657280.jar created
2024-04-19 12:33:59,784 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up multi store job
2024-04-19 12:33:59,787 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2024-04-19 12:33:59,787 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distribute cache
2024-04-19 12:33:59,787 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2024-04-19 12:33:59,811 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2024-04-19 12:33:59,811 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```

```

[17] 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 ORDER_BY,FILTER,LIMIT
Success!
Job Stats (time in seconds):
Jobid Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime
e MedianReduceTime Alias Feature Outputs
job 1709496148644 0007 1 0 8 8 8 n/a n/a n/a data MULTI QUERY,MAP ONLY
job-1709496148644-0008 1 1 7 7 7 8 8 8 1-68 SAMPLER -
job-1709496148644-0009 1 1 7 7 7 8 8 8 1-68 ORDER BY,COMBINER
job-1709496148644-0010 1 1 7 7 7 8 8 8 1-68,max height
job-1709496148644-0011 1 0 7 7 7 n/a n/a n/a largest block data.largest b
lock_timestamp 1-AP_ONLY hdfs://localhost:8020/tmp/temp-112855710/tmp-2010907623,
InputIsI:
Successfully read 1027 records (294515 bytes) from: u/user/training/output.csv
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 Job 1709496148644 0008, job 1709496148644 0009, -,
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 tried 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 tried 1 time(s); retry policy is R.retryUpToMaximumCountWithFixedSleep(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 tried 2 time(s); retry policy is R.retryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
 2024-04-19 12:37:15,418 [main] INFO org.apache.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED.

```

[17] 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 tried 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. FinalApplicationStatus=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 tried 0 time(s); 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 tried 1 time(s); 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 tried 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. FinalApplicationStatus=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 tried 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 tried 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 tried 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. FinalApplicationStatus=SUCCEEDED. 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 tried 0 time(s); 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 tried 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 tried 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. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:37:29,080 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning 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, use 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>I

```

The Current Epoch Unix Timestamp

Enter a Timestamp

Supports Unix timestamps in seconds, milliseconds, microseconds and nanoseconds.

Convert →

1713556021

SECONDS SINCE JAN 01 1970. (UTC)

2:47:16 PM

Copy

Format

Seconds

GMT

Wed Mar 20 2024 08:44:34 GMT+0000

Your Time Zone

Wed Mar 20 2024 03:44:34 GMT-0500 (Central Daylight Time)

Relative

a month ago

-- 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;
```

[f6]

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, LimitOptimizer, 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.MultiQueryOptimizer 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 Job244899998032650196.jar
2024-04-19 12:57:23,233 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler jar file Job244899998032650196.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 distribute
2024-04-19 12:57:23,250 [main] INFO org.apache.pig.data.SchemaTupleFrontend Setting key [pig.schematuple.classes] with classes to deserialize [1]

```

[f7]

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 12:57:15	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 block GROUP BV,fOMBINER
job_1709496148644_0013	1	1	7	7	7	7	7	7	1-113 SAMPLER
job_1709496148644-0014	1	1	8	8	8	8	8	8	1-113 ORDER BV,COMBINER
job_1709496148644-0015	1	1	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.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationstatus=SUCCEEDED. Redirecting to job history server

```

```
File Edit View Search Tenninal Help
ried 1 time(s);retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3,sleepTime=1000 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 RetryUpToMaximumCountWithFixedSleep(maxRetries=3,sleepTime=1000 MILLISECONDOS)
2024-04-19 13:00:13,870[main]INFO org.apache.hadoop.mapred.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=1000 MILLISECONDOS)
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 RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDOS)
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 RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDOS)
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=1000 MILLISECONDOS)
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 RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDOS)
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 RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDOS)
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=1000 MILLISECONDOS)
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 RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDOS)
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 RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDOS)
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(s).
2024-04-19 13:00:24,199 [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.tracker is 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> I
```

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;
```

Hue - Hive Editor - Query - Mozilla Firefox

localhost:8888/beeswax/execute/query/119#query/results

Most Visited ▾ Cloudera □ Hue □ YARN RM □ Spark UI (local) □ Spark Doc □ Solr Admin UI □ Kite SDK Doc

HUE Home Query Editors Data Browsers Workflows Search File Browser Job Browser training Open menu

Hive Editor **Query Editor** My Queries Saved Queries History

Assist Settings

DATABASE default

Table name... accounts_avro accounts_by_areacode device stock temp_stock webpage

```
1 select * from stock;
```

Execute Save as... Explain or create a New query

Recent queries Query Log Columns Results Chart

	stock.symbol	stock.date	stock.open	stock.high	stock.low	stock.close
0	GOOG	2024-03-18 09:30:00.0	149.37	150.47999999999999	149.34999999999999	150.16999999999999
1	GOOG	2024-03-18 09:31:00.0	150.18000000000001	150.97	149.84999999999999	150.93000000000001
2	GOOG	2024-03-18 09:32:00.0	150.96000000000001	151.25	150.49000000000001	151.185
3	GOOG	2024-03-18 09:33:00.0	151.16999999999999	151.50999999999999	150.97999999999999	151.40000000000001
4	GOOG	2024-03-18 09:34:00.0	151.38999999999999	152	151.00999999999999	151.83000000000001

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

SELECT COUNT(*) AS total_records FROM stock;

Hue - Hive Editor - Query - Mozilla Firefox

localhost:8888/beeswax/execute/query/120#query/results

Most Visited ▾ Cloudera □ Hue □ YARN RM □ Spark UI (local) □ Spark Doc □ Solr Admin UI □ Kite SDK Doc

HUE Home Query Editors Data Browsers Workflows Search File Browser Job Browser training Open menu

Hive Editor **Query Editor** My Queries Saved Queries History

Assist Settings

DATABASE default

Table name... accounts_avro accounts_by_areacode device stock temp_stock webpage

```
1 SELECT COUNT(*) AS total_records FROM stock;
```

Execute Save as... Explain or create a New query

Recent queries Query Log Columns Results Chart

	total_records
0	9776

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

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

The screenshot shows the Hue Hive Editor interface in a Mozilla Firefox browser window. The URL is `localhost:8888/beeswax/execute/query/121#query/results`. The query executed is:

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

The results table shows one row:

	number_of_different_days
0	5

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;
```

Hue - Hive Editor - Query - Mozilla Firefox

```
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;
```

Execute **Save as...** **Explain** or create a **New query**

	day	no_of_records_per_day
0	NULL	1
1	2024-03-18	1955
2	2024-03-19	1955
3	2024-03-20	1955
4	2024-03-21	1955
5	2024-03-22	1955

d. What are the symbols on the table?

SELECT DISTINCT symbol FROM stock;

Hue - Hive Editor - Query - Mozilla Firefox

```
SELECT DISTINCT symbol FROM stock;
```

Execute **Save as...** **Explain** or create a **New query**

	symbol
0	AAPL
1	AMZN
2	GOOG
3	IBM
4	MSFT
5	symbol

e. What is the highest price for each symbol?

```
SELECT symbol, MAX(high) AS highest_price
```

```
FROM stock
```

```
GROUP BY symbol;
```

The screenshot shows the Hue Hive Editor interface. In the top-left corner, it says "cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)". The title bar reads "Hue - Hive Editor - Query - Mozilla Firefox". The URL in the address bar is "localhost:8888/beeswax/execute/query/124#query/results". The main area shows a query editor with the following code:

```
1 SELECT symbol, MAX(high) AS highest_price
2 FROM stock
3 GROUP BY symbol;
4
5
```

Below the code are buttons for "Execute", "Save as...", "Explain", or create a "New query". The results section shows a table with the following data:

	symbol	highest_price
0	AAPL	178.68000000000001
1	AMZN	181.41499999999999
2	GOOG	152.93000000000001
3	IBM	194
4	MSFT	430.81999999999999
5	symbol	NULL

f. What is the lowest price for each symbol?

```
SELECT symbol, MIN(low) AS lowest_price
```

```
FROM stock
```

```
GROUP BY symbol;
```

The screenshot shows the Hue Hive Editor interface. In the top-left corner, there's a navigation bar with links to Cloudera, Hue, YARN RM, Spark UI (local), Spark Doc, Solr Admin UI, and Kite SDK Doc. Below the navigation bar is a search bar. The main area has tabs for Query Editors, Data Browsers, Workflows, and Search. The Query Editor tab is selected. On the left, there's a sidebar with 'Assist' and 'Settings' buttons, and a 'DATABASE' dropdown set to 'default'. Under 'Table name...', there's a list of tables: accounts_avro, accounts_by_areacode, device, stock, temp_stock, and webpage. The main query editor window contains the following SQL code:

```

1 SELECT symbol, MIN(low) AS lowest_price
2 FROM stock
3 GROUP BY symbol;
4
5
6

```

Below the code are buttons for 'Execute', 'Save as...', 'Explain', or create a 'New query'. The results section shows a table with two columns: 'symbol' and 'lowest_price'. The data is as follows:

symbol	lowest_price
AAPL	170.06
AMZN	173.52000000000001
GOOG	147.00999999999999
IBM	190.00999999999999
MSFT	413.77999999999997
symbol	NULL

g. What is the average price for each symbol?

`SELECT symbol, AVG(close) AS average_price`

`FROM stock`

`GROUP BY symbol;`

The screenshot shows the Hue Hive Editor interface, identical to the previous one but with a different query. The main query editor window contains the following SQL code:

```

1 SELECT symbol, AVG(close) AS average_price
2 FROM stock
3 GROUP BY symbol;
4
5
6

```

The results section shows a table with two columns: 'symbol' and 'average_price'. The data is as follows:

symbol	average_price
AAPL	174.54806649616415
AMZN	176.86847007672588
GOOG	149.36463836317154
IBM	192.19785984654732
MSFT	423.75076368286403
symbol	NULL

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

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

The screenshot shows the Hue Hive Editor interface. The top navigation bar includes links for Player, Applications, Places, System, and a search bar. Below the bar, the title 'Hue - Hive Editor - Query - Mozilla Firefox' is displayed, along with the date 'Sun Apr 21, 11:58 AM'. The main area has tabs for 'Hive Editor' and 'Query Editor', with 'Query Editor' selected. On the left, there's a sidebar for 'Assist' and 'Settings', and a 'DATABASE' dropdown set to 'default'. The central workspace contains the following SQL code:

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

Below the code are buttons for 'Execute', 'Save as...', 'Explain', and 'or create a New query'. The results section shows a table with two columns: 'symbol' and 'price_range'. The data is as follows:

	symbol	price_range
0	AAPL	8.620000000000045
1	AMZN	7.894999999999818
2	GOOG	5.920000000000159
3	IBM	3.990000000000091
4	MSFT	17.04000000000002
5	symbol	NULL

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.highest_price = mp.highest_price;
```

cloudera-trainig-capspark-student-rev_cdh5.4.3a VMwareWorkstation17 Player(Non-commercial use011ly)

Player • 151

Applications Places System 11 111 111

Sun Apr 21, 12:03 PM

Hue - Hive Editor•Query x ◊

MostVisitedv Cloude @ fltHue [:]YARNRM [:]spart::UI (local) [:]spart::Doc [:]SolrAdminUI [:]Kite SDK Doc

fltHue ti Query Editors Data Browsers Workflows Search

I File Browser raJobBrowser O training v O c

Hive Editor Query Editor My Queries Saved Queries History

Assist Settings

DATABASE

Table name...

flaccounts_avro mlaccounts_by_areacode fldevice flstock Intemp_stock mlwebpage

Saveas... Explain orcreatea Newquery

Recent queries Query Log Columns Results Change

```
WITH MaxPrices AS
R L c:yrnbol, MAX(high) AS highest_price
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;
```

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

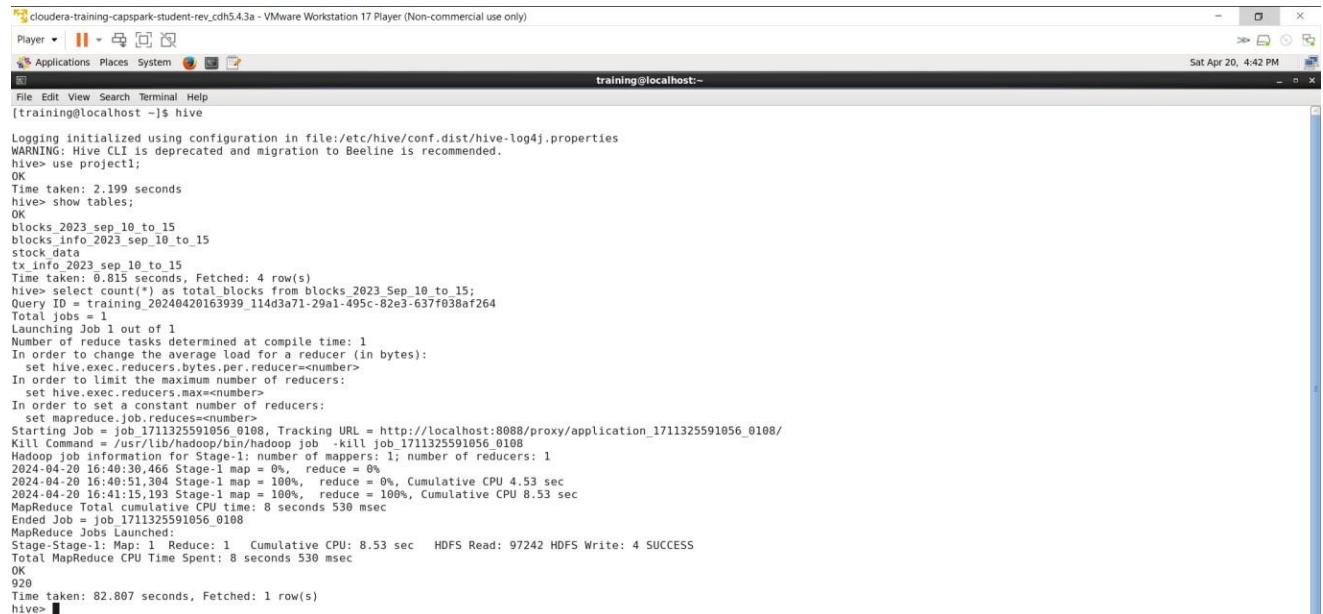
E t@ining@localhost:~ Hue - Hive Editor - Qu...

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?

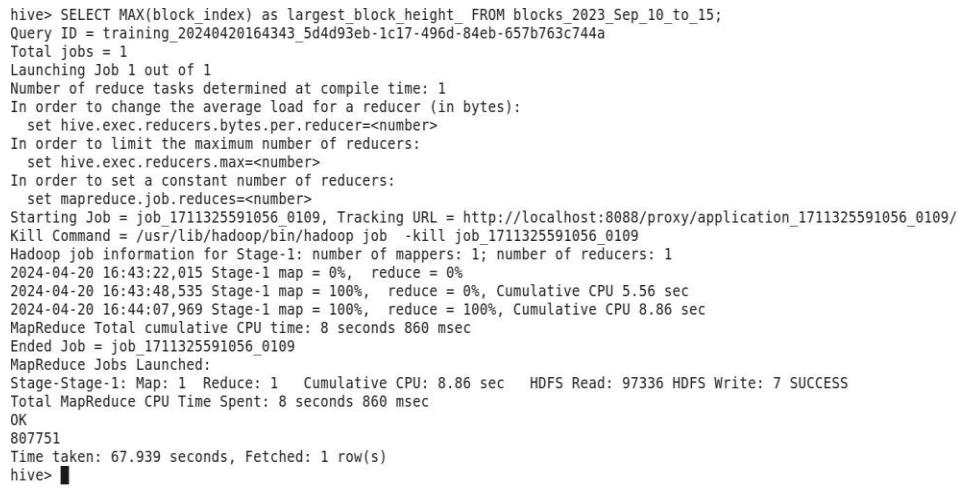
```
select count(*) as total_blocks from blocks_2023_Sep_10_to_15;
```



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player | || - & Applications Places System
File Edit View Search Terminal Help
[training@localhost ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> use project;
OK
Time taken: 2.199 seconds
hive> show tables;
OK
blocks_2023_sep_10_to_15
blocks_info_2023_sep_10_to_15
stock_data
tx_info_2023_sep_10_to_15
Time taken: 0.815 seconds, Fetched: 4 row(s)
hive> select count(*) as total_blocks from blocks_2023_Sep_10_to_15;
Query ID = training_20240420163939_114d3a71-29a1-495c-82e3-637f038af264
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1711325591056_0108, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0108/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1711325591056_0108
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-04-20 16:40:39,466 Stage-1: map = 0%, reduce = 0%
2024-04-20 16:40:51,304 Stage-1: map = 100%, reduce = 0%, Cumulative CPU 4.53 sec
2024-04-20 16:41:15,193 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.53 sec
MapReduce Total cumulative CPU time: 8 seconds 530 msec
Ended Job = job_1711325591056_0108
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.53 sec HDFS Read: 97242 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 530 msec
OK
920
Time taken: 82.807 seconds, Fetched: 1 row(s)
hive> ■
```

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;
```



```
hive> SELECT MAX(block_index) as largest_block_height_ FROM blocks_2023_Sep_10_to_15;
Query ID = training_20240420164343_5d4d93eb-1c17-496d-84eb-657b763c744a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1711325591056_0109, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0109/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1711325591056_0109
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-04-20 16:43:22,015 Stage-1 map = 0%, reduce = 0%
2024-04-20 16:43:48,535 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.56 sec
2024-04-20 16:44:07,969 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.86 sec
MapReduce Total cumulative CPU time: 8 seconds 860 msec
Ended Job = job_1711325591056_0109
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.86 sec HDFS Read: 97336 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 860 msec
OK
807751
Time taken: 67.939 seconds, Fetched: 1 row(s)
hive> ■
```

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_402389693779369930-1-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_402389693779369930-1-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  
> FROM tx_info_2023
```

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

```
Starting Job = job_1711325591056_0115, Tracking URL = http://localhost:8088/proxy/application_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
000000000000000000000000000000002dbf9d0bc1c743ac17bdb60d5c6abc8cd94f2d253621d 7252
Time taken: 236.629 seconds, Fetched: 1 row(s)
hive>
```

