BIG DATA ANALYTICS MANAGERIAL REPORT

Team 2

15th March 2016

Following are the hypothesis on which we will be performing analysis to either prove them right or wrong.

- Elasticity of growth – Growth in employee count of one sector depends on the downfall of income generation in other sector
- Computer software sector would continue to introduce more number of patents in the future as well
- Petroleum specialty sector would introduce a lot of high paying jobs in the future
- Democratic Party will be winning upcoming election.

Data collection:

Most of the revenue data was found in the excel worksheets that we already had. The main task was to get hold of patent data. All the patent zip files, containing patent information of US, Japan and European patents were extremely large and it was not possible for us to download each one of them manually. So, we used *Hadoop to extract all the zip files and XML files*. Once the data was retrieved, we exported it in *Hive, ran queries to get hold of company name and their patents*, and stored them in CSV format. Hive was used because the extraction of XML to CSV leads to the generation of **4.7 million records.**

The presidential election information was obtained on the basis of election candidates who were contesting for a particular year, the amount of votes that they received in that election, and the state wise voting information for the candidates.

Data cleaning:

Firstly, we eliminated columns which was not required for our objective and hypothesis and then removed the columns which conveyed repeated, redundant and empty information. Most of the data present in KLD report consisted of blank values, so after prior analysis, they were cleared from the dataset. Industry code was assigned to companies which missed that information.

Data amalgamation:

As it was a huge dataset, it was not possible for us to use SQL for joining these tables. So, we took up the initiative of creating the tables in Hive, importing content via CSV and based on the composite key being “year” and “Ticker” value, we performed left outer join to a create a single dataset. The patent information and presidential election information has been stored in different datasets at this point of time, which will later be coagulated with the consolidated dataset.

Statistical analysis:

The statistical analysis on the consolidated sheet is performed considering a lot of different factors, using R. The summary command in R does not provide the required information in the right way. So, we have prepared different consolidated sheets, where Mean, Median and SD is found for Employees, net income and revenue, on the basis of company, specialty, year and industry. This way, we get the segregated information in the right way, providing us the scope to represent the data in multiple scenarios.

Data Visualization:

The entire visualization was done in Tableau. One of the representations shows the annual revenue generated by each sector over the years, which also gives the information about the topmost company in that sector and the state which consists of most of these companies. Another representation has been formed to find the sector which dominates the entire industry. A visual graph has also been created to depict the financial hit taken by downfall in income over the years. One of the graphs shows representation of mean revenue, income and employees, based on specialty.