MILESTONE 2 REPORT: PULLING DATA FROM DIFFERENT SOURCES AND PUT THEM INTO APACHE HADOOP AND APACHE HIVE

MUHAMAD HAFIZ B HUSSIN

WQD170080

LINK TO MY VIDEO: <https://youtu.be/aY60X5W_R1s>

ATTACHED IS ALSO THE SOURCE CODE FOR PYTHON PROGRAMME

From the previous milestone, we are required to use web crawler to gather real time data from the internet. For this assignment, we are required to use the web crawlers that are developed in the previous task and crawled more data for the datawarehouse.

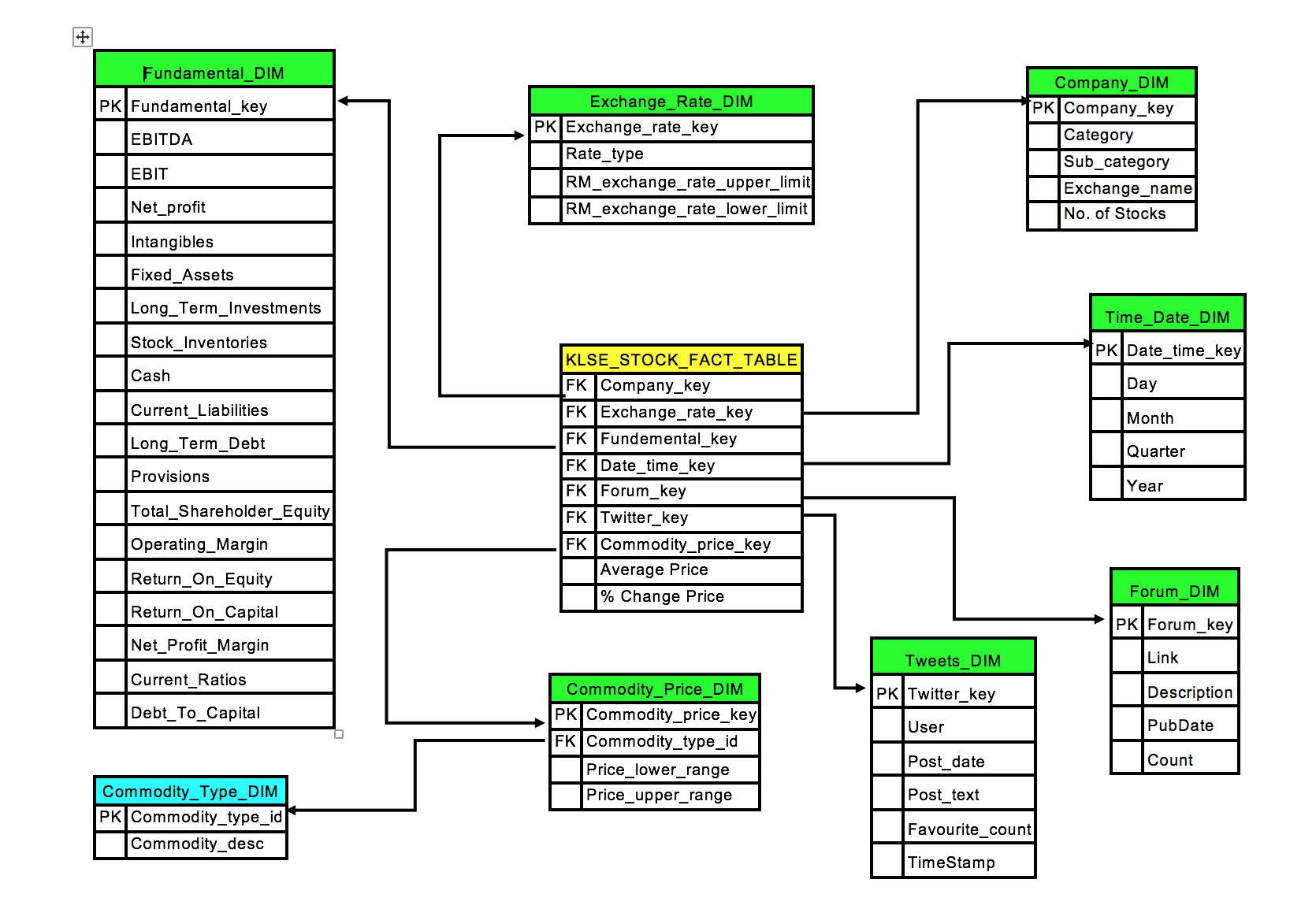
PART 1  
In order to archieve the objective of this assignment, I have developed a snowflake schema to guide me to plan for buiding the datawarehouse for KLSE stocks. Diagram 1 below is my snowflake schema.  
  


Diagram 1: KLSE STOCK Snowflake Schema

I have a KLSE Stock Fact Table as the base and 7 dimension tables namely:

1. Currency Exchange Table

2. Company Details Table

3. Social media (Forum) Table

4. Tweets Data Table

5. Time and Date Table

6. Stock Fundamental Table

7. Commodity Table

Commodity Table is further devide into Commodity Type Table. There are two commodities that might important for determining KLSE Stocks price. There are:

1. Palm Oil Price

2. Crude Oil Price

Based on the schema, I gathered data from 8 different sources to complete this assignment. I listed the tables, the link where I get the data and the python file that the crawlers were implemented. All the data are stored in Postgrest Database and CSV files.

1. Currency Exchange Table –

<http://www.bnm.gov.my/?tpl=exchangerates>

ForexRM.py

2. Company Details Table

<https://www.thestar.com.my/business/marketwatch/stock-list/>

stock\_crawl.py

3. Social media (Forum) Table

<https://klse.i3investor.com/jsp/scl/forum.jsp?fp=1&c=1>

ForumKLSE.py

4. Tweets Data Table

<https://twitter.com/login>

TweetKLSEAPI.py

5. Time and Date Table

created by the database

6. Stock Fundamental Table

provided by team member

7. Palm Oil Price

<http://www.mpoc.org.my/dailypalmoilprices.aspx?catID=b4ad7d4e-d7d0-410b-be86-9a80af0f4693&print=&ddlID=28abbe06-f695-4fd0-8bde-85d8a2ee9ccd>

CSV is available to be downloaded

8. Crude Oil Price

<https://www.macrotrends.net/2516/wti-crude-oil-prices-10-year-daily-chart>

CSV is available to be downloaded

PART 2

For the second part of this milestone, we are required to migrate the data into hadoop hdfs files. I used virtual box to install Ubuntu 16.04 Operating System in order to run Hadoop. I installed Apache Hadoop 2.7.7 and Apache Hive 2.1.0.

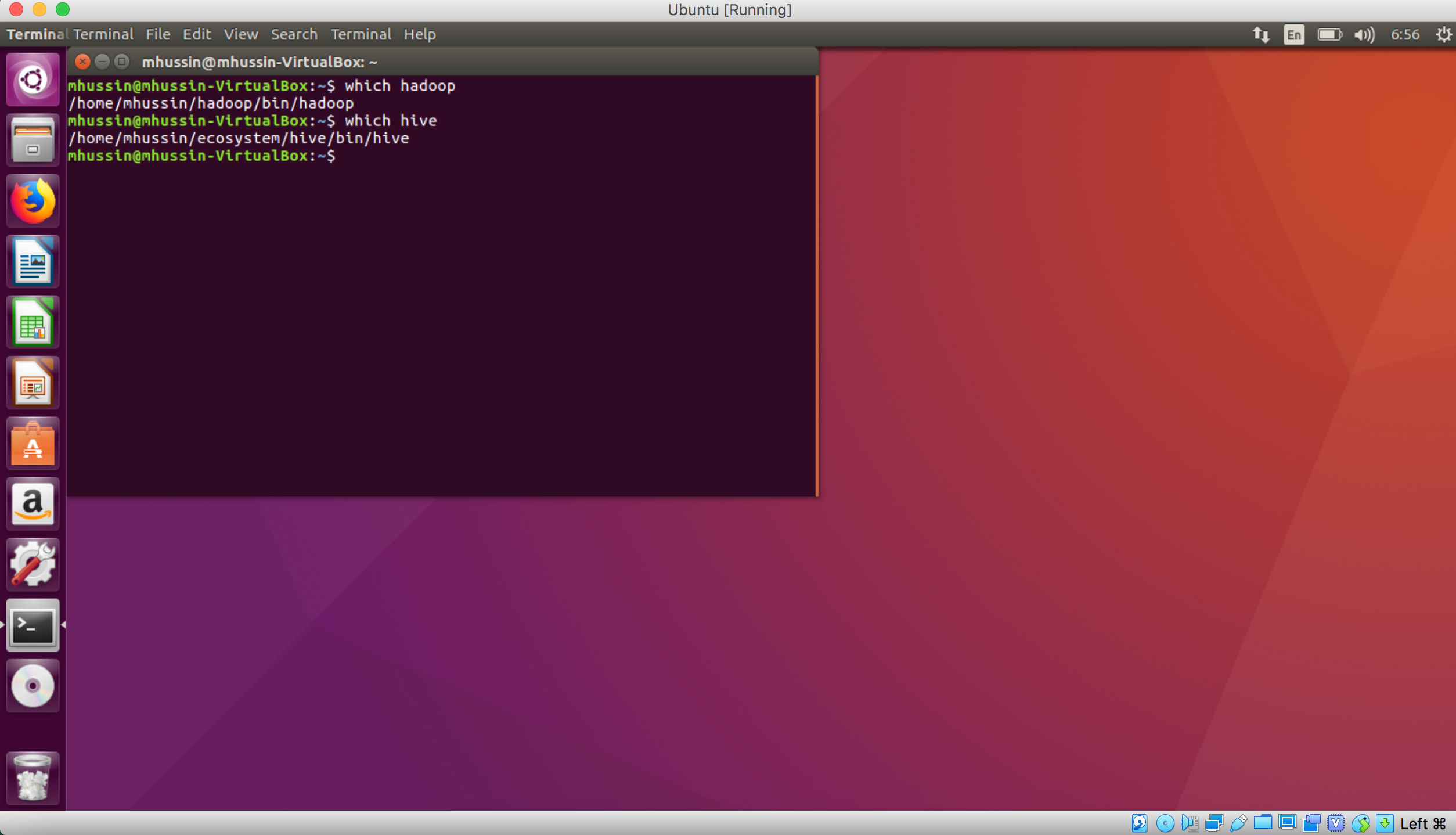


Diagram 2: Virtual box Ubuntu

After all the installation processes, I can run hadoop by start-all.sh before running Hive. In order to store data into HDFS files, I transfered all the csv files data that I collected from the crawling process from the host OS to Ubuntu. I stored them all into my\_input folder in the home directory.

First, I created all temporary folder to store the output of hdfs and hive. The folder paths are:

hdfs: mydata/mycsv/id

hive: /user/hive/warehouse

Second, I moved all the csv files into hdfs format using this command

hdfs dfs –put /home/mhussin/my\_input/\*.csv mydata/mycsv/id

Third, I started hadoop and hive. I define the tables for my hive and the load all the data into Hive tables using the command:

LOAD DATA INPATH ‘mydata/mycsv/id/file.csv’ INTO TABLE xxx;

The output of the hive tables are shown in Diagram 3 below:

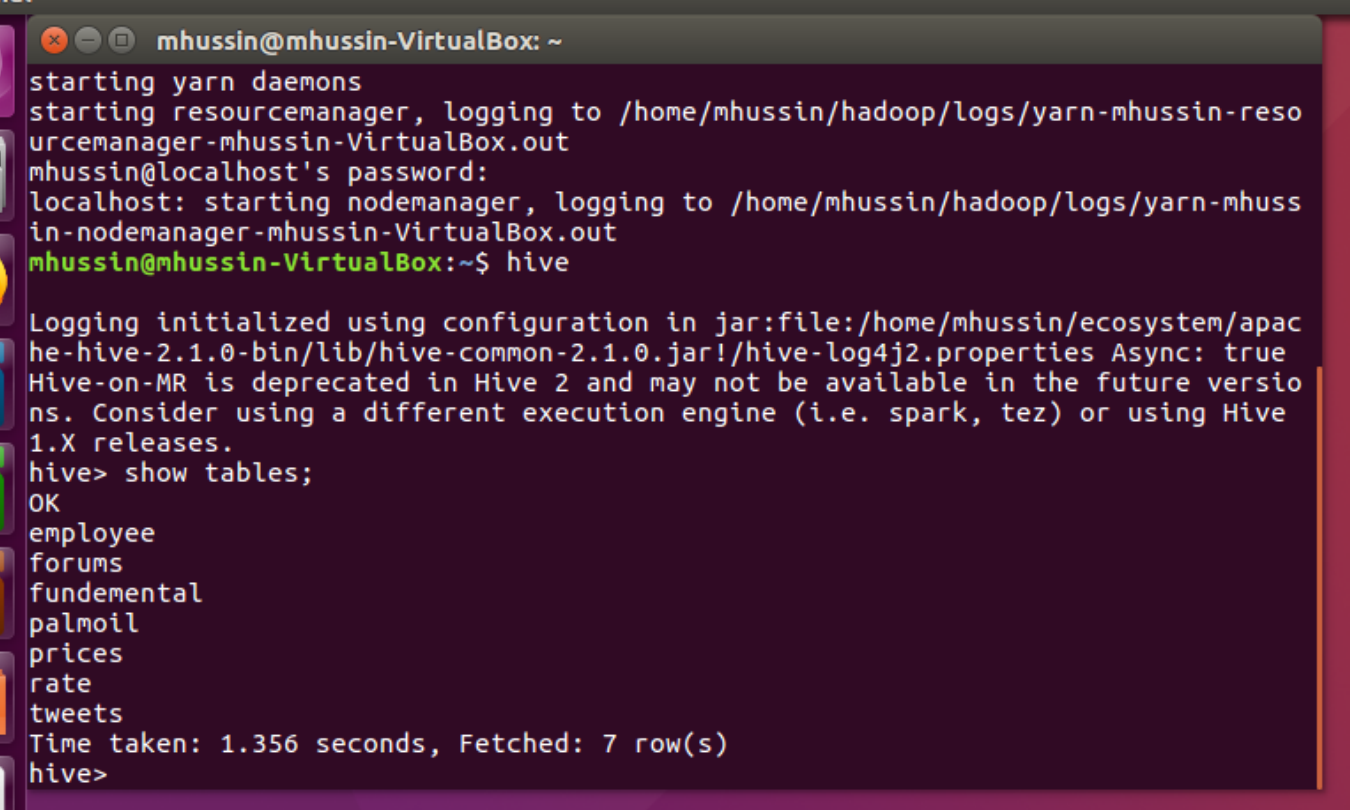


Diagram 3: Hive Tables output