In [ ]:
```
# Milestone 3 (Accessing Hive Data Warehouse)

Name : Rinashini a/p Arunasalam Sukormaru
Matric ID : WQD170077 (17013672/1)
Github Link :  https://github.com/RinashiniA/WQD7005-Group
```

In [6]:
```
# Installing hdfs, pyhive and thrift ( which have been installed)

!pip install hdfs
!pip install pyhive
!pip install thrift
```

```
Requirement already satisfied: hdfs in /Users/rinashiniarunasalam/
anaconda3/lib/python3.6/site-packages (2.5.8)
Requirement already satisfied: requests>=2.7.0 in /Users/rinashini
arunasalam/anaconda3/lib/python3.6/site-packages (from hdfs) (2.23
.0)
Requirement already satisfied: six>=1.9.0 in /Users/rinashiniaruna
salam/anaconda3/lib/python3.6/site-packages (from hdfs) (1.11.0)
Requirement already satisfied: docopt in /Users/rinashiniarunasala
m/anaconda3/lib/python3.6/site-packages (from hdfs) (0.6.2)
Requirement already satisfied: chardet<4,>=3.0.2 in /Users/rinashi
niarunasalam/anaconda3/lib/python3.6/site-packages (from requests>
=2.7.0->hdfs) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /Users/rinashiniaru
nasalam/anaconda3/lib/python3.6/site-packages (from requests>=2.7.
0->hdfs) (2.6)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.
21.1 in /Users/rinashiniarunasalam/anaconda3/lib/python3.6/site-pa
ckages (from requests>=2.7.0->hdfs) (1.22)
Requirement already satisfied: certifi>=2017.4.17 in /Users/rinash
iniarunasalam/anaconda3/lib/python3.6/site-packages (from requests
>=2.7.0->hdfs) (2018.4.16)
You are using pip version 19.0.3, however version 20.2b1 is availa
ble.
You should consider upgrading via the 'pip install --upgrade pip'
command.
Requirement already satisfied: pyhive in /Users/rinashiniarunasala
m/anaconda3/lib/python3.6/site-packages (0.6.2)
Requirement already satisfied: future in /Users/rinashiniarunasala
m/anaconda3/lib/python3.6/site-packages (from pyhive) (0.18.2)
Requirement already satisfied: python-dateutil in /Users/rinashini
arunasalam/anaconda3/lib/python3.6/site-packages (from pyhive) (2.
7.3)
Requirement already satisfied: six>=1.5 in /Users/rinashiniarunasa
lam/anaconda3/lib/python3.6/site-packages (from python-dateutil->p
yhive) (1.11.0)
You are using pip version 19.0.3, however version 20.2b1 is availa
ble.
You should consider upgrading via the 'pip install --upgrade pip'
command.
```

```
Collecting thrift
  Using cached https://files.pythonhosted.org/packages/97/1e/3284d
19d7be99305eda145b8aa46b0c33244e4a496ec66440dac19f8274d/thrift-0.1
3.0.tar.gz
Requirement already satisfied: six>=1.7.2 in /Users/rinashiniaruna
salam/anaconda3/lib/python3.6/site-packages (from thrift) (1.11.0)
Building wheels for collected packages: thrift
  Building wheel for thrift (setup.py) ... done
  Stored in directory: /Users/rinashiniarunasalam/Library/Caches/p
ip/wheels/02/a2/46/689ccfcf40155c23edc7cdbd9de488611c8fdf49ff34b17
06e
Successfully built thrift
Installing collected packages: thrift
Successfully installed thrift-0.13.0
You are using pip version 19.0.3, however version 20.2b1 is availa
ble.
You should consider upgrading via the 'pip install --upgrade pip'
command.
```

In [28]:
```
from io import BytesIO as StringIO
# For Data Lake
from hdfs import InsecureClient
# For Data Warehouse
from pyhive import hive

import pandas as pd


# To access the file stored on HDFS
with hdfs_interface.read('/wqd7005/source/000000_0', length=1024) a
s reader:
    content=reader.read()


#  Connecting to hive to access hivetables with python
host_name="localhost"
port=10000
conn=hive.Connection(host=host_name,port=port, auth='NOSASL')
cur=conn.cursor()


# Create External Table for source
cur.execute("DROP TABLE IF EXISTS source")
cur.execute("CREATE EXTERNAL TABLE IF NOT EXISTS \
            source( tdate STRING, \
                    closing_price DECIMAL(5,2), \
                    open DECIMAL(5,2), \
                    daily_high DECIMAL(5,2), \
                    daily_low DECIMAL(5,2)) \
            ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' \
            STORED AS TEXTFILE LOCATION '/wqd7005/source'")

# Accessing the file stored in the Hive data warehouse
cur.execute("SELECT * FROM source")
fetch_df=cur.fetchall()
df=pd.DataFrame(data=fetch_df)
print(df)
```

```
                              0       1       2       3       4
0                5/22/2020          33.25   33.95   34.00   30.72
1                5/21/2020          33.92   33.53   34.66   33.26
2                5/20/2020          33.49    None   33.78    None
3                5/19/2020          32.50   32.83   33.44   31.34
4                5/18/2020          31.82   29.53   33.32   29.53
...                        ...     ...     ...     ...     ...
3660              2/9/2006          62.62   62.68   63.73   62.38
3661              2/8/2006          62.55   62.96   63.44   62.29
3662              2/7/2006          63.09   64.82   64.90   62.81
3663              2/6/2006          65.11   66.35   66.50   64.77
3664              2/3/2006          65.37   64.77   65.48   63.93

[3665 rows x 5 columns]
```

In [32]:

```
## Data Cleaning

# Appending column names to the dataset
df.columns = ['Date', 'Closing Price', 'Open', 'Daily High', 'Daily
Low']
print(df.columns)
print(df)
```

```
Index(['Date', 'Closing Price', 'Open', 'Daily High', 'Daily Low']
, dtype='object')
                           Date Closing Price   Open Daily High D
aily Low
0                5/22/2020                33.25  33.95      34.00
30.72
1                5/21/2020                33.92  33.53      34.66
33.26
2                5/20/2020                33.49   None      33.78
None
3                5/19/2020                32.50  32.83      33.44
31.34
4                5/18/2020                31.82  29.53      33.32
29.53
...                           ...          ...    ...        ...
...
3660              2/9/2006                62.62  62.68      63.73
62.38
3661              2/8/2006                62.55  62.96      63.44
62.29
3662              2/7/2006                63.09  64.82      64.90
62.81
3663              2/6/2006                65.11  66.35      66.50
64.77
3664              2/3/2006                65.37  64.77      65.48
63.93

[3665 rows x 5 columns]
```

In [34]:
```
# Viewing the missing values in the dataset

print(df.shape)
print(df[1:10])
```

```
(3665, 5)
                        Date Closing Price   Open Daily High Dail
y Low
1           5/21/2020                33.92  33.53      34.66
33.26
2           5/20/2020                33.49   None      33.78
None
3           5/19/2020                32.50  32.83      33.44
31.34
4           5/18/2020                31.82  29.53      33.32
29.53
5           5/15/2020                29.43  27.64      29.92
27.24
6           5/14/2020                27.56  25.56      27.96
25.18
7           5/13/2020                25.29  25.30      26.45
24.79
8           5/12/2020                25.78  24.49      26.23
24.22
9           5/11/2020                24.14  24.49      25.58
23.67
```

In [36]:
```
## Number of missing values in each column

col_missing = df.isnull().sum()
print(col_missing)
```

```
Date             0
Closing Price    0
Open            62
Daily High      19
Daily Low       13
dtype: int64
```

In [37]:
```
## Converting the first column to datatime format and the values to
numeric form

df['Date'] = pd.to_datetime(df['Date'])
df['Closing Price'] = pd.to_numeric(df['Closing Price'])
df['Open'] = pd.to_numeric(df['Open'])
df['Daily High'] = pd.to_numeric(df['Daily High'])
df['Daily Low'] = pd.to_numeric(df['Daily Low'])
print(df.dtypes)
```

```
Date             datetime64[ns]
Closing Price           float64
Open                    float64
Daily High              float64
Daily Low               float64
dtype: object
```

In [43]:
```
## Obtaining the mean values of each column to impute the missing v
alues with its mean values

df_None = df.dropna()
Open_mean_value = round(df_None['Open'].mean(),2)
df['Open'] = df['Open'].fillna(Open_mean_value)

Daily_High_mean_value = round(df_None['Daily High'].mean(),2)
df['Daily High'] = df['Daily High'].fillna(Daily_High_mean_value)

Daily_Low_mean_value = round(df_None['Daily Low'].mean(),2)
df['Daily Low'] = df['Daily Low'].fillna(Daily_Low_mean_value)

print(df[1:10])
```

```
        Date  Closing Price   Open  Daily High  Daily Low
1 2020-05-21          33.92  33.53       34.66      33.26
2 2020-05-20          33.49  72.61       33.78      71.44
3 2020-05-19          32.50  32.83       33.44      31.34
4 2020-05-18          31.82  29.53       33.32      29.53
5 2020-05-15          29.43  27.64       29.92      27.24
6 2020-05-14          27.56  25.56       27.96      25.18
7 2020-05-13          25.29  25.30       26.45      24.79
8 2020-05-12          25.78  24.49       26.23      24.22
9 2020-05-11          24.14  24.49       25.58      23.67
```

In [44]:
```
# Exporting the cleaned dataset to a csv file

df.to_csv(r'dataset_cleaned.csv', index=False)
```