**Bus 464 Term Project Part B**

**Kaggle Walmart**

**December 9, 2015**

Kavi Sekhon 301166470

Keith Leung 301204551

Vivian Phan 301186759

Shiqi Zhang 301138309

Change Request from Part A

From part A, we reorganized the Sale and Line\_item table to include information regarding sales total for a specific sale and item. This reorganization allows the database to update the information with fewer joins as the line\_item table includes information about specific items and the quantity sold. Please follow the directions in *Access to Additional Raw Data* sections to apply these changes before implemented Part B code.

Three Questions

Discovery:

1. Which store location generates the most sales? And which one generates the least?
   * What weather event is most common in that location?
2. Which quarter generates the least sales?
3. What types of item sell the best in winter (December - March)?

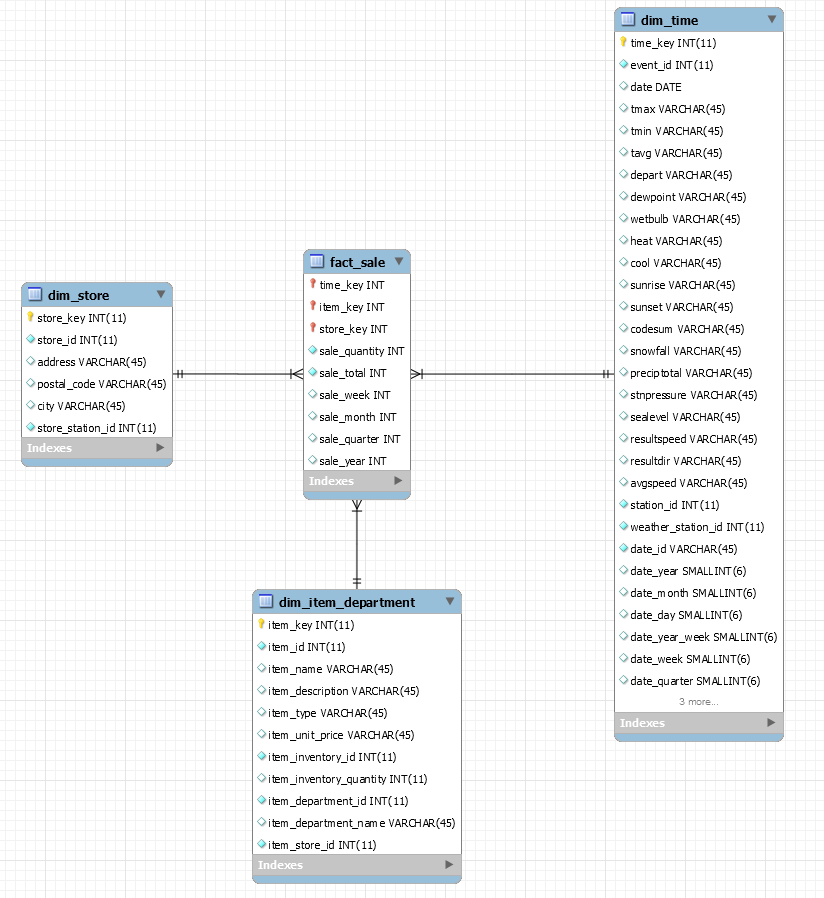
Verification:

1. More units are sold in the summer (June-August) when the sunset is later. Verify which time frame of Sunset the rate of store sales will be higher?
2. Usually when it snows heavily and rises to 2 inches on the ground, there will be traffic and driving problems. Thus, people tend to stay home and businesses are usually slower during this time. Verify that when the snowfall becomes more than 2 inches on the ground, the store’s total sales are more likely to decrease.
3. Verify that whether item types from the office and stationery sell the most during the back to school period before September (August)?

The discovery questions that a manager considers typically involve increasing sales and identifying the top performing locations. Our discovery questions determine which quarters and locations generate the most sales and indicate areas where improvement is needed. A manager also considers which item types sell best in certain periods of time, to ensure that more types of these products are available and offered at that time period.

Verification questions help managers understand and verify their assumptions. As weather was an important indicator that affects sales, we created two verification questions in order to accurately determine the positive and negative effects of weather. By verifying these two questions, a manager can anticipate and prepare for changes in demand. The last verification question identifies a specific item type. As our database divides items into departments and item types, a manager can verify and prepare for specific item types and departments to ensure adequate supply chain management.

Appendix: Star Schema



Appendix: Explanation

We have created a start schema that takes 5 tables and combines them into 3 dimensions, and then we created a *fact table* based on the dimensions of *time, store, and item\_departments.* Our time table is a combination of a *weather* table that store all the event characteristics, and *time\_dimesion* table that separates dates into multiple formats to widen the available options for queries.  This combination is necessary because our initial setup process for the sql time dimension code. After we combine that two features of the table to a single *time* table we drop boththe *time-dimension* and *weather* table, because they become redundant. In a separate occurrence, we initially call the data of store items and departments into a single table from the transaction database, since no time dimensions are present, and departments share a one to one relationship with each item.

The temp tables *tempweather, tempstation, tempstore, tempinventory, tempitem, templineitem* and *tempsale* are used as a filter to reduce the number of data running in the datawarehouse. This increases performance in each query, decreasing the runtime, to match the hardware. It is important for hardware and software to maintain performance parity, with large data sets and complex star schema.

*Note: We uploaded two files: one contains limits and one without for better performance in testing and grading with local servers and remote machines. Part B (limited).sql contains limits on the filter tables, and is best used for validating queries. If more powerful machines are present you may use the Part B (full) file that contains no limits for the strongest insights.*

Appendix: Access to Additional Raw Data

Please open up the zip file to access, that was delivered with this assignment.

1. If you need to redo the data structure from Part A, follow this guide:
   1. Go to: -> *Data* Folder -> Drag all csv. files onto the main desktop
   2. Go to: -> *Code* Folder -> Open *Part A.sql* in MySQL
2. Use the Find and Replace Function to replace the existing database name with the database you with to use
3. Change the paths of load infile statements to source from you own Desktop (where you moved all the .csv files to)
   1. Run *Part A* in MySQL
4. After *Part A* completes, run *Part B(limited)* in MySQL for validating the queries or run *Part B (full)* for retrieving all data
   1. Change the database names to your database if required