Final Exam: Question 1

1. You are required to make a user-agent that will crawl the WWW (your familiar domain) to produce dataset of a particular website.
    - the web site can be as simple as a list of webpages and what other pages they link to
    - the output does not need to be in XHTML (or HTML) form a multi-stage approach (e.g. produce the xhtml or html in csv format)

Step 1 : Data Scraping – crawl the trading date and closing price of the following Index and Stocks

- Maybank –MBB
- RHB
- KLCI – Kuala Lumpur Composite Index
- DJI – Dow Jone Index
- SNP - S&P 500 Index

```
import requests
import bs4
import json
import datetime
import pandas as pd
page = requests.get("https://finance.yahoo.com/quote/MBBM.KL/history?period1=1425945600&period2=1583798400&interval=1d&filter

from bs4 import BeautifulSoup
soup = BeautifulSoup(page.content, 'html.parser')
#print(soup.prettify())

listScript = soup.find_all("script")
for script in listScript:
    txtScript = script.string
    if type(txtScript) is bs4.element.NavigableString and txtScript.find('HistoricalPriceStore') != -1:
        MBB = pd.DataFrame(columns=['date','close'])
        txtInfo = txtScript[txtScript.find('HistoricalPriceStore'):txtScript.find('}],"isPending":false,"firstTradeDate":')]
        txtInfo = "{\"" + txtInfo + "}]}}"
        objJson = json.loads(txtInfo)
        for price in objJson['HistoricalPriceStore']['prices']:
            txtDate = ""
            txtClose = ""
            for attr,val in price.items():
                if attr == "date" and val != None:
                    txtDate = datetime.datetime.fromtimestamp(val).strftime('%Y-%m-%d')
                if attr == "close" and val != None:
                    txtClose = str(val)
            if txtDate != "" and txtClose != "":
                MBB = MBB.append({"date":txtDate, "close":txtClose}, ignore_index=True)
        print(MBB)
```

Step 1 : Data Scraping – crawl the trading date and closing price of the following Index and Stocks

MBB

	date	close
0	2020-03-09	8.239999771118164
1	2020-03-06	8.5
2	2020-03-05	8.529999732971191
3	2020-03-04	8.470000267028809
4	2020-03-03	8.40999984741211
...	...	...
1237	2015-03-16	9.09000015258789
1238	2015-03-13	9.100000381469727
1239	2015-03-12	9.119999885559082
1240	2015-03-11	9.09000015258789
1241	2015-03-10	9.229999542236328

[1242 rows x 2 columns]

CIMB

	date	close
0	2020-03-09	4.199999809265137
1	2020-03-06	4.440000057220459
2	2020-03-05	4.5
3	2020-03-04	4.539999961853027
4	2020-03-03	4.5
...	...	...
1238	2015-03-16	5.909999847412109
1239	2015-03-13	5.920000076293945
1240	2015-03-12	5.849999904632568
1241	2015-03-11	5.800000190734863
1242	2015-03-10	5.949999809265137

[1243 rows x 2 columns]

RHB

	date	close
0	2020-03-09	5.46999979019165
1	2020-03-06	5.699999809265137
2	2020-03-05	5.710000038146973
3	2020-03-04	5.710000038146973
4	2020-03-03	5.639999866485596
...	...	...
1238	2015-03-16	7.265110015869141
1239	2015-03-13	7.199659824371338
1240	2015-03-12	7.237060070037842
1241	2015-03-11	7.349259853363037
1242	2015-03-10	7.386660099029541

[1243 rows x 2 columns]

KLCI

	date	close
0	2020-03-09	1424.1600341796875
1	2020-03-06	1483.0999755859375
2	2020-03-05	1491.030029296875
3	2020-03-04	1489.949951171875
4	2020-03-03	1478.6400146484375
...	...	...
1221	2015-03-16	1780.5400390625
1222	2015-03-13	1781.75
1223	2015-03-12	1786.8699951171875
1224	2015-03-11	1778.1600341796875
1225	2015-03-10	1789.72998046875

[1226 rows x 2 columns]

DJI

	date	close
0	2020-03-09	23851.01953125
1	2020-03-06	25864.779296875
2	2020-03-05	26121.279296875
3	2020-03-04	27090.859375
4	2020-03-03	25917.41015625
...	...	...
1254	2015-03-16	17977.419921875
1255	2015-03-13	17749.310546875
1256	2015-03-12	17895.220703125
1257	2015-03-11	17635.390625
1258	2015-03-10	17662.939453125

[1259 rows x 2 columns]

SNP

	date	close
0	2020-03-09	2746.56005859375
1	2020-03-06	2972.3701171875
2	2020-03-05	3023.93994140625
3	2020-03-04	3130.1201171875
4	2020-03-03	3003.3701171875
...	...	...
1254	2015-03-16	2081.18994140625
1255	2015-03-13	2053.39990234375
1256	2015-03-12	2065.949951171875
1257	2015-03-11	2040.239990234375
1258	2015-03-10	2044.1600341796875

[1259 rows x 2 columns]

Step 2 : Data Preprocessing

1. Using the following code to create new attributes as follows:

Type of index

```
MBB['Type_of_Index'] = 'Malaysia Stock'
CIMB['Type_of_Index'] = 'Malaysia Stock'
RHB['Type_of_Index'] = 'Malaysia Stock'
KLCI['Type_of_Index'] = 'Malaysia Index'
DJI['Type_of_Index'] = 'US Index'
SNP['Type_of_Index'] = 'US Index'
```

Sector

```
MBB['Sector'] = 'Financial'
CIMB['Sector'] = 'Financial'
RHB['Sector'] = 'Financial'
KLCI['Sector'] = 'Top 30 Malaysia stock by Market Cap'
DJI['Sector'] = 'Top 30 US stock by Market Cap'
SNP['Sector'] = 'Top 500 US stock by Market Cap'
```

Market Board

```
MBB['Market board'] = 'Main Board'
CIMB['Market board'] = 'Main Board'
RHB['Market board'] = 'Main Board'
KLCI['Market board'] = 'Kuala Lumpur Composite Index'
DJI['Market board'] = 'US Stock Excahnge'
SNP['Market board'] = 'US Stock Excahnge'
```

Country

```
MBB['Country'] = 'MY'
CIMB['Country'] = 'MY'
RHB['Country'] = 'MY'
KLCI['Country']= 'MY'
DJI['Country'] = 'US'
SNP['Country'] = 'US'
```

Month/ Year Column

```
MBB['Month_Year'] = pd.to_datetime(MBB['date']).dt.to_period('M')
CIMB['Month_Year'] = pd.to_datetime(CIMB['date']).dt.to_period('M')
RHB['Month_Year'] = pd.to_datetime(RHB['date']).dt.to_period('M')
KLCI['Month_Year'] = pd.to_datetime(KLCI['date']).dt.to_period('M')
DJI['Month_Year'] = pd.to_datetime(DJI['date']).dt.to_period('M')
SNP['Month_Year'] = pd.to_datetime(SNP['date']).dt.to_period('M')
```

Name of stock and index

```
MBB['Name of Stock/ Index'] = 'MBB'
CIMB['Name of Stock/ Index'] = 'CIMB'
RHB['Name of Stock/ Index'] = 'RHB'
KLCI['Name of Stock/ Index'] = 'KLCI'
DJI['Name of Stock/ Index'] = 'DJI'
SNP['Name of Stock/ Index'] = 'SNP'
```

Index type

```
MBB['Index_Type'] = 'Stock'
CIMB['Index_Type'] = 'Stock'
RHB['Index_Type'] = 'Stock'
KLCI['Index_Type']= 'Stock'
DJI['Index_Type'] = 'Index'
SNP['Index_Type'] = 'index'
```

Monthly average of close price

```
MBB['Monthly_Average'] = MBB['close'].groupby(MBB['Month_Year']).transform('mean')
CIMB['Monthly_Average'] = CIMB['close'].groupby(CIMB['Month_Year']).transform('mean')
RHB['Monthly_Average'] = RHB['close'].groupby(RHB['Month_Year']).transform('mean')
KLCI['Monthly_Average'] = KLCI['close'].groupby(KLCI['Month_Year']).transform('mean')
DJI['Monthly_Average'] = DJI['close'].groupby(DJI['Month_Year']).transform('mean')
SNP['Monthly_Average'] = SNP['close'].groupby(SNP['Month_Year']).transform('mean')
```

Step 2 : Data Preprocessing

2. Data Cleaning :

Formatted date

```
MBB['date'] = pd.to_datetime(MBB.date)
CIMB['date'] = pd.to_datetime(CIMB.date)
RHB['date'] = pd.to_datetime(RHB.date)
KLCI['date'] = pd.to_datetime(KLCI.date)
DJI['date'] = pd.to_datetime(DJI.date)
SNP['date'] = pd.to_datetime(SNP.date)
```

Convert the data type (close price) to float

```
MBB['close'] = MBB['close'].astype(float)
CIMB['close'] = CIMB['close'].astype(float)
RHB['close'] = RHB['close'].astype(float)
KLCI['close']= KLCI['close'].astype(float)
DJI['close'] = DJI['close'].astype(float)
SNP['close'] = SNP['close'].astype(float)
```

3. Data Integration and rename the column name

```
DF = pd.concat([MBB, CIMB, RHB, KLCI, DJI, SNP], ignore_index=True)
DF.rename(columns={'date':'Trading_Date',
                   'close':'Closing_Index/Price'},
          inplace=True)
```

4. Check if data is cleaned

```
print(DF)
DF.isnull().sum()
```

```
Trading_Date           0
Closing_Index/Price    0
Type_of_Index          0
Sector                 0
Market board           0
Month_Year             0
Name of Stock/ Index   0
Monthly_Average        0
Country                0
Index_Type             0
dtype: int64
```

Step 2 : Data Preprocessing

Data frame in pandas post data preprocessing

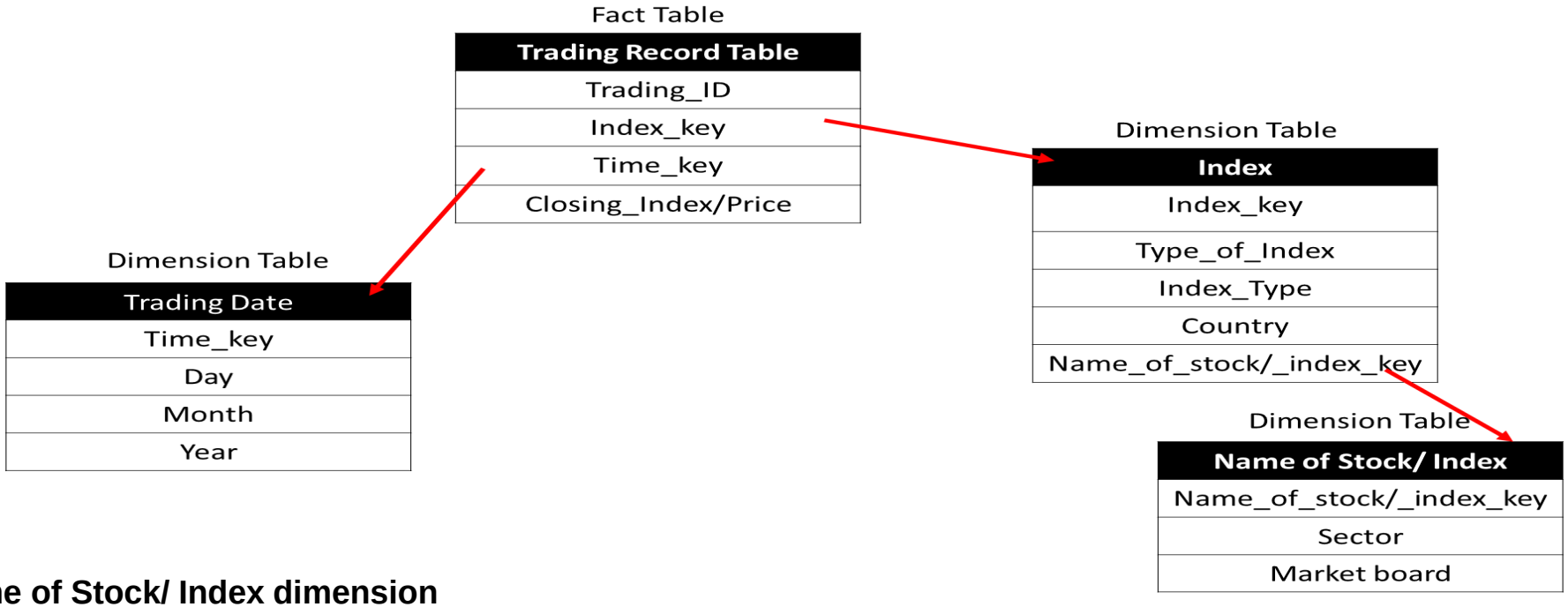
```
print(DF)
```

	Trading_Date	Closing_Index/Price	Type_of_Index	Sector	Market board	Month_Year	Name of Stock/ Index	Monthly_Average	Country	Index_Type
0	2020-03-09	8.240000	Malaysia Stock	Financial	Main Board	2020-03	MBB	8.423333	MY	Stock
1	2020-03-06	8.500000	Malaysia Stock	Financial	Main Board	2020-03	MBB	8.423333	MY	Stock
2	2020-03-05	8.530000	Malaysia Stock	Financial	Main Board	2020-03	MBB	8.423333	MY	Stock
3	2020-03-04	8.470000	Malaysia Stock	Financial	Main Board	2020-03	MBB	8.423333	MY	Stock
4	2020-03-03	8.410000	Malaysia Stock	Financial	Main Board	2020-03	MBB	8.423333	MY	Stock
...	...	...	...	...	...	...	...	...	...	...
7467	2015-03-16	2081.189941	US Index	Top 500 US stock by Market Cap	US Stock Excahnge	2015-03	SNP	2074.022484	US	index
7468	2015-03-13	2053.399902	US Index	Top 500 US stock by Market Cap	US Stock Excahnge	2015-03	SNP	2074.022484	US	index
7469	2015-03-12	2065.949951	US Index	Top 500 US stock by Market Cap	US Stock Excahnge	2015-03	SNP	2074.022484	US	index
7470	2015-03-11	2040.239990	US Index	Top 500 US stock by Market Cap	US Stock Excahnge	2015-03	SNP	2074.022484	US	index
7471	2015-03-10	2044.160034	US Index	Top 500 US stock by Market Cap	US Stock Excahnge	2015-03	SNP	2074.022484	US	index

7472 rows × 10 columns

Final Exam: Question 2

Draw snowflake schema diagram for the above dataset. Justify your attributes to be selected in the respective dimensions.



Index and Name of Stock/ Index dimension

Index dimension table contains Type_of_Index, Index_Type, Country and Name_of_stock/_index_key, the reason that I selected the attributes is because if I read the attributes (dataset) from this schema, only these attributes will be read, I will have a glance of what type_of_index,Index_type, Country and Name_of_stock/_index being stored in this disk. Sector and Market Board attribute are not important attribute to be read so I store it in a separate dimension namely Name of Stock/ Index dimension with the given key to retrieve the data only when it needed. In addition, it requires low disk storage and it will be faster when it is queried.

Trading Date Dimension

Trading date dimension store the day, month and year data. The reason is to ensure data storage and integrity issue when the format of the date is changed, one the date format is changed in this dimension, trading date will be consistent / standardized across all the reads.