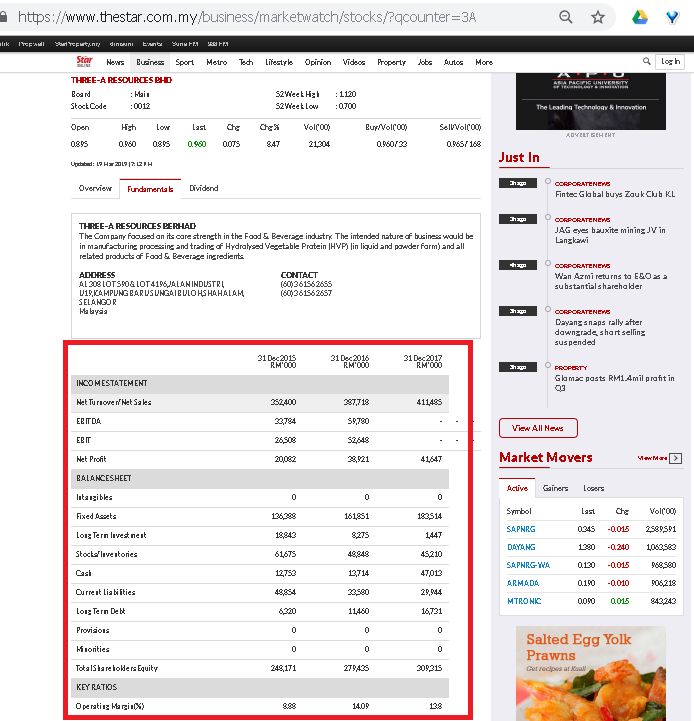
*2. Management of data (Group)*

*Pulling together large volumes of data and merging multiple sources (such as web information (stock index), social media, whatapp and news ) to create the one big analysis record for analytics. And the team should expect to do this process over and over again until a useful set of data is prepared.*

Video: https://drive.google.com/open?id=1ebVQ6jcUCeVMoXBrYzUvyskkiNflNauA

To implements this, there are steps have been done:

1. Pulling data from multiple sources:
2. Company background

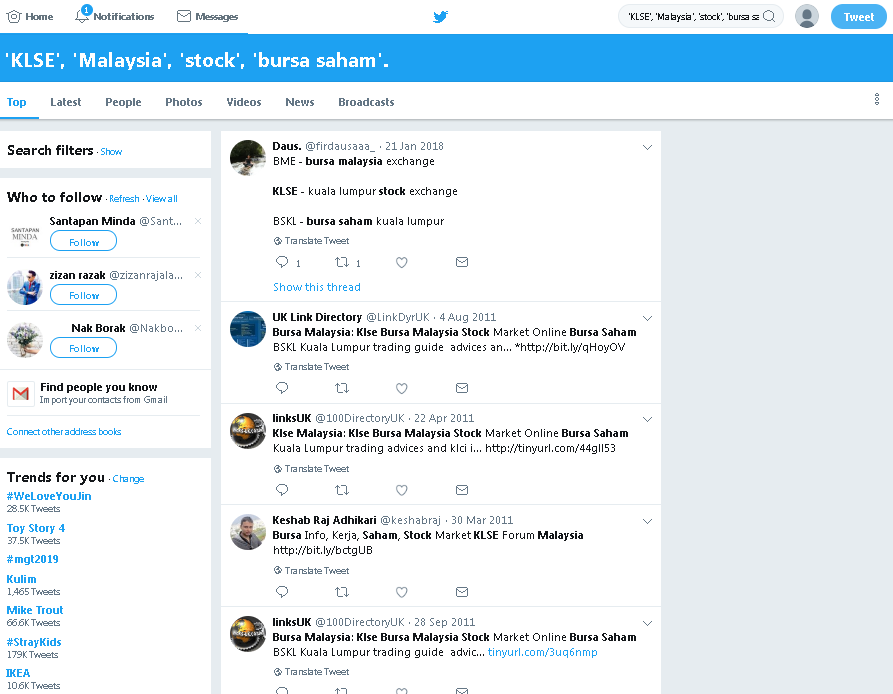


The company background/fundamental will be a factor of the company stock activities.

The python program is ‘StockFundamental\_crawl.py’.

It will refer to the compcode.txt and loop all the company code to pull fundamental info and save the data into ‘fundamental’ table in mysql database.

1. Twitter

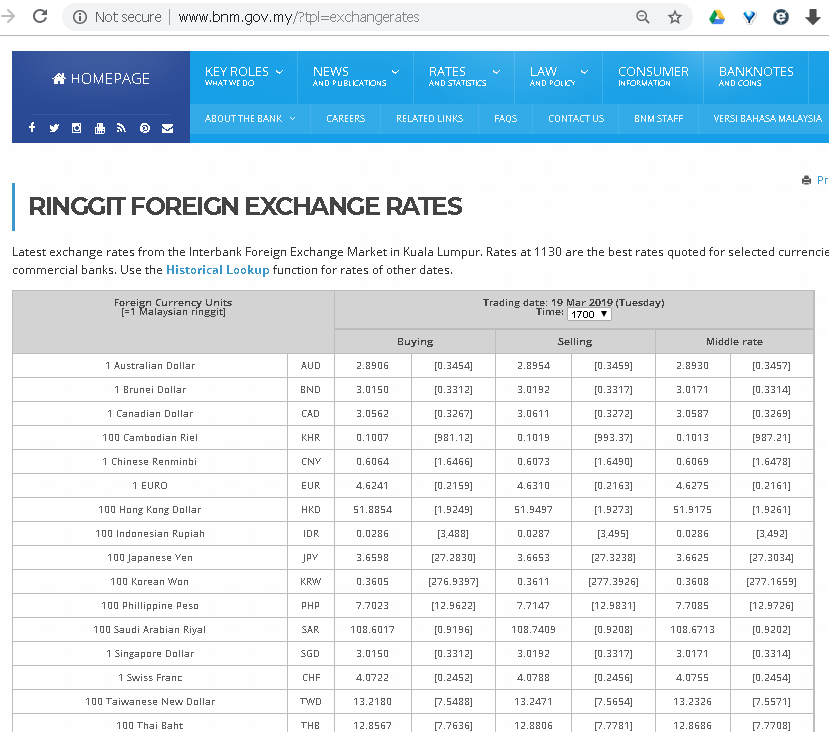


The responses / tweets related to the stock market might bring hints to the stock market.

The python program for pulling data is ‘TweetKLSEAPI.py’.

It query the tweets is focusing on the 'KLSE', 'Malaysia', 'stock', 'bursa saham'.  
All the result will be pulled and saved into ‘tweetsdata’ in mysql database

1. Daily Currency Rate



The currency rate with Malaysia Ringgit with same other main currency also a factor to the stock market activity.

1. News

The news reported also will be a factor for the market activity. The news reports what is happing in the country.

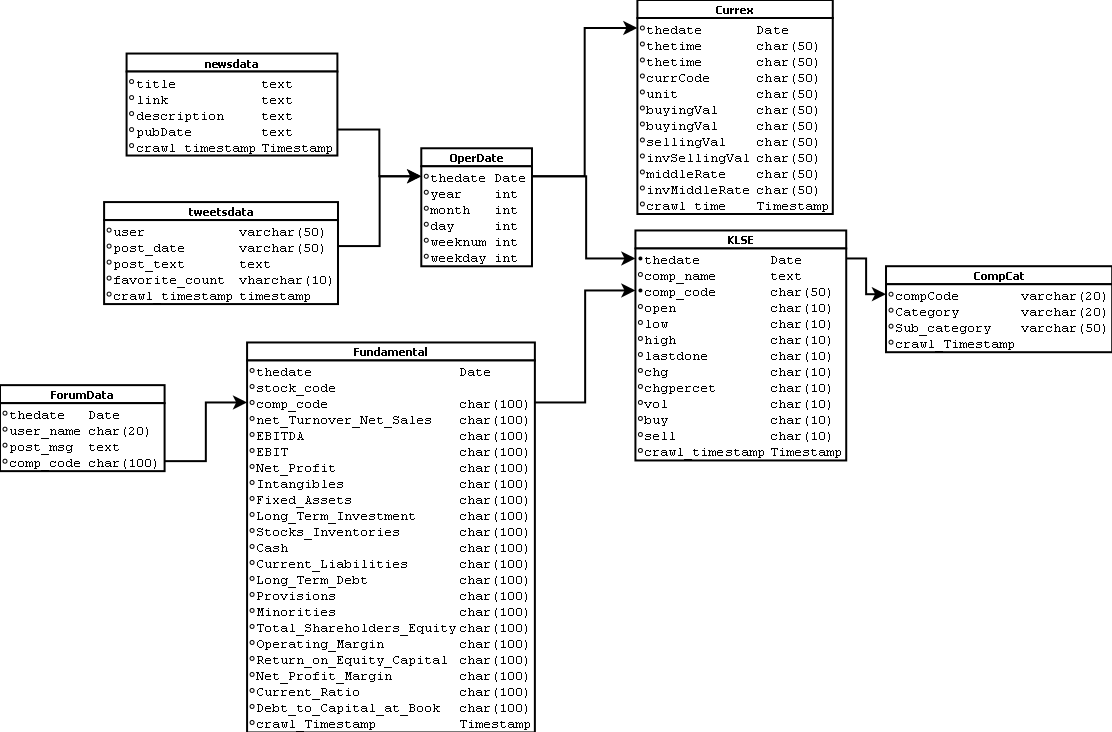
The python program for fulling data is ‘NewsFeedAccess.py’.

It query the news feed which is related to stock market such as The Edge and MalaysiaKini. The headline of the news will be pulled and save into ‘newsdata’ table mysql database. The feeds to be pulled:  
https://www.theedgemarkets.com/mytopstories.rss

https://www.malaysiakini.com/en/news.rss

1. The schema

Snowflake schema for the stock data are as below:



The description of the tables:

|  |  |
| --- | --- |
| Table | Description |
| KLSE | The daily stock market opening and closing market values. |
| Fundamental | The financial background of the companies. |
| CompCat | Company category info. |
| CurrEx | Daily current exchange data. |
| OperDate | Operational date |
| NewsData | Data pulled from the news feed. |
| TweetsData | Data pulled from twitter social websites. |
| ForumData | Data pulled from web forum pages. |

1. Setup Hadoop and Hive
2. Export data from MYSQL

There are files exported:

* compcat.txt (Data related to company category)
* currex2.txt (Data related to currency exchange)
* fundamental.txt (Data related to company background)
* KLSE2.txt (Data related to KLSE market)
* tweetsdata.txt (Data related )

1. Put data files to HDFS

The command to put the file to HDFS are as below:  
  
hadoop fs -put KLSE2.txt /user/datafile  
hadoop fs -put fundamental.txt /user/datafile  
hadoop fs -put tweetsdata.txt /user/datafile  
hadoop fs -put currEx2.txt /user/datafile

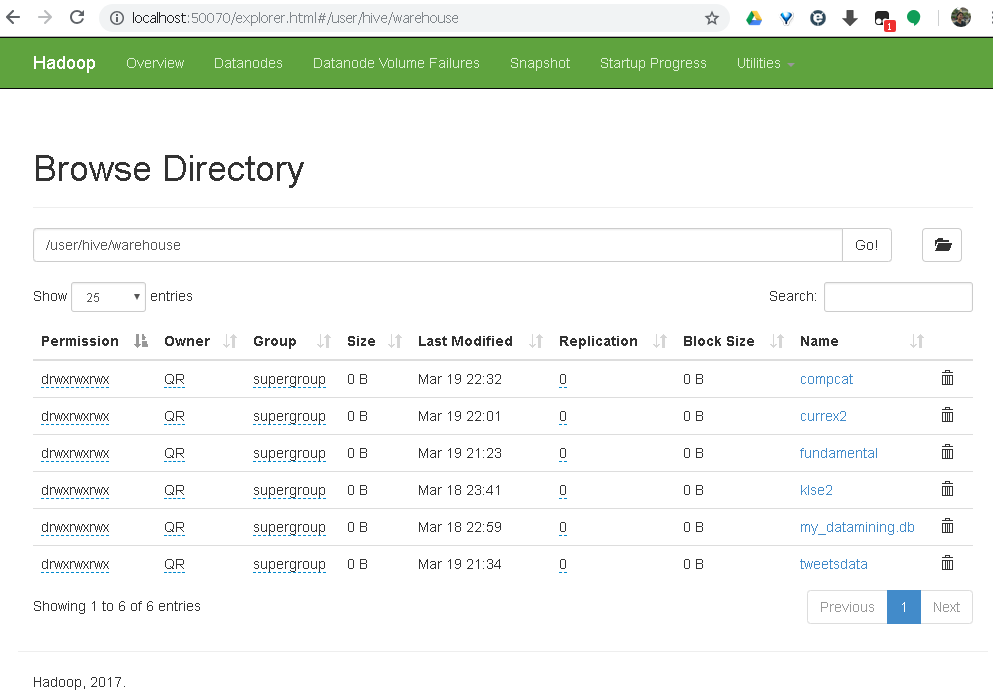
All the files will be put to /user/datafile directory in HDFS

1. Create tables in Hive and load the data files into Apache Hive

To have data stored in Hive, some tables have to be created.  
There are tables are created:

* KLSE2
* Fundamental
* tweetsData
* currEx2

Then, load the data files into the tables :  
  
LOAD DATA INPATH '/user/datafile/KLSE2.csv' INTO TABLE KLSE2;  
LOAD DATA INPATH '/user/datafile/fundamental.txt' INTO TABLE fundamental;  
LOAD DATA INPATH '/user/datafile/tweetsdata.txt' INTO TABLE tweetsData;  
LOAD DATA INPATH '/user/datafile/currEx2.txt' INTO TABLE currEx2;

After the data loaded into Hive, the information of the hive can be found in HDFS page:  


For listing Hive tables, they can be listed by   
Hive> LIST TABLES;

