REAL-TIME FINANCIAL MARKET DATA INGESTION CASE STUDY

In this case study, we will create a Nifi flow to ingest Financial Market Data from IEX Cloud using their REST API:

- Split returned JSON data into three main components (news, chart and quote),
- Store symbols charts on HDFS as Parquet file(s),
- Create a Hive/Impala table to read and generate a basic summary of the data
- Show the output results as graph in a Zeppelin note whenever it is possible.

TASK 01

- Refer to stock_data_schema.txt
- Copy of Schema:

```
{
 "type" : "record",
 "name" : "stock_data",
 "namespace" : "nifi",
 "fields" : [
 {
  "name" : "Symbol",
  "type" : "string"
 },
 {
  "name" : "CompanyName",
  "type" : "string"
 },
 {
  "name" : "Date",
  "type" : "string"
```

```
  },
  {
    "name" : "Open",
    "type" : "double"
  },
  {
    "name" : "Close",
    "type" : "double"
  },
  {
    "name" : "High",
    "type" : "double"
  },
  {
    "name" : "Low",
    "type" : "double"
  },
  {
    "name" : "Volume",
    "type" : "long"
  }]
}
```

TASK 02

- **Processors used and properties updated:**
  - *Getfile*
    - Input Directory = /home/cloudera/Downloads/
  - *UpdateAttribute*
    - Filename = ${UUID()}
    - Mime.type = application/json

TASK 03

- **Processors used and properties updated:**
  - *SplitJson*
    - JsonPath Expression = $.*.quote
  - *EvaluateJsonPath*
    - Destination = flowfile-attribute
    - Open = $.open
    - Symbol = $.symbol
  - *RouteOnAttribute*
    - *Routing Strategy = Route to Property Name*
    - *Empty = ${Open:isEmpty()}*
    - *NotEmpty = ${Open:isEmpty():not()}*
  - *PutFile*
    - Directory = /home/cloudera/market/quote/${Symbol}
    - Conflict Resolution Strategy = replace
    - Create Missing Directories = true
  - *PutHDFS*
    - Hadoop Configuration Resources = /etc/hadoop/conf/core-site.xml, /etc/hadoop/conf/hdfs-site.xml
    - Directory = /market/quote/${Symbol}
    - Conflict Resolution Strategy = replace

TASK 04

- **Processors used and properties updated:**
  - *SplitJson*
    - JsonPath Expression = $.*
  - *EvaluateJsonPath*
    - Destination = flowfile-attribute
    - Symbol = $.quote.symbol
  - *SplitJson*
    - JsonPath Expression = $.news
  - *PutHDFS*
    - Hadoop Configuration Resources = /etc/hadoop/conf/core-site.xml, /etc/hadoop/conf/hdfs-site.xml

        - Directory = /market/news/${Symbol}
        - Conflict Resolution Strategy = append

TASK 05

- **Processors used and properties updated:**
    - *SplitJson*
        - JsonPath Expression = $.*
    - *EvaluateJsonPath*
        - Destination = flowfile-attribute
        - CompanyName = $.quote.companyName
        - Symbol = $.quote.symbol
    - *SplitJson*
        - JsonPath Expression = $.chart
    - *EvaluateJsonPath*
        - Destination = flowfile-attribute
        - Close = $.close
        - Date = $.date
        - High = $.high
        - Low = $.low
        - Open = $.open
        - Volume = $.volume
    - AttributesToJson
        - Attributes List = Symbol, CompanyName, Date,Open,Close,High,Low,Volume
        - Destination = flowfile-content
    - MergeContent
        - Maximum Number of Entries = 10000
    - PutParquet
        - Hadoop Configuration Resources = /etc/hadoop/conf/core-site.xml, /etc/hadoop/conf/hdfs-site.xml
        - Record Reader = JsonTreeReader
        - Directory = /market/chart/
        - Compression Type = Snappy
        - JsonTreeReader Service Details:
            - Schema Access Strategy = Use “Schema Text” property
            - Schema Text = (pasted the schema from task 01)

TASK 06

```
%hive
--Use as needed
DROP TABLE market
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at June 01 2020, 3:22:46 PM.

FINISHED

```
%hive
--Use as needed
drop database MarketDB
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at June 01 2020, 3:22:48 PM.

FINISHED

```
%hive
--#### TASK 06 ######
-- Create Database

CREATE DATABASE MarketDB
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at June 01 2020, 3:22:50 PM. (outdated)

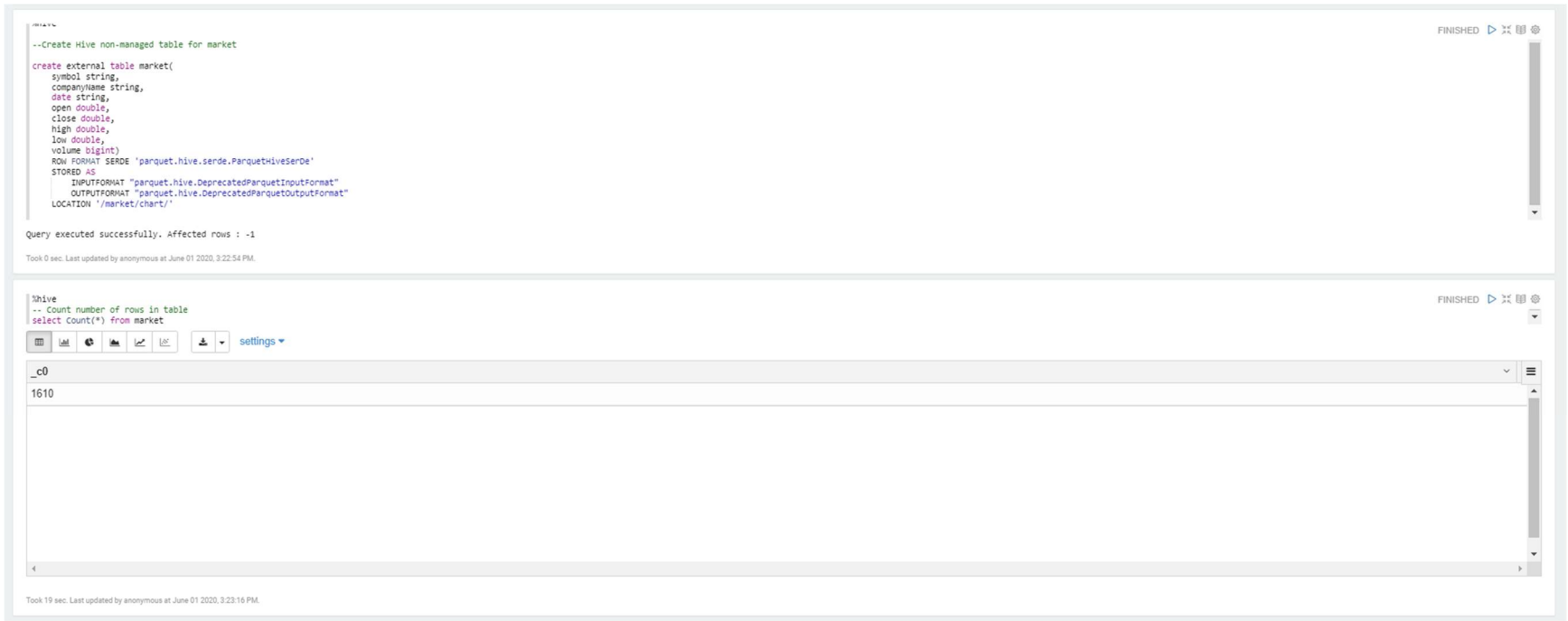
FINISHED

```
%hive
-- Use database
use MarketDB
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at June 01 2020, 3:22:52 PM.

FINISHED



```
%hive

--Create Hive non-managed table for market

create external table market(
    symbol string,
    companyName string,
    date string,
    open double,
    close double,
    high double,
    low double,
    volume bigint)
    ROW FORMAT SERDE 'parquet.hive.serde.ParquetHiveSerDe'
    STORED AS
        INPUTFORMAT "parquet.hive.DeprecatedParquetInputFormat"
        OUTPUTFORMAT "parquet.hive.DeprecatedParquetOutputFormat"
    LOCATION '/market/chart/'
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at June 01 2020, 3:22:54 PM.

FINISHED

```
%hive
-- Count number of rows in table
select Count(*) from market
```

settings

_c0
1610

Took 19 sec. Last updated by anonymous at June 01 2020, 3:23:16 PM.

FINISHED

```
%hive
--Display table properties and size
show create table market
```

FINISHED

settings

createtab_stmt
'hdfs://quickstart.cloudera:8020/market/chart'
TBLPROPERTIES (
'COLUMN_STATS_ACCURATE'='false',
'numFiles'='1',
'numRows'='-1',
'rawDataSize'='-1',
'totalSize'='42981',
'transient_lastDdlTime'='1591039374')

Took 0 sec. Last updated by anonymous at June 01 2020, 3:24:05 PM.

```
%sh
# Size of table on HDFS
hdfs dfs -ls '/market/chart/' #to match statement above
hdfs dfs -ls -h '/market/chart/'
```

FINISHED

```
Found 1 items
-rw-r--r--   1 root supergroup      42981 2020-06-01 12:03 /market/chart/0fbe789c-8c74-4ffa-b757-721b4d6d5d98
Found 1 items
-rw-r--r--   1 root supergroup     42.0 K 2020-06-01 12:03 /market/chart/0fbe789c-8c74-4ffa-b757-721b4d6d5d98
```

Took 4 sec. Last updated by anonymous at June 01 2020, 3:24:11 PM.

```
%hive
select * from market limit 5
```

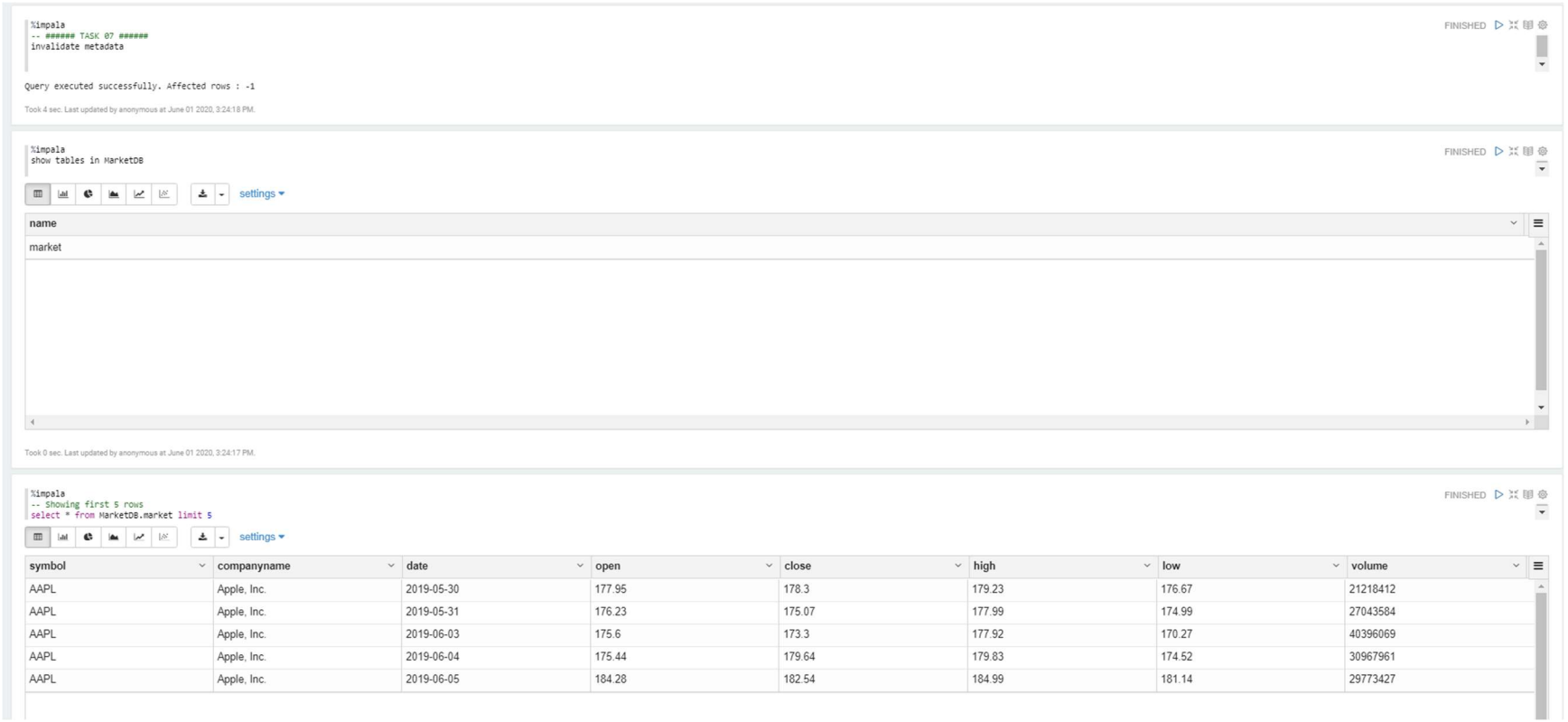
FINISHED

settings

market.symbol	market.companyname	market.date	market.open	market.close	market.high	market.low	market.volume
AAPL	Apple, Inc.	2019-05-30	177.95	178.3	179.23	176.67	21218412
AAPL	Apple, Inc.	2019-05-31	176.23	175.07	177.99	174.99	27043584
AAPL	Apple, Inc.	2019-06-03	175.6	173.3	177.92	170.27	40396069
AAPL	Apple, Inc.	2019-06-04	175.44	179.64	179.83	174.52	30967961
AAPL	Apple, Inc.	2019-06-05	184.28	182.54	184.99	181.14	29773427

Took 0 sec. Last updated by anonymous at June 01 2020, 3:24:12 PM.

TASK 07



```
%impala
-- ###### TASK 07 ######
invalidate metadata
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 4 sec. Last updated by anonymous at June 01 2020, 3:24:18 PM.

```
%impala
show tables in MarketDB
```

FINISHED

settings

name
market

Took 0 sec. Last updated by anonymous at June 01 2020, 3:24:17 PM.

```
%impala
-- Showing first 5 rows
select * from MarketDB.market limit 5
```

FINISHED

settings

symbol	companyname	date	open	close	high	low	volume
AAPL	Apple, Inc.	2019-05-30	177.95	178.3	179.23	176.67	21218412
AAPL	Apple, Inc.	2019-05-31	176.23	175.07	177.99	174.99	27043584
AAPL	Apple, Inc.	2019-06-03	175.6	173.3	177.92	170.27	40396069
AAPL	Apple, Inc.	2019-06-04	175.44	179.64	179.83	174.52	30967961
AAPL	Apple, Inc.	2019-06-05	184.28	182.54	184.99	181.14	29773427

```
%impala
-- Count number of chart for each symbol
SELECT symbol, count(symbol) as NumberChart FROM MarketDB.market GROUP BY symbol
```

FINISHED

settings

symbol	numberchart
AAPL	253
GS	253
RHT	92
MSFT	253
FB	253
SBUX	253
CLDR	253

Took 0 sec. Last updated by anonymous at June 01 2020, 3:24:29 PM.

```
%impala
SELECT symbol, min(open) as MinOpen, max(open) as MaxOpen, min(high) as MinHigh, max(high) as MaxHigh, min(low) as MinLow, max(low) as MaxLowFROM, min(volume) as MinVolmue, max(volume) as MaxVolume from MarketDB.market GROUP BY symbol
```

FINISHED

settings

symbol	minopen	maxopen	minhigh	maxhigh	minlow	maxlowfrom	minvolmue	maxvolume
AAPL	175.44	324.74	177.92	327.85	170.27	323.35	11362045	106721230
GS	136.03	250.23	141.94	250.46	130.85	248	467722	9543242
RHT	0.0	188.4	0.0	189.14	0.0	188.05	0	4591542
MSFT	121.28	190.65	123.28	190.7	119.01	186.47	8989150	97073557
FB	139.75	239.77	148.18	240.9	137.1	231.67	6046273	56059609
SBUX	55.55	98.14	57.44	99.72	50.02	97.21	1847770	28770619
CLDR	5.06	12.1	5.14	12.22	4.76	11.9	1446493	57916618

Took 0 sec. Last updated by anonymous at June 01 2020, 8:47:44 PM. (outdated)

```
%impala
-- Showing the min and max value for the open price for each symbol

SELECT symbol, min(open) as MinOpen, max(open) as MaxOpen FROM MarketDB.market GROUP BY symbol
```

FINISHED

settings

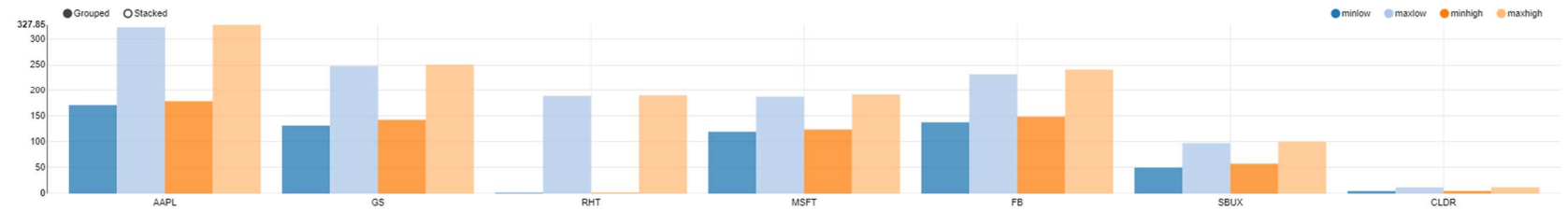
Took 0 sec. Last updated by anonymous at June 01 2020, 3:24:32 PM.

```
%impala

-- Showing the min and max of the low and high for each symbol
SELECT symbol, min(high) as MinHigh, max(high) as MaxHigh, min(low) as MinLow, max(low) as MaxLow FROM MarketDB.market GROUP BY symbol
```

FINISHED

settings



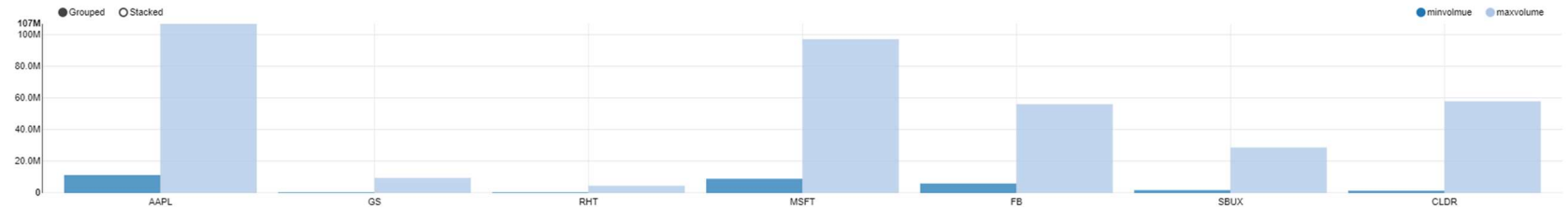
Took 0 sec. Last updated by anonymous at June 01 2020, 3:24:35 PM.

```
%impala

-- Showing min and max values for the volume for each symbol
SELECT symbol, min(volume) as MinVolmue, max(volume) as MaxVolume  FROM MarketDB.market GROUP BY symbol
```

FINISHED

settings



Took 0 sec. Last updated by anonymous at June 01 2020, 3:24:37 PM.