Link to GitHub: https://github.com/SidratulMuntaha/DataMIning.git

Video link:

Milestone 2: Management of data : Store data into hive data warehouse or store data into data lake.

Due Date: Week 5

**Title: Text mining on customer feedbacks of different laptop brands**

**Step 1:**

**Data Source:**

real time data from reddit social media is scraped using python. Six Subreddit communities : SuggestAlaptop,Dell,AcerOfficial,macos,Asus,thinkpad were chosen. Two csv file was created. One is with the comments and id.

A screenshot of a cell phone

Description automatically generated

Fig1:reddit\_comment.csv

The second one is with 8 columns :

A screenshot of a computer

Description automatically generated  
 Fig2:reddit.csv

**Step 2:**

Data Cleaning: The acquired dataset from reddit have been cleaned using R language in Rstudio. The screenshot of cleaned dataset is given below:

A screenshot of a cell phone

Description automatically generated Fig3: clean\_reddit\_comments.csv

A screenshot of a cell phone

Description automatically generated Fig4:clean\_Reddit.csv

**Step 3:**

Store Data in Hive data warehouse:

**Apache Hive** is a data warehouse infrastructure that facilitates querying and managing large data sets which resides in distributed storage system. It is built on top of Hadoop and developed by Facebook. **Hive** provides a way to query the data using a SQL-like query language called **HiveQL(Hive query Language).** Internally, a compiler translates **HiveQL** statements into **MapReduce** jobs, which are then submitted to **Hadoop framework** for execution.

1. In order to store data in hive, Hadoop environment is needed. Through Virtual machine ubuntu operating system has been installed and then in ubuntu Hadoop and hive is installed respectably.
2. First, with start-all.sh command the Hadoop environment is set up, and then ‘jps’ command which is a tool to check, whether expected Hadoop processes are up and in running state or not.

A screenshot of a cell phone

Description automatically generated

1. Creating a directory in named, ‘milestone\_2/input/’

A picture containing drawing, food

Description automatically generated

1. The datasets are placed at the desired directory.

A close up of a sign

Description automatically generated

A close up of a sign

Description automatically generated

1. Initializing hive.

A picture containing bottle, table, phone, white

Description automatically generated

1. Creating Table to put datasets on hive.

A screenshot of a cell phone

Description automatically generated

1. Created directory on hdfs :

A screen shot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

1. Internal table was created :

A screenshot of a cell phone

Description automatically generated

9)Data from external table to internal table was loaded using the following command.

A screenshot of a computer

Description automatically generated

end