Group1 – Milestone 2

Scott Walters, Md Mahedi Hasan, Md Muhtasim Billah, Joesph Patten

Data Model: Our project relies upon a relational database management system. The data focuses on user’s reviews and comments on products so the utility of a RDMS will helpful in managing, categorizing and preparing our dataset for our co-purchasing algorithm. The data comes to us categorized, so the data is smoothly integrated into a relational database.

Data Size and Statistics: The number of tuples corresponds to the number of reviews in the larger dataset. We have over 200 million tuples (224,538,272) representing the 10 attributes of each review. See below for the table defining the tuples. The complete data is 34gb. We are concerned about the size of this data and it is recommended to use a subset of this data that is a third the size (14.3gb). This data covers 75,258,750 reviews. In addition to this data is metadata that has additional information pertaining to the items reviewed. Due to the significant time it takes to upload, parse and insert into the database we have an estimate of just as many tuples as the reviews themselves. However, this metadata is separated into three tables to compartmentalize the relation between items details, an also buy relation, and an also viewed relation.

Parsing and Database Insertion: We use a posgres database to house the data locally until we can migrate the database to the ect2 instance. The tuples are inserted into the data base via the psycopg2 library which allows us to create sql statements to insert the tuples. To obtain the tuples, we were assisted with code that loads the data from the .json.gz files to pandas dataframe objects that nicely separate each tuple’s attributes into their respective columns. From here, our program creates an insert statement string made up of this information in an additional function that is quite simple and efficient. Please see the attached py file for these processes.

Time: Because of the vast amount of data, the transfer from json file to dataframe object takes considerable amount of time. The Books.json.gz file detailing 27 million reviews has taken over two hours to transfer on Scott’s 16gb ram, 8 processor cpu. Despite this being in line to a 10 million tuple data set for a previous database class using python and posgres, we would like feedback as to the expectations on performing these tasks on the raw data. We estimate that it should take a 24 hours of operation of our code to load the full dataset into postgres.