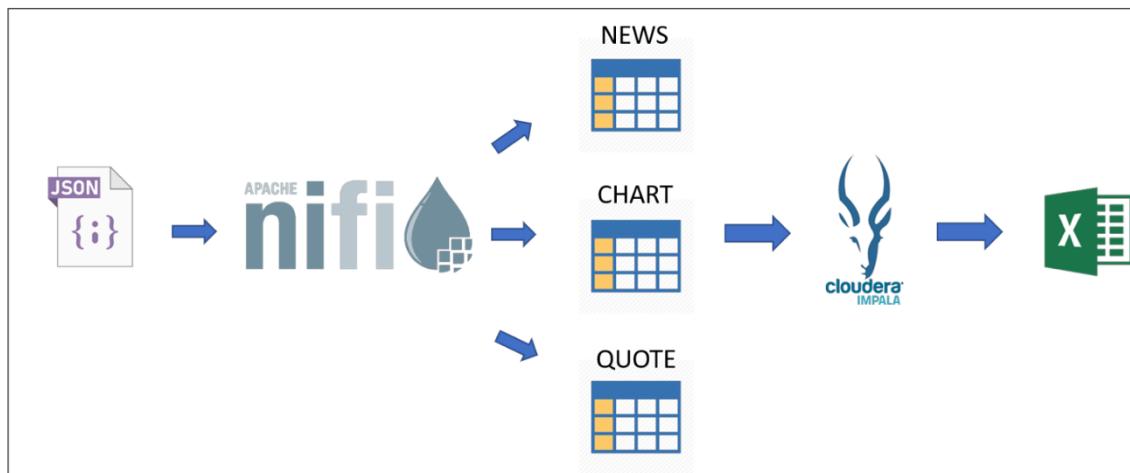


Assignment 2

Real-Time Financial Market Data Ingestion



November 23, 2021

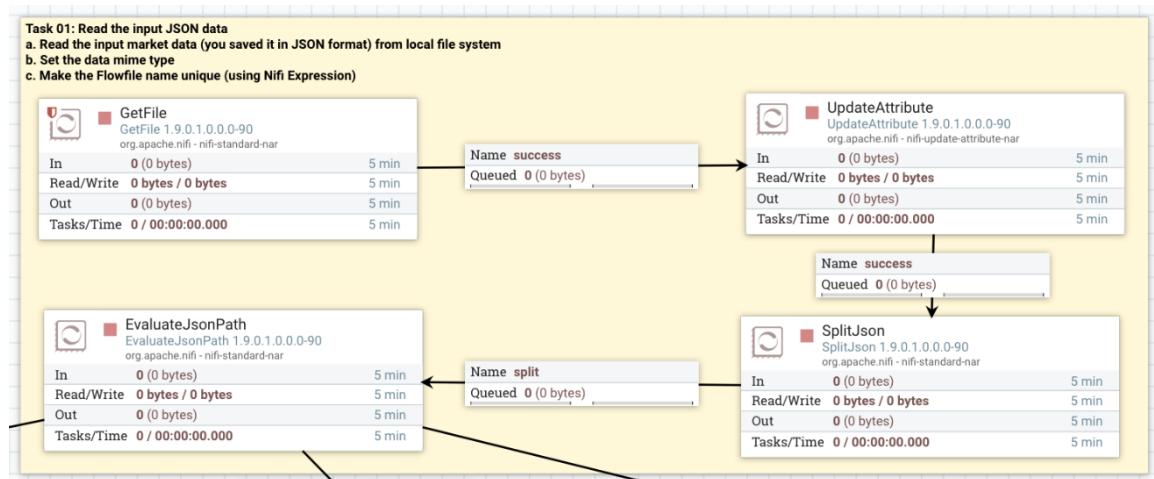
Team #5

Nazanin Hashemipoor
Sofienne Srihi
Nasim Afzali Chali
Marie-Noël Lepage
Benoit Tessier

In our team, all of us were working and participating. We all worked on code on our own and shared our progress with each other to solve issues and obtain a final working code. In parallel we produced and review the documentation to supplement the exercise.

TASK 01

Read the input JSON data



- Read the input market data (you saved it in JSON format) from your local file system.

We downloaded our data from IEX Cloud Platform as mentioned on the assignment and we save it in Json format file, then we put it in HDFS to start working with it. The credit limitation allowed us to work with 3 months of data.

We put the file in our local system at this directory: /home/cloudera/Downloads/nifi

```
[cloudera@quickstart ~]$ cd Downloads
[cloudera@quickstart Downloads]$ cd nifi
[cloudera@quickstart nifi]$ ll
total 260
-rw----- 1 cloudera cloudera 262963 Nov 21 11:54 big_data_3m_100.json
[cloudera@quickstart nifi]$
```

| Processor | Properties | |
|-----------|------------------|-------------------------------|
| GetFile | Directory | /home/cloudera/Downloads/nifi |
| | Keep Source File | true |



Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|------------------------|------------|-------------------------------|-------------------|
| Required field | | | + |
| Property | | | Value |
| Input Directory | ? | /home/cloudera/Downloads/nifi | |
| File Filter | ? | [^\\.]* | |
| Path Filter | ? | No value set | |
| Batch Size | ? | 10 | |
| Keep Source File | ? | true | |
| Recurse Subdirectories | ? | true | |
| Polling Interval | ? | 0 sec | |
| Ignore Hidden Files | ? | true | |
| Minimum File Age | ? | 0 sec | |
| Maximum File Age | ? | No value set | |
| Minimum File Size | ? | 0 B | |
| Maximum File Size | ? | No value set | |

We use scheduling at this step to take the data each hour.
For all other step we put 0 sec to run the process in real time.

Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|---------------------------------------|--------------|--------------------------------|----------|
| Scheduling Strategy ? | Timer driven | | |
| Concurrent Tasks ? | 1 | Run Schedule ? | 3600 sec |
| Execution ? | All nodes | | |

- b. Set the data mime type
- c. Make the Flowfile name unique (using Nifi Expression)

Link on success to:

| Processor | Properties | |
|-----------------|-----------------------------|------------------|
| UpdateAttribute | Add these custom attributes | |
| | filename | \${UUID()} |
| | mime.type | application/json |

| | |
|---|---|
|  | UpdateAttribute |
| | UpdateAttribute 1.9.0.1.0.0.0-90 |
| | org.apache.nifi - nifi-update-attribute-nar |
| In | 0 (0 bytes) |
| Read/Write | 0 bytes / 0 bytes |
| Out | 0 (0 bytes) |
| Tasks/Time | 0 / 00:00:00.000 |

Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|--------------------------------------|------------|---------------------------|---|
| Required field | | | |
| Property | | Value | |
| Delete Attributes Expression | ? | No value set | |
| Store State | ? | Do not store state | |
| Stateful Variables Initial Value | ? | No value set | |
| Cache Value Lookup Cache Size | ? | 100 | |
| filename | ? | \${UUID()} |  |
| mime.type | ? | application/json |  |

We used SplitJson and EvaluateJsonPath processor to create 3 attributes (companyName, symbol and Open). We will use them in the next tasks.

Link on success to:

| Processor | Properties | |
|----------------------|---------------------|------|
| SplitJson | JsonPath Expression | \$.* |
| <u>Relationships</u> | | |
| Failure | | |
| success | | |

| | |
|---|-------------------------------------|
|  | SplitJson |
| | SplitJson 1.9.0.1.0.0.0-90 |
| | org.apache.nifi - nifi-standard-nar |
| In | 0 (0 bytes) |
| Read/Write | 0 bytes / 0 bytes |
| Out | 0 (0 bytes) |
| Tasks/Time | 0 / 00:00:00.000 |

Configure Processor

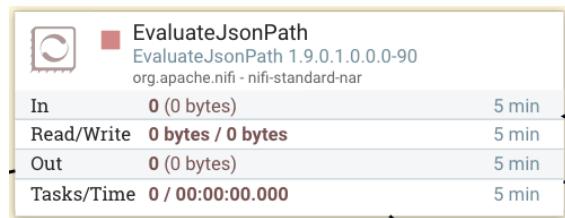
SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

| Property | Value |
|---------------------------|--------------|
| JsonPath Expression | \$.* |
| Null Value Representation | empty string |

Link on split to:

| Processor | Properties | |
|--|------------------------------------|----------------------|
| SplitJson <u>Relationships</u> Failure unmatched | Destination | Flowfile-attribute |
| | Null Value Representation | Empty string |
| | Add these custom attributes | |
| | companyName | \$.quote.companyName |
| | symbol | \$.quote.symbol |
| | Open | \$.quote.open |



Processor Details

SETTINGS SCHEDULING PROPERTIES COMMENTS

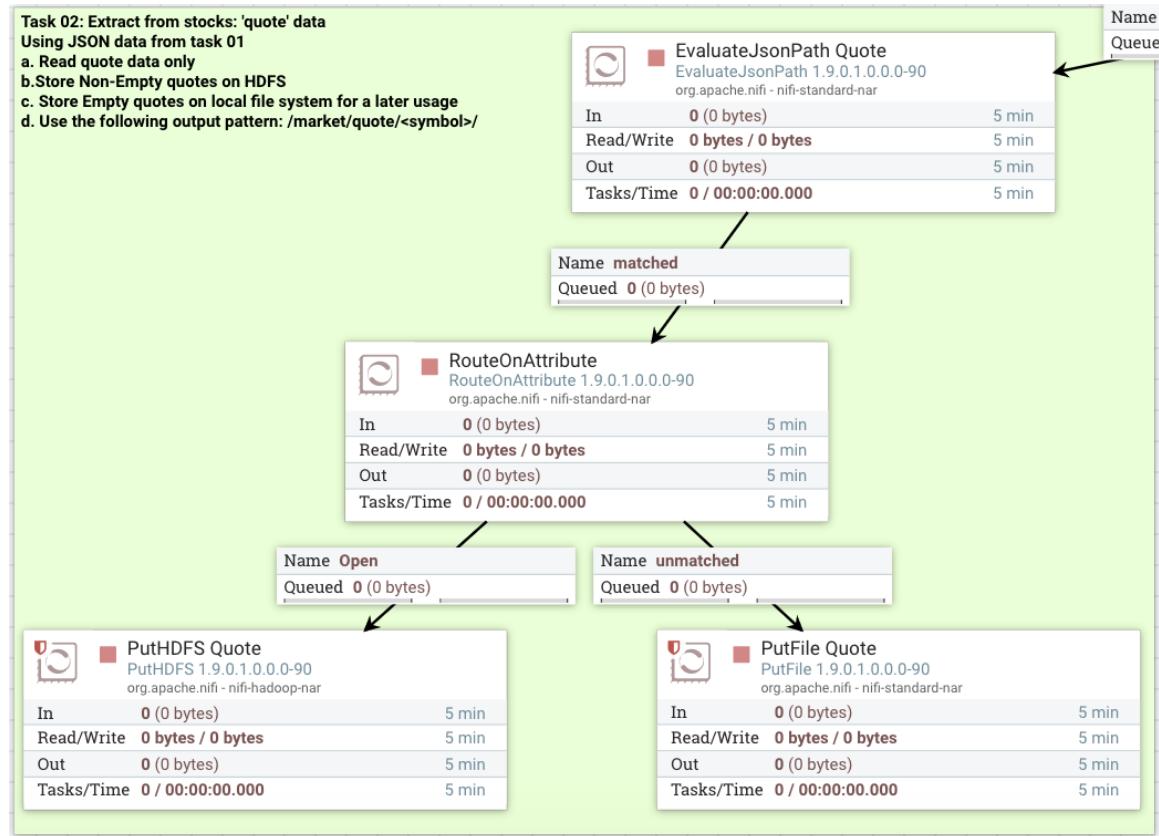
Required field

| Property | Value |
|---------------------------|----------------------|
| Destination | flowfile-attribute |
| Return Type | auto-detect |
| Path Not Found Behavior | ignore |
| Null Value Representation | empty string |
| companyName | \$.quote.companyName |
| Open | \$.quote.open |
| symbol | \$.quote.symbol |

TASK 02

Extract from stocks: 'quote' data

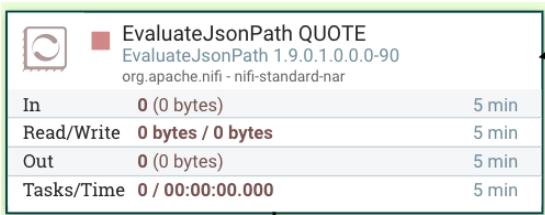
Using JSON data from task01



a. Read quote data only

Link on matched to:

| Processor | Properties | |
|---|----------------------------------|------------------|
| EvaluateJsonPath <u>Relationships</u> Failure unmatched | Destination | Flowfile-content |
| | Null Value Representation | Empty string |
| | Add this custom attribute | |
| | quote | \$quote |



Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

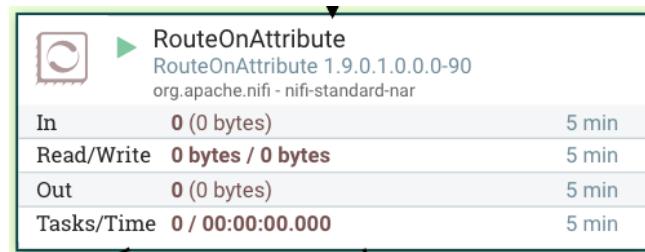
| Property | Value |
|---------------------------|------------------|
| Destination | flowfile-content |
| Return Type | auto-detect |
| Path Not Found Behavior | ignore |
| Null Value Representation | empty string |
| quote | \$.quote |

- b. Store Non-Empty quotes on HDFS
- c. Store Empty quotes on local system for a later usage
- d. Use the following output pattern: /market/quote/<symbol>/

Hint: To check if a stocks quote is empty, check the value 'open' attribute.

Link on matched to:

| Processor | Properties | |
|------------------|---------------------------|---------------------------|
| EvaluateJsonPath | Routing Strategy | Route to Property name |
| | Add this custom attribute | |
| | Open | \$.{Open:isEmpty():not()} |



Processor Details

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|-----------------------|------------|--------------------------|----------|
| Required field | | | |
| Property | | Value | |
| Routing Strategy | ? | Route to Property name | |
| Open | ? | \$(Open:isEmpty():not()) | |

Link on Open / unmatched to:

| Processor | Properties | |
|----------------------------------|---------------------------------|---|
| PutHDFS (open) | Hadoop Configuration Ressources | /etc/hadoop/conf/core-site.xml,/etc/hadoop/conf/hdfs-site.xml |
| Relationships Failure success | Directory | /market/quote/\${symbol}/ |
| | Conflict Resolution Strategy | Replace |
| | Permissions unmask | 000 |
| | Directory | /home/cloudera/Downloads/market/quote/\${symbol}/ |
| Relationships Failure success | Conflict Resolution Strategy | replace |
| | Create missing directory | true |
| | Permissions | 777 |
| | | |

| | |
|---|-----------------------------------|
|  | PutHDFS Quote |
| | PutHDFS 1.9.0.1.0.0.0-90 |
| | org.apache.nifi - nifi-hadoop-nar |
| In | 0 (0 bytes) |
| Read/Write | 0 bytes / 0 bytes |
| Out | 0 (0 bytes) |
| Tasks/Time | 0 / 00:00:00.000 |

Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|--------------------------------|------------|---|--------------|
| Required field | | | + |
| Property | | | Value |
| Hadoop Configuration Resources | ? | /etc/hadoop/conf/core-site.xml,/etc/hadoop/conf/hdfs-sit... | |
| Kerberos Credentials Service | ? | No value set | |
| Kerberos Principal | ? | No value set | |
| Kerberos Keytab | ? | No value set | |
| Kerberos Relogin Period | ? | 4 hours | |
| Additional Classpath Resources | ? | No value set | |
| Directory | ? | /market/quote/\${symbol}/ | |
| Conflict Resolution Strategy | ? | replace | |
| Block Size | ? | No value set | |
| IO Buffer Size | ? | No value set | |
| Replication | ? | No value set | |
| Permissions umask | ? | 000 | |

For the PutHDFS, we see the open is not empty for the result quote:

```

43      "oddLotDelayedPrice" : 160.58,
44      "oddLotDelayedPriceTime" : 1637355599827,
45      "open" : 157.87,
46      "openTime" : 1637332202707,
47      "openSource" : "official",
48      "peRatio" : 14.3,
49      "previousClose" : 157.87,
```

Two files and directory are created (for AAPL and MSFT) in HDFS on directory:

```

# Check the directory recursively
hdfs dfs -ls -R -h /market
drwxrwxrwx  - root supergroup          0 2021-11-22 06:07 /market/quote
drwxrwxrwx  - root supergroup          0 2021-11-22 06:07 /market/quote/AAPL
-rw-rw-rw-  1 root supergroup       1.3 K 2021-11-22 06:07 /market/quote/AAPL/c963fbad-eff0-411a-b16e-01b34804b224
drwxrwxrwx  - root supergroup          0 2021-11-22 06:07 /market/quote/MSFT
-rw-rw-rw-  1 root supergroup       1.3 K 2021-11-22 06:07 /market/quote/MSFT/c963fbad-eff0-411a-b16e-01b34804b224
```



Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS | | | | | | | | | | | | | | | | | | |
|---|---|------------|----------|-------|-----------|---|------------------------------|---------|----------------------------|------|--------------------|--------------|--------------------|--------------|-------------|-----|-------|--------------|-------|--------------|--|
| Required field | | | | | | | | | | | | | | | | | | | | | |
| <table border="1"><thead><tr><th>Property</th><th>Value</th></tr></thead><tbody><tr><td>Directory</td><td>/home/cloudera/Downloads/market/quote/\${symbol}/</td></tr><tr><td>Conflict Resolution Strategy</td><td>replace</td></tr><tr><td>Create Missing Directories</td><td>true</td></tr><tr><td>Maximum File Count</td><td>No value set</td></tr><tr><td>Last Modified Time</td><td>No value set</td></tr><tr><td>Permissions</td><td>777</td></tr><tr><td>Owner</td><td>No value set</td></tr><tr><td>Group</td><td>No value set</td></tr></tbody></table> | | | Property | Value | Directory | /home/cloudera/Downloads/market/quote/\${symbol}/ | Conflict Resolution Strategy | replace | Create Missing Directories | true | Maximum File Count | No value set | Last Modified Time | No value set | Permissions | 777 | Owner | No value set | Group | No value set | |
| Property | Value | | | | | | | | | | | | | | | | | | | | |
| Directory | /home/cloudera/Downloads/market/quote/\${symbol}/ | | | | | | | | | | | | | | | | | | | | |
| Conflict Resolution Strategy | replace | | | | | | | | | | | | | | | | | | | | |
| Create Missing Directories | true | | | | | | | | | | | | | | | | | | | | |
| Maximum File Count | No value set | | | | | | | | | | | | | | | | | | | | |
| Last Modified Time | No value set | | | | | | | | | | | | | | | | | | | | |
| Permissions | 777 | | | | | | | | | | | | | | | | | | | | |
| Owner | No value set | | | | | | | | | | | | | | | | | | | | |
| Group | No value set | | | | | | | | | | | | | | | | | | | | |

For the PutFile, we see the open is null for the result quote:

```
-- 37 "latestUpdate" : 1562616178268,
-- 38 "latestVolume" : 0,
-- 39 "low" : null,
-- 40 "lowSource" : null,
-- 41 "lowTime" : null,
-- 42 "marketCap" : 0,
-- 43 "oddLotDelayedPrice" : null,
-- 44 "oddLotDelayedPriceTime" : null,
-- 45 "open" : null,
-- 46 "openTime" : null,
```

Two files and directory are created (for CLDR and RHT) in our local system:

```
[cloudera@quickstart ~]$ cd Downloads
[cloudera@quickstart Downloads]$ ll -R market
market:
total 4
drwxr-xr-x 4 root root 4096 Nov 22 06:00 quote

market/quote:
total 8
drwxr-xr-x 2 root root 4096 Nov 22 06:00 CLDR
drwxr-xr-x 2 root root 4096 Nov 22 06:00 RHT

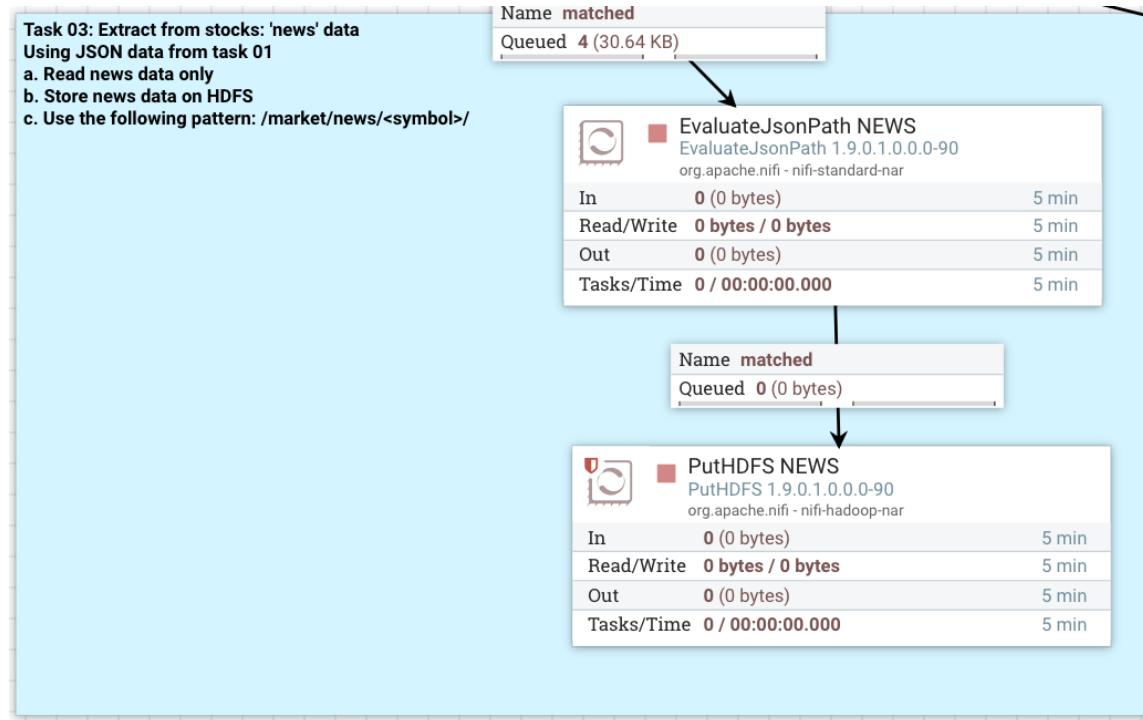
market/quote/CLDR:
total 4
-rwxrwxrwx 1 root root 1296 Nov 22 06:00 c963fbad-eff0-411a-b16e-01b34804b224

market/quote/RHT:
total 4
-rwxrwxrwx 1 root root 1157 Nov 22 06:00 c963fbad-eff0-411a-b16e-01b34804b224
[cloudera@quickstart Downloads]$
```

TASK 03

Extract from stocks: 'news' data

Using JSON data from task01



a. Read news data only

Link on matched to:

| Processor | Properties | |
|---|---------------------------|------------------|
| EvaluateJsonPath <u>Relationships</u> Failure unmatched | Destination | Flowfile-content |
| | Null Value Representation | Empty string |
| | Add this custom attribute | |
| | news | \$news |



Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|---------------------------|------------|------------------|--------------|
| Required field | | | |
| Property | | | Value |
| Destination | ? | flowfile-content | |
| Return Type | ? | auto-detect | |
| Path Not Found Behavior | ? | ignore | |
| Null Value Representation | ? | empty string | |
| news | ? | \$news | |

- b. Store news data on HDFS
- c. Use the following pattern: /market/news/<symbol>/

Link on matched to:

| Processor | Properties | |
|----------------------|-----------------------------------|---|
| PutHDFS | Hadoop Configuration Resources | /etc/hadoop/conf/core-site.xml,/etc/hadoop/conf/hdfs-site.xml |
| <u>Relationships</u> | | |
| Failure | Directory | /market/news/\${symbol}/ |
| success | Conflict Resolution Strategy | replace |
| | Permissions unmask | 000 |



Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|--------------------------------|------------|---|--------------|
| Required field | | | |
| Property | | | Value |
| Hadoop Configuration Resources | ? | /etc/hadoop/conf/core-site.xml,/etc/hadoop/conf/hdfs-sit... | |
| Kerberos Credentials Service | ? | No value set | |
| Kerberos Principal | ? | No value set | |
| Kerberos Keytab | ? | No value set | |
| Kerberos Relogin Period | ? | 4 hours | |
| Additional Classpath Resources | ? | No value set | |
| Directory | ? | /market/news/\${symbol}/ | |
| Conflict Resolution Strategy | ? | replace | |
| Block Size | ? | No value set | |
| IO Buffer Size | ? | No value set | |
| Replication | ? | No value set | |
| Permissions umask | ? | 000 | |

We see on HDFS four files and directory are created for the result news:

- For stock RHT, we don't have news data into the JSON.

```
# Check the directory recursively|
hdfs dfs -ls -R -h /market

drwxrwxrwx - root supergroup          0 2021-11-22 06:15 /market/news
drwxrwxrwx - root supergroup          0 2021-11-22 06:15 /market/news/AAPL
-rw-rw-rw-  1 root supergroup        74.1 K 2021-11-22 06:15 /market/news/AAPL/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
drwxrwxrwx - root supergroup          0 2021-11-22 06:15 /market/news/CLDR
-rw-rw-rw-  1 root supergroup        19.1 K 2021-11-22 06:15 /market/news/CLDR/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
drwxrwxrwx - root supergroup          0 2021-11-22 06:15 /market/news/MSFT
-rw-rw-rw-  1 root supergroup        80.2 K 2021-11-22 06:15 /market/news/MSFT/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
drwxrwxrwx - root supergroup          0 2021-11-22 06:15 /market/news/RHT
-rw-rw-rw-  1 root supergroup        2 2021-11-22 06:15 /market/news/RHT/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
```

```
[
  {
    "datetime" : 1637449582000,
    "headline" : "Heres where to meet New Yorks young(ish) crypto millionaires",
    "source" : "New York Post",
    "url" : "https://cloud.iexapis.com/v1/news/article/3fmFvCnNhSFNhtYgPc2kSu2TLSNG04o1EVHxEWVBPP",
    "summary" : "Young cryptocurrency millionaires -- and those who want to learn more about their hot assets -- converge weekly in the Big Apple.",
    "related" : "AAPL",
    "image" : "https://cloud.iexapis.com/v1/news/image/3fmFvCnNhSFNhtYgPc2kSu2TLSNG04o1EVHxEWVBPP",
    "lang" : "en",
    "hasPaywall" : false
  },
  {
    "datetime" : 1637448933000,
    "headline" : "Apple affirms employees right to speak about working conditions, in a win for #AppleToo movement",
    "source" : "Techtelegraph",
    "url" : "https://cloud.iexapis.com/v1/news/article/151hayiSGYjgrMCHWJ0TEDV5QXwwWztKattjKbfBIKuI",
    "summary" : "Apple is deeply committed to providing employees with a workplace where they feel safe, respected, and inspired to do their best work",
    "related" : "AAPL",
    "image" : "https://cloud.iexapis.com/v1/news/image/151hayiSGYjgrMCHWJ0TEDV5QXwwWztKattjKbfBIKuI",
    "lang" : "en",
    "hasPaywall" : false
  }
]
```

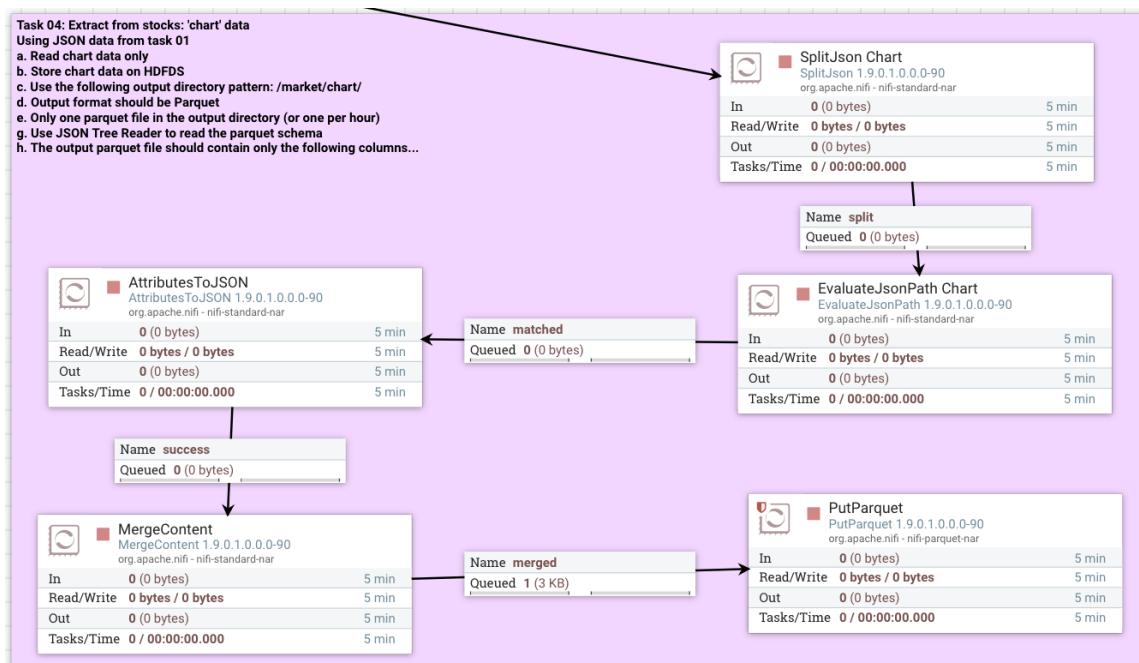
TASK 04

Extract from stocks: 'chart' data

Using JSON data from task01

- a) Read chart data only
- b) Store chart data on HDFS
- c) Use the following output directory pattern: /market/chart/
- d) Output format should be Parquet
- e) Parquet file should be compressed using SNAPPY compression
- f) Only one parquet file in the output directory (or one per hour)
- g) Use JSON Tree Reader to read the parquet schema
- h) The output parquet file should contain only the following columns:

| Source: quote | | Source: chart | | | | | |
|---------------|-------------|---------------|--------|--------|--------|--------|--------|
| symbol | companyName | date | open | close | high | low | volume |
| string | string | string | double | double | double | double | long |



Link on matched to:

| Processor | Properties | |
|---|---------------------------|--------------|
| SplitJson | JsonPath Expression | \$.chart |
| <u>Relationships</u> Failure original | Null Value Representation | Empty string |



Configure Processor

SETTINGS SCHEDULING **PROPERTIES** COMMENTS

Required field +

| Property | Value |
|---------------------------|--------------|
| JsonPath Expression | \$.chart |
| Null Value Representation | empty string |

Link on split to:

| Processor | Properties | |
|--|---------------------------|--------------------|
| EvaluateJsonPath | Destination | Flowfile-attribute |
| <u>Relationships</u> Failure unmatched | Null Value Representation | Empty string |
| Add these custom attributes | | |
| close | \$.close | |
| date | \$.date | |
| high | \$.high | |
| low | \$.low | |
| open | \$.open | |
| volume | \$.volume | |

| | | |
|---|-------------------------------------|-------|
|  | EvaluateJsonPath Chart | |
| | EvaluateJsonPath 1.9.0.1.0.0.0-90 | |
| | org.apache.nifi - nifi-standard-nar | |
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|---------------------------|------------|--------------------|---|
| Required field | | | |
| Property | | | Value |
| Destination | ? | flowfile-attribute | |
| Return Type | ? | auto-detect | |
| Path Not Found Behavior | ? | ignore | |
| Null Value Representation | ? | empty string | |
| close | ? | \$.close |  |
| date | ? | \$.date |  |
| high | ? | \$.high |  |
| low | ? | \$.low |  |
| open | ? | \$.open |  |
| volume | ? | \$.volume |  |

Link on matched to:

| Processor | Properties | |
|-----------------|-----------------|--|
| AttributeToJson | Attributes List | Symbol,companyName,date,open,close,high,low,volume |
| Relationships | Destination | Flowfile-content |
| Failure | | |

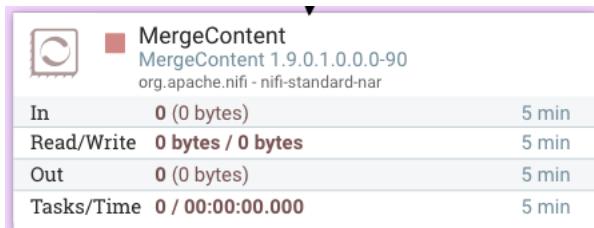
| | | |
|---|-------------------------------------|-------|
|  | AttributesToJson | |
| | AttributesToJson 1.9.0.1.0.0.0-90 | |
| | org.apache.nifi - nifi-standard-nar | |
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|-------------------------------|------------|--|----------|
| Required field | | | |
| Property | | | Value |
| Attributes List | ? | symbol,companyName,date,open,close,high,low,volume | |
| Attributes Regular Expression | ? | No value set | |
| Destination | ? | flowfile-content | |
| Include Core Attributes | ? | true | |
| Null Value | ? | false | |

Link on success to:

| Processor | Properties | |
|--|--------------------|--------------------------------|
| MergeContent <u>Relationships</u> Failure original | Merge Strategy | Bin-Packing Algorithm |
| | Merge Format | Binary Concatenation |
| | Attribute Strategy | Keep Only Common Attributes |
| | Metadata Strategy | Do not merge uncommon metadata |
| | Max Bin Age | 60 min |
| | Delimiter Strategy | Text |
| | Demarcator | Shift + enter |



Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|----------------------------|------------|--------------------------------|----------|
| Required field | | | |
| Property | | | Value |
| Merge Strategy | ? | Bin-Packing Algorithm | |
| Merge Format | ? | Binary Concatenation | |
| Attribute Strategy | ? | Keep Only Common Attributes | |
| Correlation Attribute Name | ? | No value set | |
| Metadata Strategy | ? | Do Not Merge Uncommon Metadata | |
| Minimum Number of Entries | ? | 1 | |
| Maximum Number of Entries | ? | 1000 | |
| Minimum Group Size | ? | 0 B | |
| Maximum Group Size | ? | No value set | |
| Max Bin Age | ? | 60 min | |
| Maximum number of Bins | ? | 5 | |
| Delimiter Strategy | ? | Text | |
| Header | ? | No value set | |

| | | | |
|-------------------|-------------------|---|--|
| Footer | ? | No value set | |
| Demarcator | ? | | |
| Compression Level | ? | 1 | |
| Keep Path | ? | false | |
| Tar Modified Time | ? | <code> \${file.lastModifiedTime}</code> | |

Link on merged to:

| Processor | Properties | |
|--|--------------------------------|---|
| PutParquet <u>Relationships</u> Failure Retry success | Hadoop Configuration Resources | /etc/hadoop/conf/core-site.xml,/etc/hadoop/conf/hdfs-site.xml |
| | Record Reader | Json TreeReader |
| | Directory | /market/chart |
| | Compression Type | SNAPPY |
| | Overwrite Files | true |
| | Permissions unmask | 000 |



Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|--|------------|-----------------------|-------------------|
| Required field | | | |
| Hadoop Configuration Resources | | Value | |
| Kerberos Credentials Service | | Value | |
| Kerberos Principal | | Value | |
| Kerberos Keytab | | Value | |
| Kerberos Relogin Period | | Value | |
| Additional Classpath Resources | | Value | |
| Record Reader | | Value | → |
| Directory | | Value | |
| Compression Type | | Value | |
| Overwrite Files | | Value | |
| Permissions umask | | Value | |

For the Record Reader (JsonTreeReader):

| GENERAL | | CONTROLLER SERVICES | | |
|----------------|---------------------------------|--|---------|--------------|
| Name | Type | Bundle | State | Scope |
| JsonTreeReader | JsonTreeReader 1.9.0.1.0.0.0-90 | org.apache.nifi - nifi-record-serialization... | Enabled | Assignment 2 |

View Configuration:

| Property | Value |
|------------------------|-----------------------------------|
| Schema Access Strategy | Use 'Schema Text' Property |
| Schema Text | See below the view of schema text |

Controller Service Details

SETTINGS PROPERTIES COMMENTS

Required field

| Property | Value |
|------------------------|--|
| Schema Access Strategy | Use 'Schema Text' Property |
| Schema Registry | No value set |
| Schema Name | \$(schema.name) |
| Schema Version | No value set |
| Schema Branch | No value set |
| Schema Text | {"type": "record", "namespace": "chart", "name": "d... |
| Schema Inference Cache | No value set |
| Date Format | No value set |
| Time Format | No value set |
| Timestamp Format | No value set |

Schema text:

```

1  {
2      "type": "record",
3      "namespace": "chart",
4      "name": "data",
5      "fields": [
6          { "name": "symbol", "type": "string" },
7          { "name": "compagnyName", "type": "string" },
8          { "name": "date", "type": "string" },
9          { "name": "open", "type": "double" },
10         { "name": "close", "type": "double" },
11         { "name": "high", "type": "double" },
12         { "name": "low", "type": "double" },
13         { "name": "volume", "type": "long" }
14     ]
15 }
```

We need to enable the processors after these steps.

Controller Service Details

| SETTINGS | PROPERTIES | COMMENTS |
|--|------------|---|
| Name JsonTreeReader | | Referencing Components ? ▼ Processors (1) ■ PutParquet PutParquet |
| Id 017d12a1-98df-1464-8211-4554a21ee3b8 | | |
| Type JsonTreeReader 1.9.0.1.0.0.0-90 | | |
| Bundle org.apache.nifi - nifi-record-serialization-services-nar | | |
| Supports Controller Service | | |
| • RecordReaderFactory 1.9.0.1.0.0.0-90 from org.apache.nifi - nifi-standard-services-api-nar | | |

For this step we run schedule for one hour.

Configure Processor

| SETTINGS | SCHEDULING | PROPERTIES | COMMENTS |
|---------------------------------------|--------------|--------------------------------|----------|
| Scheduling Strategy ? | Timer driven | | |
| Concurrent Tasks ? | 1 | Run Schedule ? | 3600 sec |
| Execution ? | All nodes | | |

We see on HDFS one file and directory is created for the result chart:

- For stock RHT, we don't have chart data into the JSON.

```
# Create a new Impala user manager table named market that describes the following structure
# Check the directory recursively
hdfs dfs -ls -R -h /market

drwxrwxrwx - root supergroup          0 2021-11-22 06:24 /market/chart
-rw-rw-rw-  1 root supergroup      6.9 K 2021-11-22 06:24 /market/chart/64981d19-6511-4c3b-b5c2-dec1ba087923

61 {"date": "2021-08-26", "volume": "48597195", "symbol": "AAPL", "high": "149.12", "low": "147.51", "companyName": "Apple Inc", "close": "147.54", "open": "148.35"},  
62 {"date": "2021-08-25", "volume": "58991297", "symbol": "AAPL", "high": "150.32", "low": "147.8", "companyName": "Apple Inc", "close": "149.36", "open": "149.81"},  
63 {"date": "2021-08-24", "volume": "48606428", "symbol": "AAPL", "high": "150.86", "low": "149.15", "companyName": "Apple Inc", "close": "149.62", "open": "149.45"},  
64 {"date": "2021-08-23", "volume": "60131810", "symbol": "AAPL", "high": "150.19", "low": "147.89", "companyName": "Apple Inc", "close": "149.71", "open": "148.31"},  
65 {"date": "2021-11-19", "volume": "21963403", "symbol": "MSFT", "high": "345.1", "low": "342.2", "companyName": "Microsoft Corporation", "close": "343.11", "open": "342.64"},  
66 {"date": "2021-11-18", "volume": "22463533", "symbol": "MSFT", "high": "342.45", "low": "337.12", "companyName": "Microsoft Corporation", "close": "341.27", "open": "338.18"},  
67 {"date": "2021-11-17", "volume": "19053380", "symbol": "MSFT", "high": "342.19", "low": "338", "companyName": "Microsoft Corporation", "close": "339.12", "open": "338.94"}...
```

TASK 05

Create a new Impala user-managed table named market that satisfies the following criteria:

- o The table schema should match stock ‘chart’ parquet schema

Check the directory recursively

```
%sh
```

```
hdfs dfs -ls -R -h /market
```

The parquet file is created in /market/chart:

```
drwxrwxrwx  - root supergroup      0 2021-11-22 06:24 /market/chart
-rw-rw-rw-  1 root supergroup  6.9 K 2021-11-22 06:24 /market/chart/64981d19-6511-4c3b-b5c2-dec1ba087923
drwxrwxrwx  - root supergroup      0 2021-11-22 06:15 /market/news
drwxrwxrwx  - root supergroup      0 2021-11-22 06:15 /market/news/AAPL
-rw-rw-rw-  1 root supergroup  74.1 K 2021-11-22 06:15 /market/news/AAPL/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
drwxrwxrwx  - root supergroup      0 2021-11-22 06:15 /market/news/CLDR
-rw-rw-rw-  1 root supergroup  19.1 K 2021-11-22 06:15 /market/news/CLDR/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
drwxrwxrwx  - root supergroup      0 2021-11-22 06:15 /market/news/MSFT
-rw-rw-rw-  1 root supergroup  80.2 K 2021-11-22 06:15 /market/news/MSFT/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
drwxrwxrwx  - root supergroup      0 2021-11-22 06:15 /market/news/RHT
-rw-rw-rw-  1 root supergroup   2 2021-11-22 06:15 /market/news/RHT/92233b5d-8ef5-4640-9c5b-b6c862ff4c3c
drwxrwxrwx  - root supergroup      0 2021-11-22 06:07 /market/quote
drwxrwxrwx  - root supergroup      0 2021-11-22 06:07 /market/quote/AAPL
-rw-rw-rw-  1 root supergroup  1.3 K 2021-11-22 06:07 /market/quote/AAPL/c963fbad-eff0-411a-b16e-01b34804b224
drwxrwxrwx  - root supergroup      0 2021-11-22 06:07 /market/quote/MSFT
-rw-rw-rw-  1 root supergroup  1.3 K 2021-11-22 06:07 /market/quote/MSFT/c963fbad-eff0-411a-b16e-01b34804b224
```

Create a database stock

```
%impala
```

```
create database stock
```

Use the database stock

```
%impala
```

```
use stock
```

Create a parquet schema with the file

```
%sh
```

```
hdfs dfs -get /market/chart/03f4aba6-348a-4307-a1a6-ef1e90b997a8
/home/cloudera/Downloads/market.parquet
```

Create an avro schema with the file

```
parquet-tools schema -d /home/cloudera/Downloads/market.parquet >
/home/cloudera/Downloads/market.parquet.avsc
```

Create a directory /market/schema on HDFS to put the schema

```
%sh
```

```
hdfs dfs -mkdir /market/schema
```

Put the schema in the directory on HDFS

```
%sh
```

```
hdfs dfs -put /home/cloudera/Downloads/market.* /market/schema
```

Check of the schema is on the HDFS directory

```
%sh
```

```
hdfs dfs -ls -h /market/schema
```

Found 2 items

| | | | | | | |
|------------|---|------|------------|-------|------------------|-------------------------------|
| -rw-r--r-- | 1 | root | supergroup | 6.9 K | 2021-11-22 06:51 | /market/schema/market.parquet |
| -rw-r--r-- | 1 | root | supergroup | 1.9 K | 2021-11-22 06:51 | /market/schema/market.avsc |

Read the avro parquet (more reading)

```
%sh
```

```
hdfs dfs -cat /market/schema/market.parquet.avsc | head -n 15
```

```
message chart.data {  
    required binary symbol (UTF8);  
    required binary companyName (UTF8);  
    required binary date (UTF8);  
    required double open;  
    required double close;  
    required double high;  
    required double low;  
    required int64 volume;  
}  
  
creator: parquet-mr version 1.10.0 (build 03106654009e3b82020012a18434c582bd74c73a)  
extra: writer.model.name = avro  
extra: parquet.avro.schema = {"type": "record", "name": "data", "namespace": "chart", "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"}]}
```

Create Impala user-managed table

```
%impala
```

```
create external table market  
like PARQUET '/market/schema/market.parquet'  
stored as parquet  
location '/market/chart/'  
TBLPROPERTIES ("parquet.compression"="SNAPPY");
```

o Report the rows count and the size for this table

Compute stats of the table

```
%impala
```

```
compute stats market;
```

Show stats of the table

```
%impala
```

```
show table stats market;
```

| #Rows | #Files | Size |
|-------|--------|--------|
| 161 | 1 | 6.92KB |

| Row Count | Table Size |
|-----------|------------|
| 161 | 6.92KB |

TASK 06

Use Impala to run the following analysis on the market table

Q1. Show the first 5 line of the table.

```
%impala  
select * from market limit 5;
```

| symbol | companyname | date | open | close | high | low | volume |
|--------|-------------|------------|---------|--------|---------|----------|-----------|
| AAPL | Apple Inc | 2021-11-19 | 157.65 | 160.55 | 161.02 | 156.5328 | 117305597 |
| AAPL | Apple Inc | 2021-11-18 | 153.71 | 157.87 | 158.67 | 153.05 | 137827673 |
| AAPL | Apple Inc | 2021-11-17 | 150.995 | 153.49 | 155 | 150.99 | 88807000 |
| AAPL | Apple Inc | 2021-11-16 | 149.94 | 151 | 151.488 | 149.34 | 59256210 |
| AAPL | Apple Inc | 2021-11-15 | 150.37 | 150 | 151.88 | 149.43 | 59222803 |

Q2. Find the count of charts per stock symbol.

```
%impala  
select symbol, count(*) as count  
from market  
group by symbol
```

| symbol | count |
|--------|-------|
| AAPL | 64 |
| MSFT | 64 |
| CLDR | 33 |

Q3. Find max and min per stock symbol for the following columns: open, high, low, volume

```
%impala  
select symbol, min(open) as min_open, max(open) as max_open,  
min(high) as min_high, max(high) as max_high,  
min(low) as min_low, max(low) as max_low,  
min(volume) as min_volume, max(volume) as max_volume  
from market  
group by symbol
```

| symbol | min_open | max_open | min_high | max_high | min_low | max_low | min_volume | max_volume |
|--------|----------|----------|----------|----------|---------|----------|------------|------------|
| AAPL | 139.47 | 157.65 | 141.4 | 161.02 | 138.27 | 156.5328 | 40999950 | 140893235 |
| MSFT | 282.1217 | 342.64 | 286.77 | 345.1 | 280.25 | 342.2 | 14751610 | 52588690 |
| CLDR | 15.9 | 15.98 | 15.91 | 16 | 15.81 | 15.98 | 1006851 | 15203961 |

Note: The RHT stock is empty for the chart in the JSON.

TASK 07

For task 7, we used two different methods. The first one consists of using the Impala built-in visualization tools. The second one consists of creating an Impala view. Both methods are presented below.

Method 1

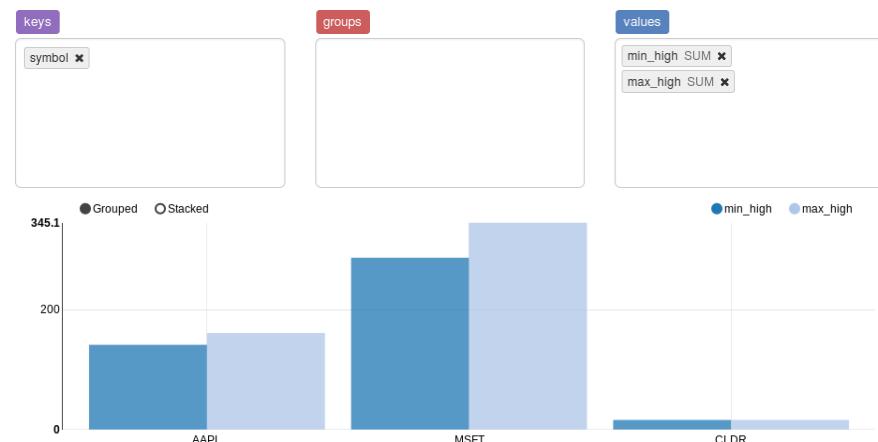
- a) Create an Impala view that report max and min per stock symbol for the following columns: open, high, low, volume.

```
%impala
select symbol, min(open) as min_open, max(open) as max_open,
       min(high) as min_high, max(high) as max_high,
       min(low) as min_low, max(low) as max_low,
       min(volume) as min_volume, max(volume) as max_volume
from market
group by symbol
```

1) Open



2) High



3) Low



4) Volume



b) Use an ODBC connection to read this view and import data into MS Excel or any external visualization tool you might have on your system.

We used two different external visualization tools, Tableau and MS Excel.

1) Tableau

ODBC connection

iODBC Data Source Administrator (Intel 64bit)

System Data Sources

| Name | Description | Driver |
|------------|-------------|---------------------------------|
| Sample DSN | - | Cloudera ODBC Driver for Impala |

Add Remove Configure Test

System Data Source Names create systems-wide ODBC Data Names that resolve to Data Source connection information.

Cancel OK

Setup of Sample DSN

Data Source Name (DSN) **Sample DSN**

Comment

| Keyword | Value |
|-------------|-----------|
| Host | localhost |
| Port | 21050 |
| Database | default |
| ServiceName | impala |
| UID | cloudera |
| PWD | cloudera |

+ - Cancel Ok

Connexions Ajouter

localhost Autres bases de données (ODBC)

Base de données IMPALA

Schéma Sélectionnez le schéma

Table Entrez le nom de la table

Exacte Contient Comme...

- market (stock.market)
- nyctaxi (default.nyctaxi)
- nyctaxi (nyc.nyctaxi)
- nyctaxi_part...axi_part_bkt

market

Connexion En direct Extrait

Filtres 0 Ajouter

market

Vous avez besoin de données supplémentaires ? Faites glisser les tables ici pour les relier. [En savoir plus](#)

Trier les champs Ordre de la source de donnée: Afficher les alias Afficher les ch

| symbol | companyname | date | market open | market close | market high |
|--------|-------------|------------|-------------|--------------|-------------|
| AAPL | Apple Inc | 2021-11-19 | 157,650 | 160,550 | 161,020 |
| AAPL | Apple Inc | 2021-11-18 | 153,710 | 157,870 | 158,670 |
| AAPL | Apple Inc | 2021-11-17 | 150,995 | 153,490 | 155,000 |

Autres bases de données (ODBC)

Connexion via

ODBC générique a besoin d'une configuration supplémentaire pour que la publication aboutisse. Sélectionnez le nom de la source de données (DSN) pour assurer la portabilité entre les plates-formes. Un DSN portant le même nom doit être configuré dans Tableau Server.

DSN: Sample DSN

Pilote:

Attributs de connexion

Serveur: localhost Port: 21050

Base de données: default

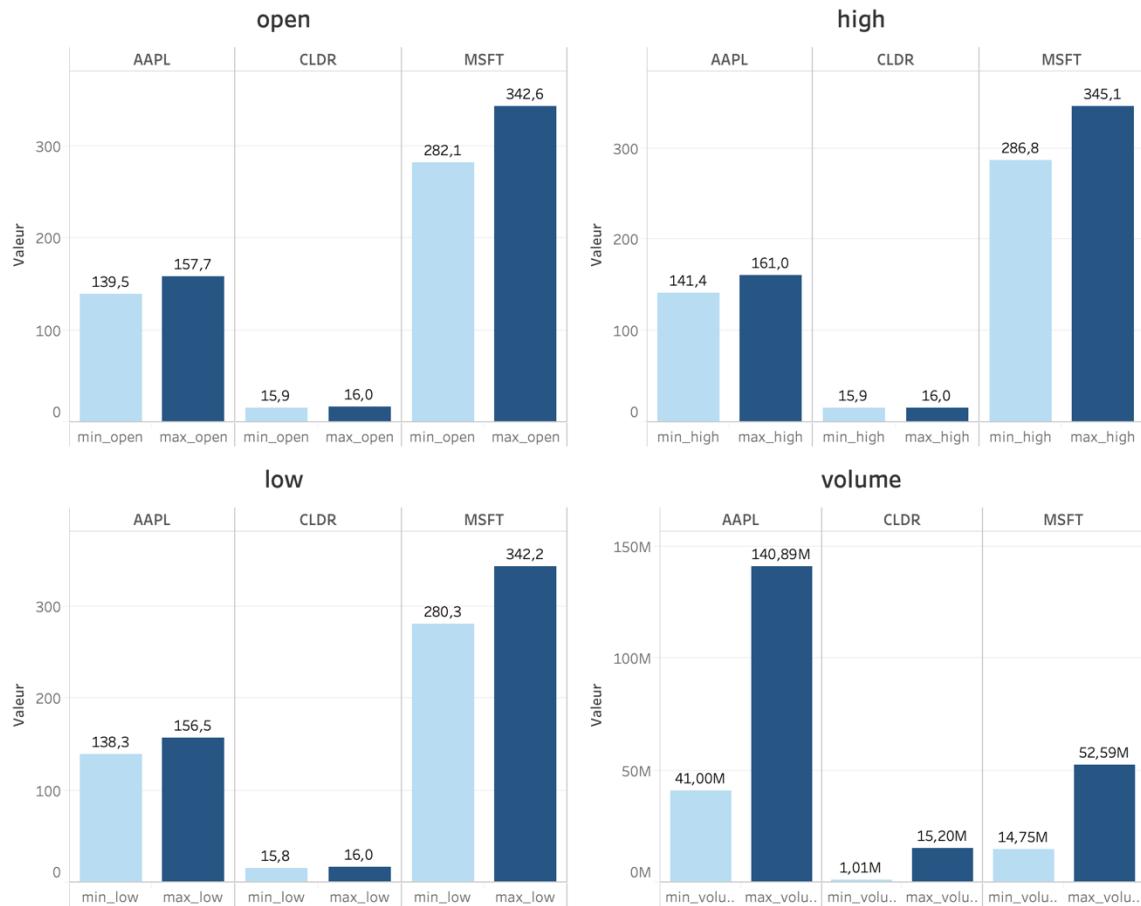
Nom d'utilisateur: cloudera

Mot de passe: *****

Chaines supplémentaires:

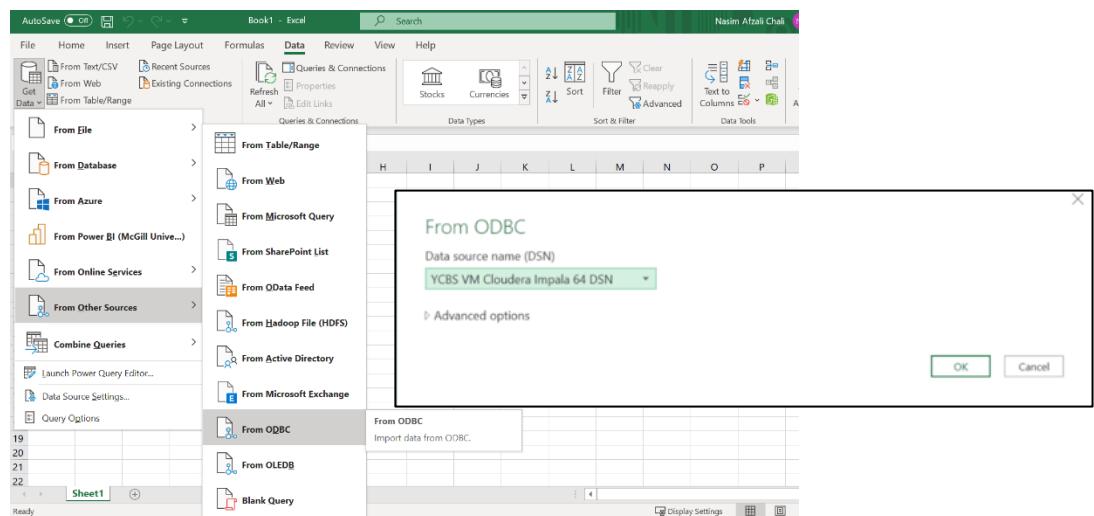
Connexion

Charts



2) MS Excel

ODBC connection



Screenshot of the Power BI Data Tools interface showing the Navigator pane and a table preview.

Navigator

- Display Options: Select multiple items
- ODBC (dsn=YCBS VM Cloudera ...)
- IMPALA [5]
- bixi
- covid_19
- default
- market_stock
- stock [1]
- market

market

| symbol | companyname | date | open | close | high | low |
|--------|-------------|------------|---------|--------|---------|-------|
| AAPL | Apple Inc | 2021-11-19 | 157.65 | 160.55 | 161.02 | 156.5 |
| AAPL | Apple Inc | 2021-11-18 | 153.71 | 157.87 | 158.67 | 151.5 |
| AAPL | Apple Inc | 2021-11-17 | 150.995 | 153.49 | 155 | 150 |
| AAPL | Apple Inc | 2021-11-16 | 149.94 | 151 | 151.488 | 149 |
| AAPL | Apple Inc | 2021-11-15 | 150.37 | 150 | 151.88 | 149 |
| AAPL | Apple Inc | 2021-11-12 | 148.43 | 149.99 | 150.4 | 147 |
| AAPL | Apple Inc | 2021-11-11 | 148.96 | 147.87 | 149.43 | 147 |
| AAPL | Apple Inc | 2021-11-10 | 150.02 | 147.92 | 150.13 | 147 |
| AAPL | Apple Inc | 2021-11-09 | 150.2 | 150.81 | 151.428 | 150.0 |
| AAPL | Apple Inc | 2021-11-08 | 151.41 | 150.44 | 151.57 | 150 |
| AAPL | Apple Inc | 2021-11-05 | 151.89 | 151.28 | 152.2 | 150 |
| AAPL | Apple Inc | 2021-11-04 | 151.58 | 150.96 | 152.43 | 150 |
| AAPL | Apple Inc | 2021-11-03 | 150.39 | 151.49 | 151.97 | 149 |
| AAPL | Apple Inc | 2021-11-02 | 148.66 | 150.02 | 151.57 | 148 |
| AAPL | Apple Inc | 2021-11-01 | 148.985 | 148.96 | 149.7 | 148 |
| AAPL | Apple Inc | 2021-10-29 | 147.215 | 149.8 | 149.94 | 146.4 |
| AAPL | Apple Inc | 2021-10-28 | 149.82 | 152.57 | 153.165 | 149 |
| AAPL | Apple Inc | 2021-10-27 | 149.36 | 148.85 | 149.73 | 148 |
| AAPL | Apple Inc | 2021-10-26 | 149.33 | 149.32 | 150.84 | 149.0 |
| AAPL | Apple Inc | 2021-10-25 | 148.68 | 148.64 | 149.37 | 147.6 |
| AAPL | Apple Inc | 2021-10-22 | 149.69 | 148.69 | 150.18 | 141 |
| AAPL | Apple Inc | 2021-10-21 | 148.81 | 149.48 | 149.64 | 141 |
| AAPL | Apple Inc | 2021-10-20 | 149.7 | 149.76 | 149.73 | 140 |

Screenshot of Microsoft Excel showing the 'market' table loaded from Power BI.

Table Name: market

Table Style Options:

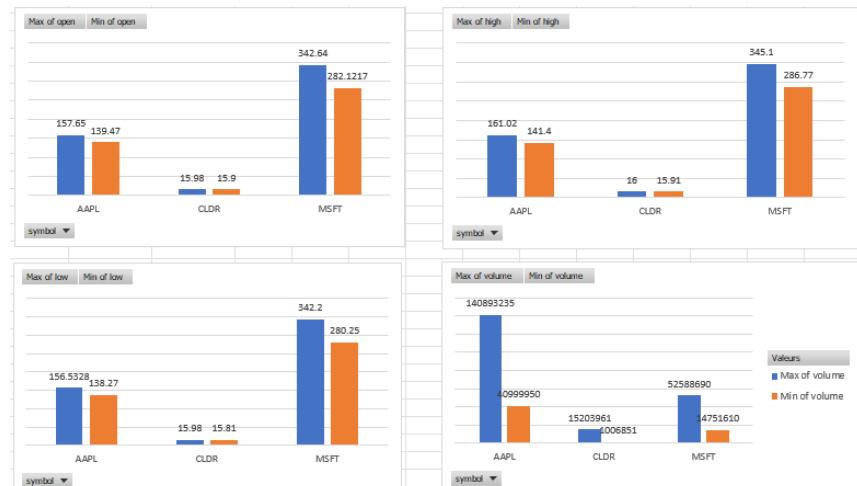
- Header Row
- Total Row
- Banded Rows
- First Column
- Last Column
- Banded Columns
- Filter Button
- Table Style Options

Queries & Connections

1 query
market
161 rows loaded.

| symbol | companyname | date | open | close | high | low | volume |
|--------|-------------|------------|---------|---------|---------|----------|-----------|
| AAPL | Apple Inc | 2021-11-19 | 157.65 | 160.55 | 161.02 | 156.5328 | 117805597 |
| AAPL | Apple Inc | 2021-11-18 | 153.71 | 157.87 | 158.67 | 150.5 | 137827673 |
| AAPL | Apple Inc | 2021-11-17 | 150.995 | 153.49 | 155 | 150.99 | 888070200 |
| AAPL | Apple Inc | 2021-11-16 | 149.94 | 151.488 | 149.34 | 92156210 | |
| AAPL | Apple Inc | 2021-11-15 | 150.37 | 150 | 151.89 | 149.43 | 59222803 |
| AAPL | Apple Inc | 2021-11-12 | 148.43 | 149.99 | 150.4 | 147.4 | 638040008 |
| AAPL | Apple Inc | 2021-11-11 | 148.96 | 147.87 | 149.43 | 147.681 | 40999950 |
| AAPL | Apple Inc | 2021-11-10 | 150.02 | 147.92 | 150.13 | 147.85 | 65187092 |
| AAPL | Apple Inc | 2021-11-09 | 150.2 | 150.81 | 151.428 | 150.0601 | 56787930 |
| AAPL | Apple Inc | 2021-11-08 | 151.41 | 150.44 | 151.57 | 150.16 | 55020868 |
| AAPL | Apple Inc | 2021-11-05 | 151.88 | 151.28 | 152.2 | 150.06 | 65463883 |
| AAPL | Apple Inc | 2021-11-04 | 151.56 | 150.96 | 152.43 | 150.64 | 60394616 |
| AAPL | Apple Inc | 2021-11-03 | 150.38 | 151.49 | 151.97 | 149.82 | 54511534 |
| AAPL | Apple Inc | 2021-11-02 | 148.66 | 150.02 | 151.57 | 148.65 | 69121987 |
| AAPL | Apple Inc | 2021-11-01 | 148.985 | 148.96 | 149.7 | 147.8 | 74588258 |
| AAPL | Apple Inc | 2021-10-29 | 147.215 | 149.8 | 149.94 | 146.4128 | 124953168 |
| AAPL | Apple Inc | 2021-10-28 | 149.82 | 152.57 | 153.165 | 149.72 | 100077888 |
| AAPL | Apple Inc | 2021-10-27 | 149.36 | 148.85 | 149.73 | 148.49 | 56094929 |
| AAPL | Apple Inc | 2021-10-26 | 149.33 | 149.32 | 150.84 | 149.0101 | 60893395 |
| AAPL | Apple Inc | 2021-10-25 | 148.68 | 148.64 | 149.37 | 147.6211 | 50720556 |
| AAPL | Apple Inc | 2021-10-22 | 149.69 | 148.69 | 150.18 | 148.64 | 58883443 |

Charts



Method 2

- a) Create an Impala view that report max and min per stock symbol for the following columns: open, high, low, volume.

```
%impala
create view market_view as
select `symbol`,min(open) as min_open, max(open) as max_open,
       min(high) as min_high, max(high) as max_high,
       min(low) as min_low, max(low) as max_low,
       min(volume) as min_volume, max(volume) as max_volume
from market
group by `symbol`
```

- b) Use an ODBC connection to read this view and import data into MS Excel or any external visualization tool you might have on your system.

We used two different external visualization tools, Tableau and MS Excel.

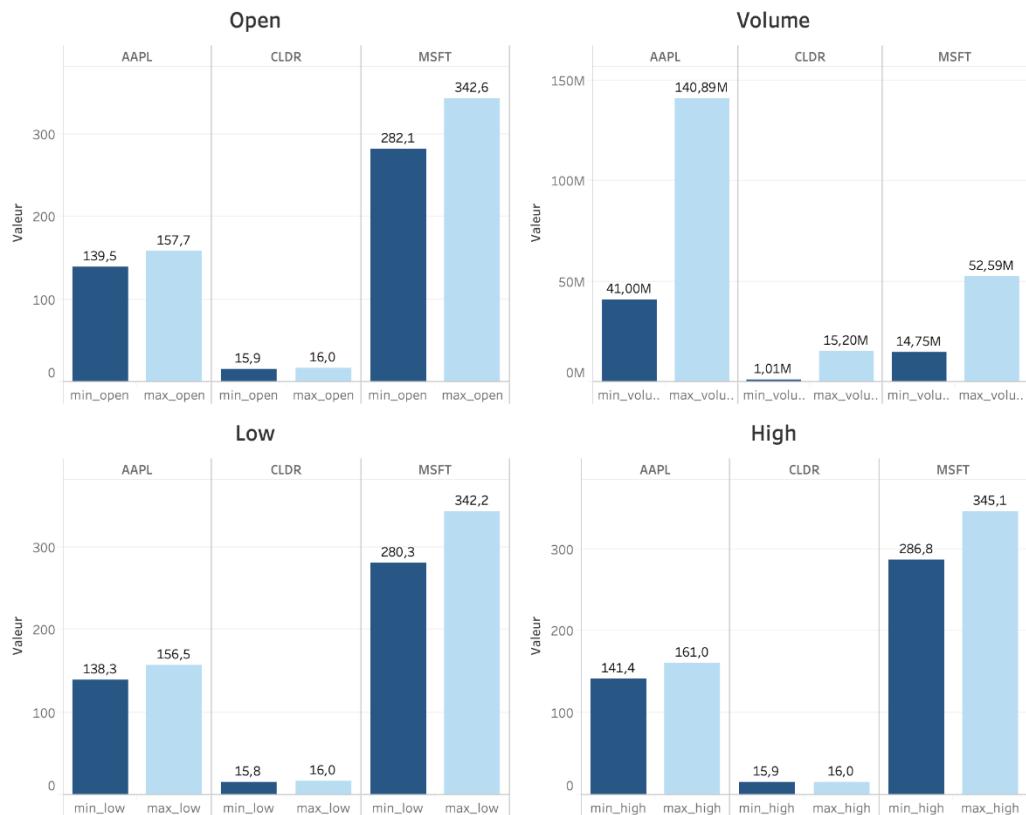
1) Tableau

ODBC connection

The screenshot shows the Tableau Data Source configuration window. At the top, it displays the connection name "market (stock.market) (IMPALA)" with options for "Connexion" (En direct) and "Filtres". Below this, there's a preview area showing a single table named "market_view". A message below the preview asks if additional tables are needed, with a note to "Faites glisser les tables ici pour les relier". The main pane shows the schema of the "market_view" table with columns: symbol, market_view_min_open, market_view_max_open, market_view_min_high, market_view_max_high, market_view_min_low, market_view_max_low, and market_view_min_volume. The data preview shows three rows: AAPL, MSFT, and CLDR, with their respective values for each column.

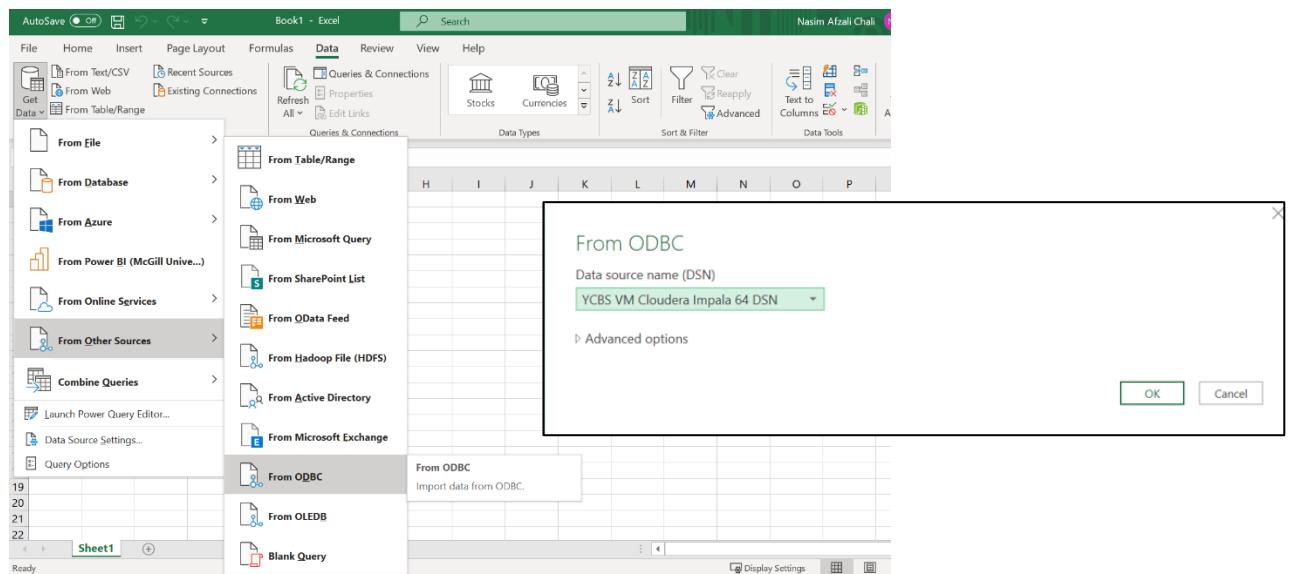
| symbol | market_view_min_open | market_view_max_open | market_view_min_high | market_view_max_high | market_view_min_low | market_view_max_low | market_view_min_volume |
|--------|----------------------|----------------------|----------------------|----------------------|---------------------|---------------------|------------------------|
| AAPL | 139,470 | 157,650 | 141,400 | 161,020 | 138,270 | 156,533 | 40 999 950 |
| MSFT | 282,122 | 342,640 | 286,770 | 345,100 | 280,250 | 342,200 | 14 751 610 |
| CLDR | 15,900 | 15,980 | 15,910 | 16,000 | 15,810 | 15,980 | 1 006 851 |

Charts



2) MS Excel

ODBC connection



Navigator

Select multiple items

Display Options

ODBC (dsn=YCBS VM Clou...)

IMPALA [5]

- bixi
- covid_19
- default
- market_stock
- stock [2]
- market
- market_view**

market_view

| min_open | max_open | min_high | max_high | min_low | max_low | min_volume | max_volume |
|----------|----------|----------|----------|---------|----------|------------|------------|
| 139.47 | 157.65 | 141.4 | 161.02 | 138.27 | 156.5328 | 40999950 | 140 |
| 15.9 | 15.98 | 15.91 | 16 | 15.81 | 15.98 | 1006851 | 15 |
| 282.1217 | 342.64 | 286.77 | 345.1 | 280.25 | 342.2 | 14751610 | 52 |

Select Related Tables

Load Transform Data Cancel

Charts

