

BIG DATA PROJECT-2

BUAN 6346.504

Group members:

Barath Kumar Dhnasekar-bxd220033

Jagadeep Nandagopal-jxn220044

Rahul Sadineni-rxs230051

Kosuri Durga Sravya-dxk220083

Nikitha Masineni-nxm230033

Navya Sahithi Surapaneni-nxs230031

Data Analysis:

A. Data Analysis using Pig

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

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

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

```
>> [training@localhost ~]$ pig -x mapreduce  
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).  
log4j:WARN Please initialize the log4j system properly.  
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.  
2024-04-19 11:52:12,746 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.3 (rexported) compiled Jun 24 2015, 19:36:38  
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, use  
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;

```

```

[main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
[main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitTop
erter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushDownFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
[main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
[main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.CombinerOptimizer - Choosing to have algebraic foreach to combiner
[main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1

Job DAG:
job_1713722356088_0012

2024-04-21 12:13:37,476 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Encountered Warning FIELD DISCARDED TYPE CONVERSION FAILED 2022 time(s).
2024-04-21 12:13:37,476 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Encountered Warning TOO LARGE FOR INT 400 time(s).
2024-04-21 12:13:37,476 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Success!
2024-04-21 12:13:37,476 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-04-21 12:13:37,476 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.job.tracker.address
2024-04-21 12:13:37,477 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-04-21 12:13:37,489 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-04-21 12:13:37,489 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
118871
grunt> █

```

1027

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

-- 2) Find the largest block height max_height = FOREACH (ORDER data BY height DESC) GENERATE height; largest_block_height = LIMIT max_height 1;

-- Output the result
DUMP largest_block_height;

```

File Edit View Search Terminal Help
ried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:55,115 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58171. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:55,232 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:18:57,797 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:51714. Already t
ried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:58,802 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:51714. Already t
ried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:59,806 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:51714. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:18:59,920 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:19:01,314 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58982. Already t
ried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:02,415 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58982. Already t
ried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:03,418 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:58982. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:03,530 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:19:04,972 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:60722. Already t
ried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:05,977 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:60722. Already t
ried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:06,982 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: localhost/127.0.0.1:60722. Already t
ried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=3, sleepTime=1000 MILLISECONDS)
2024-04-19 12:19:07,096 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicati
onStatus=SUCCEEDED. Redirecting to job history server
2024-04-19 12:19:07,305 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - Encountered Wa
rning FIELD DISCARDED TYPE CONVERSION FAILED 1 time(s).
2024-04-19 12:19:07,305 [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 - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-04-19 12:19:07,306 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.job.tracker.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> █

```

835509

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

-- 3) Find the largest block height top_block_date_time = FOREACH top_block_height GENERATE ToDate(time * 1000) AS date_time;

-- Output the result

DUMP top_block_date_time;

```
//
grunt> top_block_date_time = FOREACH top_block_height GENERATE ToDateTime * 1000 AS date_time;
2024-04-21 14:29:54.847 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> dump top_block

top_block_height    top_block_date_time
grunt> dump top_block

top_block_height    top_block_date_time
grunt> dump top_block_date_time;
2024-04-21 14:30:41.939 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
2024-04-21 14:30:41.939 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY_LIMIT
2024-04-21 14:30:41.944 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - [RULES_ENABLED]=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnWrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitTop
Timmer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushDownFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionsSimplifier, PartitionFilterOptimizer]
}

2024-04-21 14:32:20.039 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-04-21 14:32:20.039 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-04-21 14:32:20.039 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-04-21 14:32:20.040 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-04-21 14:32:20.042 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-04-21 14:32:20.042 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2024-02-24T22:04:56.000-08:00)
grunt> █
```

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

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

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

-- Output the result

DUMP top_tx_count;

```
grunt> max_tx_count = FOREACH data GENERATE n_tx;
2024-04-21 14:33:05.192 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> ordered_tx_count = ORDER max_tx_count BY n_tx DESC;
2024-04-21 14:33:05.237 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> top_tx_count = LIMIT ordered_tx_count 1;
2024-04-21 14:33:05.261 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> dump top_tx_count

top_tx_count    top_block_height    top_block_date_time
grunt> dump top_tx_count;
2024-04-21 14:33:24.104 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY_LIMIT
2024-04-21 14:33:24.111 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - [RULES_ENABLED]=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnWrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitTop
Timmer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushDownFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionsSimplifier, PartitionFilterOptimizer]
}

2024-04-21 14:33:24.112 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for data: $0, $1, $2, $3, $4, $5, $6, $7, $8, $10, $11, $12, $13
2024-04-21 14:33:24.117 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - File concatenation threshold: 100 splitSize? false

2024-04-21 14:37:13.471 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 1 time(s).
2024-04-21 14:37:13.471 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning TOO_LARGE_FOR_INT 490 time(s).
2024-04-21 14:37:13.471 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-04-21 14:37:13.471 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-04-21 14:37:13.471 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-04-21 14:37:13.472 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-04-21 14:37:13.474 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-04-21 14:37:13.474 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2141853579)
grunt> █
```

2141853579

B. Data Analysis using Hive – Part 1

```
CREATE TABLE stock_market (  
    stock STRING,  
    timestamp TIMESTAMP,  
    open DECIMAL(10,4),  
    high DECIMAL(10,4),  
    low DECIMAL(10,4),  
    close DECIMAL(10,4),  
    volume BIGINT  
)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
STORED AS TEXTFILE;  
  
LOAD DATA INPATH 'hdfs://localhost:8020/flume/data/stock_market_data.txt'  
INTO TABLE stock_market;  
  
SELECT * FROM stock_market;
```

The screenshot shows the Hue web interface running on a VM. The 'Query Editor' tab is active, displaying the query `SELECT * FROM stock_market;`. Below the editor, the 'Results' tab shows the output of the query. The results are displayed in a table with 8 columns: `stock_market.stock`, `stock_market.timestamp`, `stock_market.open`, `stock_market.high`, `stock_market.low`, `stock_market.close`, and `stock_market.volume`. The table contains 10 rows of data for AAPL stock on 2024-03-18.

	stock_market.stock	stock_market.timestamp	stock_market.open	stock_market.high	stock_market.low	stock_market.close	stock_market.volume
3	AAPL	2024-03-18 09:31:00.0	175.425	175.77	175.2	175.75	498142
4	AAPL	2024-03-18 09:32:00.0	175.76	175.91	175.54	175.77	566280
5	AAPL	2024-03-18 09:33:00.0	175.76	176.168	175.68	175.975	720968
6	AAPL	2024-03-18 09:34:00.0	175.972	176.05	175.64	175.86	427842
7	AAPL	2024-03-18 09:35:00.0	175.87	176.34	175.82	176.33	692365
8	AAPL	2024-03-18 09:36:00.0	176.33	176.4	175.88	176.019	684740
9	AAPL	2024-03-18 09:37:00.0	176.001	176.19	175.96	176.005	455199
10	AAPL	2024-03-18 09:38:00.0	176.01	176.29	176.01	176.24	387677

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

The screenshot shows the Hive Editor interface. On the left, there is a sidebar with 'Assist' and 'Settings' tabs. Below 'Assist', there is a 'DATABASE' dropdown set to 'default' and a 'Table name...' input field. A list of tables is visible: accounts_avro, accounts_by_areac..., device, new_accounts_parg..., sample_07, sample_08, stock_market, stocksy, and webpage. The main area contains a query editor with the following SQL query:

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

Below the query editor, there are buttons: 'Execute', 'Save', 'Save as...', 'Explain', 'or create a', and 'New query'. The 'Execute' button is highlighted. Below the buttons, there is a tabbed interface with 'Recent queries', 'Query', 'Log', 'Columns', 'Results', and 'Chart'. The 'Results' tab is active, showing a table with one column 'total_records' and one row with the value 9848.

total_records
9848

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

The screenshot shows the Hive Editor interface. On the left, there is a sidebar with 'Assist' and 'Settings' tabs. Below 'Assist', there is a 'DATABASE' dropdown set to 'default' and a 'Table name...' input field. A list of tables is visible: accounts_avro, accounts_by_areac..., device, new_accounts_parg..., sample_07, sample_08, stock_market, stocksy, and webpage. The main area contains a query editor with the following SQL query:

```
1 SELECT COUNT(DISTINCT CAST(timestamp AS DATE)) AS no_of_different_days FROM stock_market;
```

Below the query editor, there are buttons: 'Execute', 'Save as...', 'Explain', 'or create a', and 'New query'. The 'Execute' button is highlighted. Below the buttons, there is a tabbed interface with 'Recent queries', 'Query', 'Log', 'Columns', 'Results', and 'Chart'. The 'Results' tab is active, showing a table with one column 'no_of_different_days' and one row with the value 5.

no_of_different_days
5

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

cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)

Player ▾ | [Icons] | [Icons]

HUE | Query Editors ▾ | Data Browsers ▾ | Workflows ▾ | Search | File Browser | Job Browser | Barath0103 | [Icons]

Hive Editor | Query Editor | My Queries | Saved Queries | History

Assist | Settings

DATABASE

default ▾

Table name...

- accounts_avro
- accounts_by_areac...
- device
- new_accounts_parq...
- sample_07
- sample_08
- stock_market
- stocksy
- webpage

```

1 SELECT CAST(timestamp AS DATE) AS day, COUNT(*) AS no_of_records_per_day
2 FROM stock_market
3 GROUP BY CAST(timestamp AS DATE)
4 ORDER BY day;
5

```

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

Recent queries Query Log Columns Results Chart

	day	no_of_records_per_day
0	NULL	98
1	2024-03-18	1950
2	2024-03-19	1950
3	2024-03-20	1950
4	2024-03-21	1950
5	2024-03-22	1950

D. SELECT DISTINCT stock FROM stock_market;

HUE | Query Editors ▾ | Data Browsers ▾ | Workflows ▾ | Search | File Browser | Job Browser | Barath0103 | [Icons]

Hive Editor | Query Editor | My Queries | Saved Queries | History

Assist | Settings

DATABASE

default ▾

Table name...

```

1 SELECT DISTINCT stock
2 FROM stock_market;
3

```

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

Recent queries Query Log Columns Results Chart

	stock
0	
1	AAPL
2	GOOGL
3	IBM
4	MSFT
5	TSLA
6	stock

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

The screenshot shows the Hive Editor interface. On the left, there's a sidebar with 'Assist' and 'Settings' tabs. Under 'Assist', there's a 'DATABASE' section with a dropdown set to 'default' and a list of tables including 'accounts_avro', 'accounts_by_areac...', 'device', 'new_accounts_parq...', 'sample_07', 'sample_08', 'stock_market', 'stocksy', and 'webpage'. The main area displays a SQL query in a text editor:

```
1 SELECT STOCK, MAX(high) AS highest_price
2 FROM stock_market group by stock;
3
```

Below the query editor are buttons for 'Execute', 'Save as...', 'Explain', 'or create a', and 'New query'. The 'Execute' button is highlighted. Below the query editor, there's a tabbed interface with 'Recent queries', 'Query', 'Log', 'Columns', 'Results', and 'Chart'. The 'Results' tab is active, showing a table with two columns: 'stock' and 'highest_price'.

	stock	highest_price
0		NULL
1	AAPL	178.67
2	GOOGL	152.15
3	IBM	193.98
4	MSFT	430.82
5	TSLA	178.18
6	stock	NULL

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

The screenshot shows the Hive Editor interface. On the left, there's a sidebar with 'Assist' and 'Settings' tabs. Under 'Assist', there's a 'DATABASE' section with a dropdown set to 'default' and a list of tables including 'accounts_avro', 'accounts_by_areac...', 'device', 'new_accounts_parq...', 'sample_07', 'sample_08', 'stock_market', 'stocksy', and 'webpage'. The main area displays a SQL query in a text editor:

```
1 SELECT STOCK, MIN(high) AS lowest_price
2 FROM stock_market group by stock;
3
```

Below the query editor are buttons for 'Execute', 'Save as...', 'Explain', 'or create a', and 'New query'. The 'Execute' button is highlighted. Below the query editor, there's a tabbed interface with 'Recent queries', 'Query', 'Log', 'Columns', 'Results', and 'Chart'. The 'Results' tab is active, showing a table with two columns: 'stock' and 'lowest_price'.

	stock	lowest_price
0		NULL
1	AAPL	170.268
2	GOOGL	146.17
3	IBM	190.31
4	MSFT	415.571
5	TSLA	166.3
6	stock	NULL

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

The screenshot shows the Hive Editor interface. On the left, there's a sidebar with 'Assist' and 'Settings' tabs. Below 'Assist', there's a 'DATABASE' section with a dropdown set to 'default' and a list of tables including 'accounts_avro', 'accounts_by_areac...', 'device', 'new_accounts_parg...', 'sample_07', 'sample_08', 'stock_market', 'stocksy', and 'webpage'. The main area displays a SQL query in a text editor:

```
1 SELECT STOCK, AVG(CLOSE) AS average_price
2 FROM stock_market group by stock;
```

Below the query editor are buttons: 'Execute', 'Save as...', 'Explain', 'or create a', and 'New query'. The 'Execute' button is highlighted. Below the query editor, there's a tabbed interface with 'Recent queries', 'Query', 'Log', 'Columns', 'Results', and 'Chart'. The 'Results' tab is active, showing a table with two columns: 'stock' and 'average_price'. The table contains 7 rows of data:

stock	average_price
	NULL
AAPL	174.54863282
GOOGL	148.42665282
IBM	192.19748359
MSFT	423.74882615
TSLA	172.12119128
stock	NULL

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

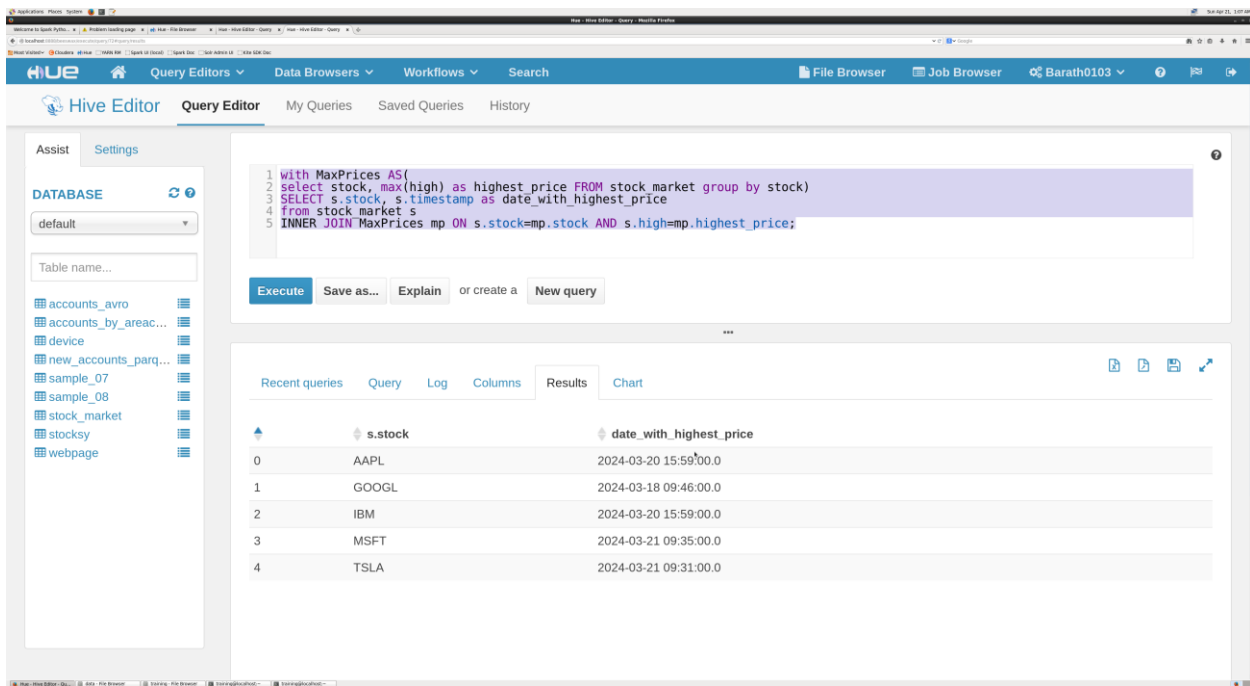
The screenshot shows the Hive Editor interface. On the left, there's a sidebar with 'Assist' and 'Settings' tabs. Below 'Assist', there's a 'DATABASE' section with a dropdown set to 'default' and a list of tables including 'accounts_avro', 'accounts_by_areac...', 'device', 'new_accounts_parg...', 'sample_07', 'sample_08', 'stock_market', 'stocksy', and 'webpage'. The main area displays a SQL query in a text editor:

```
1 SELECT STOCK, MAX(high)-MIN(low) AS price_range
2 FROM stock_market group by stock;
```

Below the query editor are buttons: 'Execute', 'Save as...', 'Explain', 'or create a', and 'New query'. The 'Execute' button is highlighted. Below the query editor, there's a tabbed interface with 'Recent queries', 'Query', 'Log', 'Columns', 'Results', and 'Chart'. The 'Results' tab is active, showing a table with two columns: 'stock' and 'price_range'. The table contains 7 rows of data:

stock	price_range
	NULL
AAPL	8.61
GOOGL	6.07
IBM	3.97
MSFT	17.04
TSLA	12.28
stock	NULL

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



C. Data Analysis using Hive – Part 2

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

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

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

```
[training@localhost ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> use project1;
OK
Time taken: 2.199 seconds
hive> show tables;
OK
blocks_2023_sep_10_to_15
blocks_info_2023_sep_10_to_15
stock_data
tx_info_2023_sep_10_to_15
Time taken: 0.015 seconds, Fetched: 4 row(s)
hive> select count(*) as total_blocks from blocks_2023_sep_10_to_15;
Query ID = training_20240420163939_114d3a71-29a1-495c-82e3-637f038af264
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1711325591056_0108, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0108/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1711325591056_0108
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-04-20 16:40:30,466 Stage-1 map = 0%, reduce = 0%
2024-04-20 16:40:51,384 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.53 sec
2024-04-20 16:41:15,193 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.53 sec
MapReduce Total cumulative CPU time: 8 seconds 530 msec
Ended Job = job_1711325591056_0108
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.53 sec HDFS Read: 97242 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 530 msec
OK
920
Time taken: 82.807 seconds, Fetched: 1 row(s)
hive>
```

920

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

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

```

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>

```

807751

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

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

```

Query ID = training_20240420164848_39b3121a-5e1e-4ac3-922f-17e37dfc1fc7
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1711325591056_0112, Tracking URL = http://localhost:8088/proxy/application_1711325591056_0112/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1711325591056_0112
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-04-20 16:49:19,741 Stage-1 map = 0%, reduce = 0%
2024-04-20 16:49:40,136 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.78 sec
2024-04-20 16:49:57,909 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.9 sec
MapReduce Total cumulative CPU time: 8 seconds 900 msec
Ended Job = job_1711325591056_0112
Execution log at: /tmp/training/training_20240420164848_39b3121a-5e1e-4ac3-922f-17e37dfc1fc7.log
2024-04-20 04:50:09 Starting to launch local task to process map join; maximum memory = 1013645312
2024-04-20 04:50:14 Dump the side-table for tag: 0 with group count: 920 into file: file:/tmp/training/45d2901d-f8e3-40d1-8179-3779a9f05131/hive_2024-04-20_16-48-58_814_4023896937779369930-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile10--.hashtable
2024-04-20 04:50:14 Uploaded 1 File to: file:/tmp/training/45d2901d-f8e3-40d1-8179-3779a9f05131/hive_2024-04-20_16-48-58_814_4023896937779369930-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile10--.hashtable (40938 bytes)
2024-04-20 04:50:14 End of local task; Time Taken: 4.518 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 2 out of 2
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 16:58:33,390 Stage-4 map = 0%, reduce = 0%
2024-04-20 16:58:49,822 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>

```

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

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

```

hive> SELECT block_hash, COUNT(tx_hash) as num_transactions
> FROM tx_info 2023_Sep_10_to_15
> GROUP BY block_hash
> ORDER BY num_transactions
> DESC LIMIT 1;
Query ID = training_20240420165252_e2e70a65-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
0000000000000000000000002dbf9d0b1c743ac17bdb60d5c6abc8cd94f2d253621d 7252
Time taken: 236.629 seconds, Fetched: 1 row(s)
hive> █

```