**Name : Jeremy Chin**

**Matric Number : WQD170033**

**Course : WQD 7005 Data Mining**

**Group Members :**

WQD180063 Yong Wai Shin

WQD180069 Yap Boon Hui

WQD170033 Jeremy Chin

**Assignment Milestone 2**

The table below shows the sources of data, brief description of data, scripts to scrap and process the data, output files and attributes of the output files.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | Description | Script | File(s) | Attributes in File(s) |
| The Star  website | Financial results | 20190320-crawl-thestar3.py | stock-financial-results.txt | * Stock symbol * Board type * Stock name * Stock code * Fundamental * Date announced * Financial year end * QTR * Revenue RM(0,000) * P&L RM(,000) * EPS |
| Malaysiastock.biz website | Daily stock price from 3 March to 19 March 2019 | 20190302-crawl-msiastockbiz.py | stock\_20190303.txt  stock\_20190304.txt  stock\_20190305.txt  stock\_20190306.txt  stock\_20190307.txt  stock\_20190308.txt  stock\_20190311.txt  stock\_20190312.txt  stock\_20190313.txt  stock\_20190314.txt  stock\_20190315.txt  stock\_20190318.txt  stock\_20190319.txt | * Stock code * Stock name * Stock ref * Open price * Last price * Change * Change (%) * Volume |
| The Edge website | Stock news headlines | edgecrawler.py | edgenews\_row\_wise.csv | * News |
| WhatsApp | Stock highlights shared by brokers | whatsapp.py | whatsapp.csv | * Date time * Sender * Keyword(s) * Message |

The files containing the scrapped and curated data were imported into Apache Hive tables using ‘*hive -f /home/waishin/hive\_command’* command in Linux Ubuntu’s terminal.