# Project 3: Hadoop Application

Charlie Mei cm3947

Question 1:

I would ingest the URL DB using Sqoop because the data is largely structured and from one main data source. To interact with URL DB, I would use Apache Hive because it provides an easy to use SQL like interface for a relatively structured database.

Question 2:

I would use Apache Flume to ingest the log file because it is appropriate for the high velocity of log file data in real time.

Question 3:

The crawler returns text data which is highly unstructured. Because of this, I would use Apache Flume to ingest the crawler data since it handles unstructured data more reliably. Flume also allows for scalability, so if the number of webpages increase, Flume will be able to handle that too.

Question 4:

I would use only Apache Hive to do the MapReduce procedures in each of the jobs in steps 4 to 6. Using an interface like Hive will make the MapReduce process easier and done in the background which we don’t have to worry about. Sticking to one interface has the benefit of consistency, especially as the MapReduce process is largely similar in steps 4-6.

Hive works better in this instance because we want to conduct data analysis – developing a report that summarizes the observed relationship between sentiment and click hit ratios.

While Hive will not store the data after each step, we can write the metadata that is linked to the data. These can be stored as separate commands and then executed in a SQL-like step by step manner to get our final desired outputs. In this manner, the intermediary outputs themselves will not need to be stored, just the final outputs.