**UTILIZING BIGDATA METHODOLOGIES AND DATA ENGINEERING TO OBTAIN AN ANALYSIS AND INSIGHT FOR BUSINESS ACUMEN**

A Summer Internship Report submitted in partial fulfilment of the requirements for the degree of

Master of Business Administration

**By**

**EMANUEL PETER THOMAS**

**REGISTER NUMBER**

**1828610**

**Under the Guidance of**

**Prof. REENA RAJ**

****

**Institute of Management**

**CHRIST (Deemed to be University), Bangalore**

**JUNE 2019**

**SOLVING BUSINESS PROBLEMS USING BIGDATA METHODOLOGIES AND DATA ENGINEERING TO OBTAIN AN ANALYSIS AND INSIGHT FOR BUSINESS ACUMEN**

A Summer Internship Report submitted in partial fulfilment of the requirements for the degree of

Master of Business Administration

**By**

**EMANUEL PETER THOMAS**

**REGISTER NUMBER**

**1828610**

**Under the Guidance of**

**Prof. REENA RAJ**

****

**Institute of Management**

**CHRIST (Deemed to be University), Bangalore**

**JUNE 2019**

**Declaration**

I hereby declare that the Summer Internship report entitled “UTLIZING BIG DATA METHODOLOGIES AND DATA ENGINEERING TO OBTAIN AN ANALYSIS AND INSIGHT FOR BUSINESS ACCUMEN AT MUSIGMA ”has been undertaken by me for the award of Master of Business Administration. I have completed this study under the guidance of Prof. Reena Raj .

I also declare that this Summer Internship Project report has not been submitted for the award of any Degree, Diploma, Associate ship, Fellowship or any other title, in Christ (Deemed to be University) or in any other university.

Place: Bengaluru \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: JUNE 2019 Emanuel peter Thomas

1828610

****

**Certificate**

This is to certify that the Summer Internship report submitted by Emanuel peter Thomas on the title “ UTLIZING BIG DATA METHODOLOGIES AND DATA ENGINEERING TO OBTAIN AN ANALYSIS AND INSIGHT FOR BUSINESS ACCUMEN AT MUSIGMA” is a record of Summer Internship work done by him during the academic year 2019-20 under my guidance and supervision in partial fulfilment of Master of Business Administration.

Place: Bengaluru \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: JUNE 2019 Prof. Reena Raj

Professor

Institute of Management

Christ (Deemed to be University)

Bengaluru

**Acknowledgement**

I am indebted to many people who helped me accomplish this Summer Internship programme successfully.

First, I thank the Vice Chancellor Dr Fr. Abraham V M, Christ (Deemed to be University) for giving me the opportunity to do my project.

I thank Father Abhraham, Vice Chancellor ,Dr. Jain Mathew, Dean, Prof. Georgy Kurien, Associate Dean, Prof. Lakshmi Iyer Head of the Department, Institute of Management, Christ (Deemed to be University) for their kind support.

I thank Prof. Reena Raj for her support and guidance during the course of my internship. I remember her with much gratitude for her patience and motivation, but for which I could not have submitted this work.

I wish to express my sincere thanks to my corporate mentor Mr. Abhinao , Miss Akhila, Mr. Ashish Sawant and Mr. Shanmugathan for giving me an opportunity to work under their guidance and successfully complete my internship.

I also thank Mr. Dhiraj C Rajaram the founder, chairman and CEO of Mu Sigma Inc, to whom I am extremely obliged and indebted for his expertise, sincere and valuable guidance and encouragement extended to me.

I thank my parents for their blessings and constant support, without which this internship project would not have seen the light of day.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Emanuel Peter Thomas

1828610

**EXECUTIVE SUMMARY**

Data has grown exponentially with time .At a certain juncture in time it was used to keep track of events and proceeding ,however now in an era where data is the new oil it gives insights into revelations that would have not been noted in hindsight by regular operation .Handling such data has become an ominous task as the 4Vs Volume ,Variety ,veracity and Velocity of data has become important as well .Technology has however aided to this progression with technologies aiding parallel distributed processing ,fault tolerance ,retrieval speed time and other benefits which makes life easier for the analysts .Statistical analysis packages have been incorporated across multifarious tools .Hypothesis testing would be disregarded as data manipulation would not be possible without effective ETL (Extract Transform Load) and transformation. Getting the data and structuring it and tailoring it for perfect and accurate analysis such that all the assumptions are met while performing the analysis is crucial.

Hardware capacity has also grown as now companies are making a trade off with cost to storing data on RAM to obtain faster processing of data .Hypothesis testing requires diligent and strict assumptions .Applying these perquisites for better exploratory data analysis is necessary , but applying these requisites across data bound by the constraints of the 4Vs is a challenge

Statistical techniques which were founded in the 1970s have been rejuvenated with the advent of big data

**CERTIFICATE**



**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Chapter Number | TITLE | Page No. |
| 1 | INTRODUCTION | 1 |
| 2 | INDUSTRY PROFILE | 4 |
| 3 | COMPANY PROFILE | 10 |
| 4 | OBJECTIVES OF STUDY | 18 |
| 5 | DATA ANALYSIS / PROCESS STUDY | 20 |
| 6 | FINDINGS & RECOMMENDATIONS | 57 |
| 7 | LEARNING | 60 |
| 8 | REFERENCES | 64 |
| 9 | APPENDIX |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Serial Number | Title | Page No. |
| 1 | Organizational Structure | 2 |
| 2 | Business Intelligence in analytics | 6 |
| 3 | Problem solving flow technology | 13 |
| 4 | OBI model | 14 |
| 5 | OBI | 16 |
| 6 | AOPS(Art Of Problem Solving ) | 17 |
| 7 | Data Dictionary | 19 |
| 9 | Mu-OBI | 21 |
| 10 | Types of retail outlets | 22 |
| 11 | Pivotal moments of retail | 23 |
| 12 | Revenue of companies in retail | 24 |
| 13 | Revenue of companies in retail | 24 |
| 14 | MECE tree node-1 | 26 |
| 15 | MECE tree node -2 | 22 |
| 16 | MECE tree node -3 | 27 |
| 17 | MECE tree node-4 | 28 |
| 18 | MECE tree node-5 | 28 |
| 19 | HDFS Architecture | 34 |
| 20 | HDFS Environment | 35 |
| 21 | Schema | 40 |
| 22 | Locations arranged in descending order | 43 |
| 23 | Locations arranged in increasing order | 44 |
| 24 | Most frequent pickup locations | 45 |
| 25 | Least frequent pickup locations | 46 |
| 26 | Sentimental analysis | 47 |
| 27 | Analytical Value Chain | 61 |
| 28 | HDFS architecture | 63 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| Table Number | TABLES | Page No. |
| 1 | Challenges and opportunities | 25 |
| 2 | Data Description | 38 |
| 3 | Customer Feedback | 53 |
| 4 | Customer Feedback polarity | 55 |
| 5 | 3.5 Normalized table | 57 |

**LIST OF CHARTS**

|  |  |  |
| --- | --- | --- |
| Table No. | CHARTS | Page No. |
| 1 | Flow if data from HDFS to SPARK | 37 |
| 2 | Flow of data from HDFS to HIVE | 37 |
| 3 | Data Architecture | 38 |
| 4 | Consumer feedback for sentimental analysis | 51 |

**CHAPTER 1**

**INTRODUCTION**

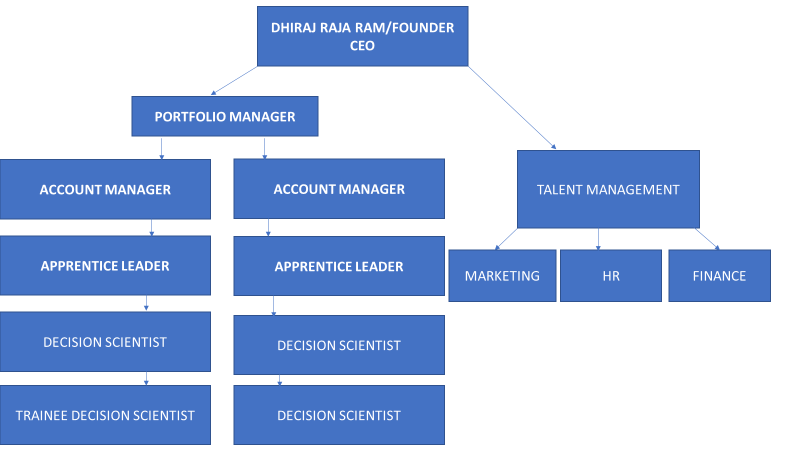
“I suspect that over the next couple of years, it will become a lot easier to hire data scientists. A wide variety of universities-including Harvard, Berkeley, Stanford, Columbia, NC State, and others-are offering data sciences courses and starting degree programs.(Berkeley is reputed to be starting three different programs, which I’m not sure represents progress.)In the meantime, there are a number of consulting organizations that can offer data science services. The big analytics providers, including Accenture , Deloitte , and IBM, are all gearing up on the topic, and there are a variety of boutique firms that have formed. The large offshore providers, of which Mu Sigma is the largest , are also getting in on the big data game.” (**Thomas H.Davenport, 2012)**

**Mu Sigma** is an Indian [management consulting](https://en.wikipedia.org/wiki/Management_consulting) firm that offers a plenty of [data analytics](https://en.wikipedia.org/wiki/Data_analytics) services pertaining to multifarious domains. The firm's name is derived from the statistical terms the "Mu (μ)" and "Sigma (σ)" which symbolize the [mean](https://en.wikipedia.org/wiki/Mean) and the [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) respectively, truly befitting of a billion dollar business analytics giant to be named in such a manner .It holds the unique distinction of securing the largest funding round by a business analytics company. The rewards that Mu Sigma has accumulated over the years is a rapidly growing number. Synonymous to all great companies, the core of the company is built around the beliefs and ideas of a very inquisitive individual. Founder, Dhiraj C Rajaram, had a brilliant idea which bore proclivity towards entrepreneurship.

Mu-sigma follows a flat hierarchy , where freshers are employed and given a training in statistical methodologies and technology which come in unison to give a better understanding of problem solving. Once a trainee decision scientist received significant training and has proved him or herself over multiple project , he/she has to clear certain examinations where he/she is promoted to a Decision Scientist . A decision scientist allocates the coding to be done by the trainee decision scientists while the statistical inferences pertaining to the business decisions are made by the decision scientist. It is the job of a decision scientist to guide the trainee decision scientists in effective coding and getting proper statistical inferences .Once a decision scientist has proved his mettle and once the apprentice leader feels he is capable of managing people in an effective manner , he is promoted to a apprentice leader. The responsibilities of an apprentice leader include managing the mu-OBI, mu-Universe ,gauging the hypothesis set up the decision scientist and dictating which tools and technologies are to be used as per client preference . An apprentice leader can approach an account manager and intimate him/her of how many decision scientists and trainee decision scientists would be required for a particular project .

The account manager manages a particular account .A single individual can be an account manager for multiple accounts .The account manager reports to a portfolio manager . The portfolio manager is selected based on the vertical .For pharmaceutical clients, all the pharmaceutical clients (account managers) have a single portfolio manager . For the automobile industry , all the account managers have a single portfolio manager .Mu-Sigma have thus improvised on their organizational structure as per the whims , fancies and benefits of the clientele.

Fig .1.1: Mu-Sigma Organizational Structure



Source: Compiled by the author

**CHAPTER 2**

**INDUSTRY PROFILE**

**2. CONSULTING INDUSTRY:**

Management consulting is the Golden Chip that help organizations to improve their performance, operating primarily through the analysis of existing organizational problems and the development of plans for improvement. Many Indian and International Organizations Make use of efficient Consulting Services to to Boom Up their Business, The era of Consulting is Knew to India and seems to be growing fast over the last decade. There are numerous types of Management Consulting Services which includes Operations Advisor, Strategy Advisory, Data Analytic Consulting and many more.

The significance of management consulting in India can’t be undervalued. In India Management consulting is estimated to be a Rs. 30,000 crore industry, which is growing at a CAGR of 30 percent.

**2.1 DATA ANALYTICS CONSULTING SERVICES:**

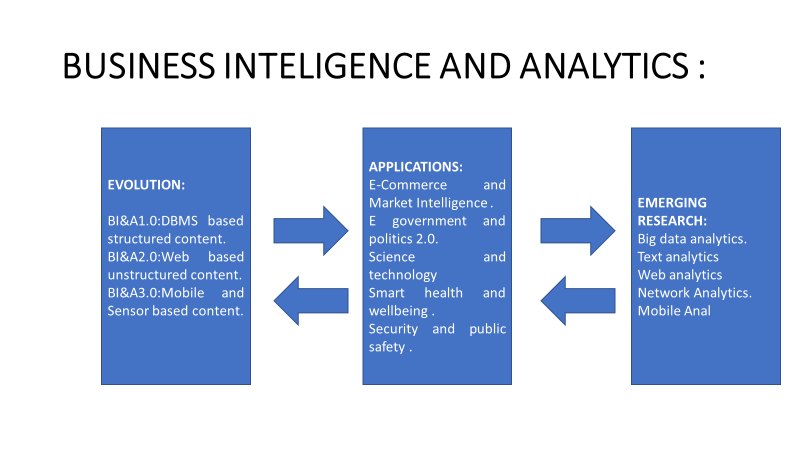
Data Analytics Consulting services Known as Analytics Consulting in Short is Important Role that combines and then balances Analytics and Hard Core business with the aim to prove best of Analytical Solutions to many of the Companies Problems.

Initially the consulting industry was more focused only on physical Operations and processes of the company but now with the digital revolution in the consulting industry, many tremendous opportunities to reinforce the Clients Experience ids Opening up..  
  
Data analytics is a quickly evolving technology that harnesses statistics, AI and advanced market insights to spot meaningful patterns in massive data-sets. Smart and Efficient Data Analysts  give premium insights into an organization’s performance metrics, and also Impart Knowledge on the companies past performance and thus Suggest Areas of Improvements.  
  
Globally, the Annual company spending on analytics consulting soared to $43 billion in 2017.Experts predict that In the near future data Analytics is going to be the only Most Efficient Problem Identifier and Solver for Companies across industries.

Inclusive growth has led to the burgeoning of the industry and several other industries as well. Industry 4.0 has led to improved control-centric optimization and intelligence. Greater intelligence can be achieved by interacting with different surrounding systems that have a direct impact to machine performance. Business intelligence has grown due to the improved methods of data collection, extraction and analysis. Systems are using techniques that have evolved from the main statistical methods developed in the 1970s and data mining techniques developed in the 1980s.Data warehousing and Extract Transform and Load functions are essential for converting and integrating enterprise specific data. Combining these processes with the techniques for statistical analysis and techniques incorporating mining of data are adopted for association analysis, classification and regression analysis, anomaly detection, predictive modeling and clustering in a plethora of applications pertaining to business.

**2.2 BUSINESS INTELLIGENCE IN ANALYTICS**

Fig 2.1: Business Intelligence and it’s applications



Source: Compiled by the author

Organizations can present their businesses online and present and interact with their customers directly, such phenomena are characteristic of the HTTP WEB 1.0. BI&A 2.0 have led to the development of web analytics, web intelligence, understanding patterns of users from the user generated content. By analyzing customer orders and transactions, google and amazon analytics can evaluate the customer’s buying patterns and behavior. Social media analytics has led to a growth of tremendous business opportunities. Several new data maiming techniques are being employed in this domain. The continuous development has led to the advent of BI&A3.0 and this has been complimented by the burgeoning number of cellphones, tablets, smaller CPUs etc. Integration of machine learning with the internet of things i.e. evaluating and noticing patterns from RFIDs, radio tags and barcodes. Integration of various technologies with the analytics domain is on the rise, as user patterns can be evaluated, and better predictions and progress can be made. Pivotal steps have been taken by the business community towards embracing these new technologies and are taking elaborate steps towards implementation of these technologies in the industry. With the increase in the volume, variety, velocity and veracity of data and the big data distributed processing systems that have evolved to extract and herness the maximum potential of petabytes and terabytes of data, several other industries have benefitted because of this i.e. e-commerce and market intelligence-government policies, science and technology, smart health and security.

**2.2.1 E-Commerce:**

A tremendous market transformation has been accomplished by e-commerce vendors such as amazon, Flipkart and eBay. Social media analytics, market basket analysis and web analytics have become domains for major development by companies such as amazon, google and Facebook. Traditional transactions capture no data whatsoever, however, online transaction can effectively predict the entire behaviors and demeanor of the users. Social media analytics has led to embracing text analytics and sentiment analysis. Netflix have developed a program in python that carefully predict the shows and movies the user is interested in. Having once being rejected a deal with blockbuster, where in blockbuster would promote Netflix and Netflix would promote blockbuster’s campaign online. Blockbuster went bankrupt in 2018 while Netflix is a 28-billion-dollar company.

**2.2.2 E-governance:**

Successful policy discussions, campaign advertising, voter mobilization, event announcements, online donations have been reached by the exploiting the multimedia platforms. Opinion mining, social network analysis and other techniques are used to support online political participation.

**2.2.3 Science and Technology**

Sensors are being used for increased automation. National Science Foundation has made it compulsory for every project to have a data plan. An advancement in data processing, gathering past insights and making visualizations and graphs has made it possible to enhance steps needed towards scientific progress. Iplant initiative was funded by NSF to facilitate the research by users towards plant biology. Biologist are using datasets to evaluate and understand the factors the best affect plant biology. The Sloan Digital Sky Survey (SDSS) has made it possible to make discoveries and with greater precision towards the field of exploration.

**2.2.4 Pharmaceutical and health:**

The sources of information are of two types being genomic driven data (study of genes and cells ,tissues and neurons ) and payer driven data( which encapsulates patient data, health records of individuals and families , smart watches such as Fitbit that provide users with a dashboard of his/her activities and movements) The data that can be gathered from each individual is supposed to be in excess of 4 terabytes . Sensor technology, networking machine learning technology, modeling cognitive processes, system and process modeling, are finding huge application in the healthcare industry. Prediction of tumors and whether it could be malignant and benign are made possible by analyzing the different parameters of the tumor. In addition to analysis of Emergency Health Records, unique research opportunities are being made for support in certain diseases such Parkinson’s, cancer and Alzheimer’s.

**2.2.5 Security:**

With an increase in the volume, velocity, variety and veracity of data, handling and understanding the different data formats is crucial to prevent online terrorism and enhance cyber security. Gathering data in in multilingual formats and then examining the data closely is necessary. Protection of intellectual assets and infrastructure is pivotal to organizations as they attempt to strengthen their cyber security domain. A consistent framework to prevent against cyber-attacks is possible only by collaboration with the data analytics domain. Selected BI&A technologies such as criminal association rule mining and clustering, criminal network analysis, spatial-temporal analysis and visualization, multilingual text analytics, sentiment and affect analysis, and cyber-attacks analysis and attribution should be considered for security informatics research.

**2.2.6 Banking Financial Services and insurance:**

Used to segment, target, attract and retain customers more efficiently and this would not be possible without the intervention of analytics. North America commands the largest market share in the analytics domain due to an ever-increasing application in the banking and financial services domain. Analyzing fraudulent transaction and then checking if the loan seekers are viable and measuring the consistency and reliability of the customers is one amongst the basic applications. Regulators are pushing the banks to have a better understanding of the data they possess. Understanding the data leads to insights in risk evaluation and getting accurate estimates of beta (company relativity to change in market) and Jensen’s alpha. Customer profitability can be enhanced by offering personalized loans. Customer analytics and risk analytics is facing an ever-increasing demand.

The analytics industry in India is valued to generate an annual revenue of 3bn$ as of 2019 and a CAGR of 34%.11% of the analytical industry is a proclivity towards predictive modelling and data science while a sizeable 22% can be attributed to big data.

**CHAPTER 3**

**COMPANY PROFILE**

**3. MU SIGMA**

With these intentions Dhiraj started Mu-Sigma in 2004. The beginning of the company was a struggle as is most with startups, however today it is a burgeoning million-dollar company. Aware of the competition from companies such as Accenture and TCS did not deter Dhiraj from entering this domain. Dhiraj initially invested 80 % of his money into the business. After surviving four years on his own, Mu Sigma raised its first round of investment worth $30 million from FT Ventures (now FTV Capital) in 2008. Subsequently, in April 2011, the company raised an additional $25 million from Sequoia Capital, followed by**the third round of $108 million again from Sequoia and a private equity investor General Atlantic, which is the highest investment ever made in an analytics company.**

Mu Sigma is headquartered in [Chicago](https://en.wikipedia.org/wiki/Chicago) and has a global delivery center in [Bangalore](https://en.wikipedia.org/wiki/Bangalore). Mu-Sigma is unique in the sense of solving problems. It is the largest Indian company that deals with analytics in the purest sense. It follows a flat hierarchy and Mu-Sigma is in the stage of releasing an IPO in the next 5 years. Being an analytics company, it helps its clients institutionalize data driven decision making, harnessing big data analytics. Having worked with a plethora of market leading organizations across multiple industry verticals, it is using the experience for solving high impact business problems in key horizontals such as marketing, risk and supply chain. It provides its clients with a holistic ecosystem of proprietary technology platforms, processes and people.

Being a major player in the pure analytics domain in the industry, Mu-Sigma specializes in data analytics and machine learning. One area of specialization which mu-sigma is keen to build-up and bring to the top position of providing services in the data analysis industry is in the data engineering field. The company has made it clear to improve and provide clients with a portfolio of services pertaining to the D3 stack i.e. Data engineering, Decision Science and Data scientists. The company is keen to fill this void by investing significantly in data engineering and training new joiners since 2018.

The problems in the context of business shortcomings i.e. the need of the clients is of 4 categories: -

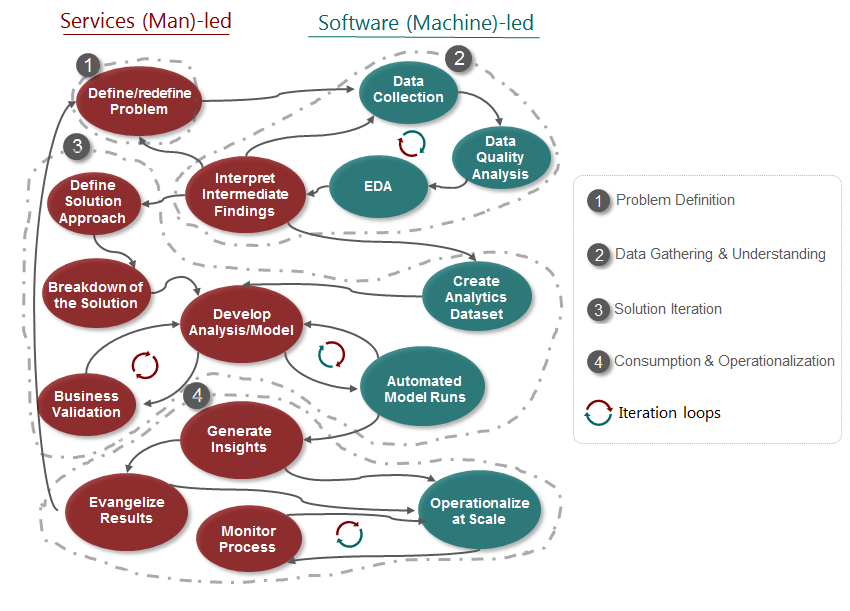
1. Demand Forecasting.
2. Churn. (how soon the customer base is prone to change)
3. Digital Marketing which is evaluated from clicks per ad, sales etc.
4. Promotional Campaigns-

80% of the problems that mu-sigma receive from their clients are of the above-mentioned nature.

Mu-sigma believes in building their employees from 2 major prospects i.e. Technosphere having a good understanding of the necessary tools and software and problem space understanding i.e. application of the tools and software necessary to solve the business problems from a purely statistical sense.

**3.1 MU SIGMA’S UNIQUE OPERATING MODEL:**

As business complexity increases, “Software” and “Software as a Service (SaaS)” will need to push to new limits to a new example of man and machine working in collaboration. Mu-Sigma calls this “Service as a Software”. Analytical decision making is achieved through data science consulting and solutions.

Fig .3.1: Problem solving flow methodology

Source: Mu-Sigma Homepage

Problems are tackled in 3 stages: -

1. Understanding the needs of the client
2. Identify the problem and gather insights.
3. Build a narrative.

Apart from the passion exuberated from the employees devoted towards solving problems, employees of Mu-sigma also build models pertaining to several problems which include designing dashboards (interactive visual storytelling of data). Dashboards in this day is of great value. Pictorial representation of data in the right sense is that it grabs the observers cognitive attention. For example, Fitbit the health app leads in the market and this is what distinguishes a brand like MI from Fitbit.

In solving problems, Mu-Sigma uses and follows four types of analytical methods and they are namely:

1. Descriptive
2. Inferential
3. Predictive, and
4. Prescriptive.

Mu-Sigma has certain artifacts that provides their employees a greater sense of understanding of the problem.

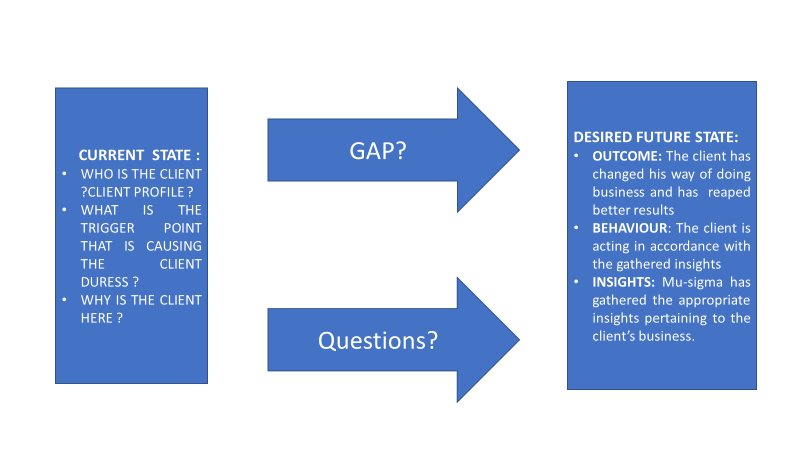
**ARTIFACT NUMBER 1**

**Mu-PDNA (Problem DNA)**

Mu-PDNA analyzes the problem from different scenarios:

1. **Design:**

Fig .3.2: Mu-Problem Definition



Source: Compiled by author

This consists of the current state of the problem which involves finding out: -

* Who is client?
* Trigger: What is upsetting the client or what are goals of the client?
* What is the client doing right now?

The future state which generates

* Insights
* Behavior to be followed by the client.
* Outcome of the client.

Mu-Problem Definition lies at the very core of problem solving .Ideally calculating the gap between the existing present state and the final future state is of paramount importance. This gives insights as to what to must be accomplished to help the client meet his/her desired set of objectives.

1. **Mu-Search:**

This is used to give a detailed report involving investigation of the problem by breaking it down into smaller segments. Sources and tags relating to the study of the problem are posted within this section.

1. **Representation:**

The representation of the problem is done using a MECE (Mutually Exclusive Collectively Exhaustive) tree. This involves framing the hypothesis at each level of the tree. This methodology is important as it teaches the employee to ask the right questions. Mu-Sigma considers problem definition as a very important part of problem solving. While the events recorded pertaining to the problem definition are mutually exclusive, they should contain all the problems within the sample space. The gap is referred to as Sequential generated gap.

**ARTIFACT NUMBER 2:**

**Mu-Decision Science** Chain keeps a track of how much of each task pertaining to the problem statement is solved. Problems are broken down into smaller chunks and red indicates a task that is incomplete, and green indicates the task at hand that has been accomplished. It maps the problem based on its complexity to an individual of relevant expertise.

It builds a communication within the team keeping them focused on the problem at hand while simultaneously identifying potential bottlenecks and simplifying the problem allowing the employees to devise methods around the potential bottlenecks without wasting any time .

**ARTIFACT NUMBER 3:**

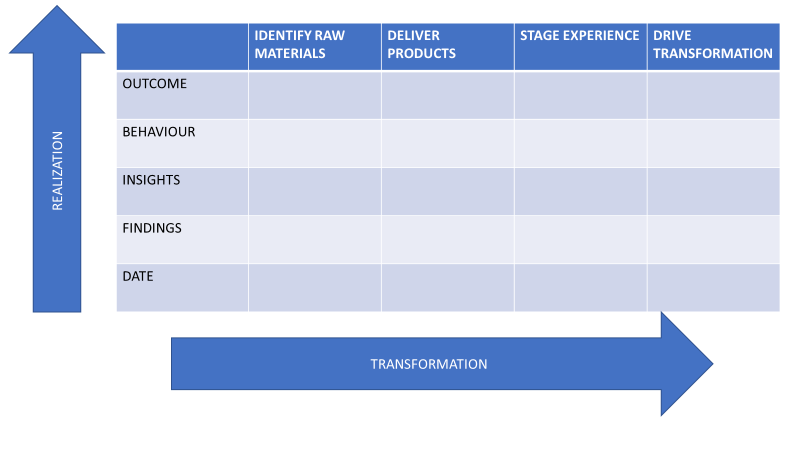
**Mu-Universe** was built on the ideology that many problems are interconnected. As the problems are broken down into little chunks, it maps the problems to similar problems based on the complexity function. Algorithms used are text matching and string-matching algorithms that are used to match the strings together. Large Scale integrations of problem and the algorithms used to tackle the problems are described.

Clustering and classification are done through the fore mentioned fuzzy mentioned algorithms. A plethora of algorithms generated and solved in R and python are available in the knowledge repository that allow the users to use relevant hardcoded algorithms to generate insights swiftly .Dhiraj Rajaram’s vision is to fill the grey matter that is left uncovered by google , and his ideologies have transgressed into the company as a pivotal backbone of problem solving . A single problem could be solved by solving other miniscule and minor problem . Moreover, more data leads to better insight that can be deciphered by the employee.

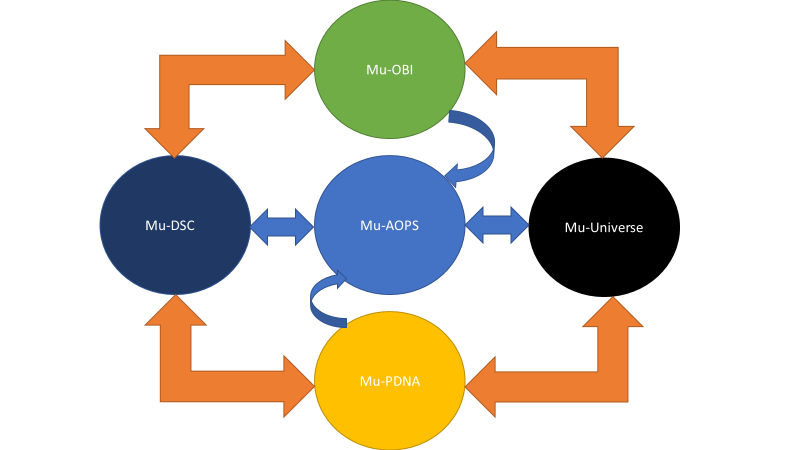
**ARTIFACT NUMBER 4:**

Mu-OBI (Outcome Behavior and Insights) is used to classify data into a tabular structure and realize what is expected of the client. Although sharing a degree of similarity with artifact number 1 , it analyzes the OBI across several horizontals of the business. This helps in analyzing the interrelation of problems across the horizontal that could help in solving the problem of the company and across similar verticals .

Fig .3.3: Mu-Outcome Behaviour Interface



Source: Compiled by author

Together these artifacts come together to form a full circle that define the art of problem solving. Fig .3.4: Mu-AOPS ecosystem.

Source: Compiled by author

This forms the backbone of a technological knowledge repository of concepts and past case studies that can used as a guidance for future fulfillment of projects in the most efficient manner.

**CHAPTER 4**

**OBJECTIVES**

**THE OBJECTIVES OF THE STUDY ARE:**

1. **TO PERFORM DATA PROCESSING BY EXTRACTING AND LOADING DATA THROUGH SPARK .**
2. **TO ANALYZE AND INTERPRET CUSTOMER TRANSACTION DATA OF A CAB AGGREGATOR TO IDENTIFY THE LOCATION GENERATING MAXIMUM REVENUE WITH HIGHEST FREQUENCY**.
3. **TO IDENTIFY AND ASSIGN POLARITY TO DATA BASED ON CUSTOMER REVIEWS ACROSS A E-RETAIL WEBSITE.**

**CHAPTER 5**

**METHODOLOGY**

* 1. **. THE HADOOP DISTRIBUTED FILE ECOSYSTEM**

Data engineering as a subject was created due to lackluster database maintenance. Real world decision making comprises a combination of knowledge, information and wisdom. Feedback comprises knowledge, information and data. Three categories of databases exist namely unstructured, semi-structured and structured databases. Unstructured databases comprise of PDF, JPEG, MP3 and movies, while semi structured databases consist of CSV files, mongo DB etc. Structured database consists strictly of two-dimensional data. Semi structured data requires the attributes be mentioned again and again. Based on problem complexity we can opt for structured or semi-structured databases. Structured data can be converted to unstructured data, however the vice versa cannot be done.

Trade-offs exist even in this domain. If we want a database that gives faster retrieval, then we opt for an unstructured database. If we want a database that is meant for analytics, then we use the relational model. Structured data sets are analytical data sets.

**5.1.1 CRISP model:**

The data understanding is dependent on the business understanding. The steps following data understanding are: -

1. Data understanding
2. Data preparation.
3. Modelling
4. Evaluation
5. Deployment.

**Data ingestion** is getting data, transformation and loading. **Data lake** is a dump of data which could be structured and unstructured. **Data catalog** is useful for determining the source and it gives an overview of basic information of the database.

* + - 1. **DATA MODELLING CONCEPTS:**

1. Entity and relationships
2. Degrees of relationship
3. Normalization
4. Dimension and fact tables: The is done to achieve 3.5 degree of normalization
5. Schema types: Snowflake and star
6. OLAP AND OLTP
7. Slowly changing dimensions which involve merge, update.

Big data involves data that is costly and extremely important to an organization. The limit is decided by the technology. If python can't handle it, then some other technology is used. Majority of the data comes in the form of photos.

The 4V’s of data comprise of **Volume, Variety, Velocity and Veracity**. The train explains that Veracity is the most critical as unauthentic information is available across the internet. Concepts of Hadoop includes effective usage of unused space and distributed processing. Fault tolerance is improved due to replication of data.

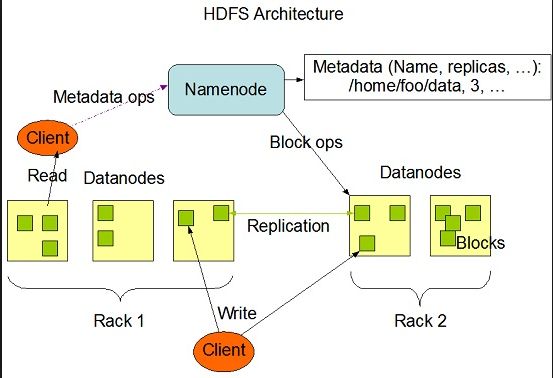
There exist **3** components of distributed processing:

1. HDFS (Hadoop Distributed File System) used for storage: - placed in a master node i.e. the main computer.
2. MapReduce which defines the logic of data processing. It is used for mapping the nodes across which the data would be stored and for returning the values to the user, it zeroes down to it.
3. YARN: Framework to run the processing task (resource allocator)

Hadoop ecosystem comprises of HIVE (SQL like query interface), HBASE (a different kind of a database), PIG (Unstructured to structured database). It uses the concept of distributed processing where data is stored in parallel across several nodes. Parallelism is promoted as it improves efficiency and faster processing.

At Mu-Sigma, each node allows a storage of 128 MB, the data is split into chunks and stored across the nodes. Mu-sigma believes that the MapReduce slows down the speed of retrieval and hence this induction of apache spark which operates on python. Essentially apache SPARK works in collaboration with HADOOP. HADOOP runs on commodity hardware, which involves the slave nodes.

Fig .5.1: HDFS architecture.



Source: <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>

* + 1. **IMPORTING DATA INTO HDFS: -CODE FOR SQOOPING THE DATA”**

---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

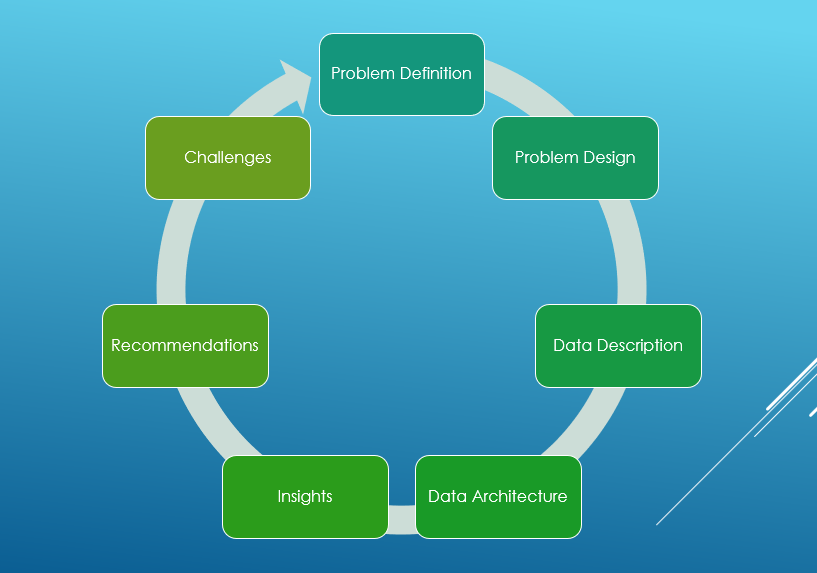
Sqoop import --connect jdbc:postgresql://10.1.2.60/postgres --username ###### --password ###### --table train\_geo\_dist --target-dir /user/Emanuel.thomas/geodisntc -m 1 -- --schema data\_engineering

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

When the data is viewed using: hdfs dfs -tail /user/Emanuel.Thomas/geodisntc/part-m-00000, following is the result:

* + 1. **UNDERSTANDING THE HDFS ENVIRONMENT:**

Fig .5.2: HDFS ecosystem



Source: <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>

* + - 1. **PROBLEM DEFINITION:**
* **Context:**

To streamline the performance of cab aggregator, partnering with major vendors by building a data pipeline and presenting clear-cut representation of road blocks with respect to Geography.

* **Constraints:**

The constraints while solving the problem are as follows:

* Huge amount of data hence tools and applications are required to handle such huge amount of data.
* Location not specified as the geographic points were explained with latitudes and longitudes.
* **Trigger:**

To build a data warehouse on top of HDFS which will ingest data from sources and store it in HDFS. Also we need to find the insights for the client business so that we can take the correct decision which will generate profit and help in growth of the business.

* + - 1. **PROBLEM DESIGN**

The steps for solving the problem and get the desired result.

1. The problem has multiple steps which are required for finding the end result:
2. The data was extracted from source (PostGreSQL) and loaded in HDFS (Hadoop Distributed File System) in the Hadoop Ecosystem.
3. The data was then extracted from HDFS and loaded in Hive in an External Table and then we run SQL queries to get the desired output.
4. The other method to find the desired output is written below:
5. The data was also loaded in Spark both as a Data frame and a RDD (Resilient Distributed Datasets) by extracting it from HDFS.
6. Perform transformations on Data frames and RDD to get the desired output.

Frequency will be calculated  by: passengers count as per pickup and drop-off latitude and longitude .

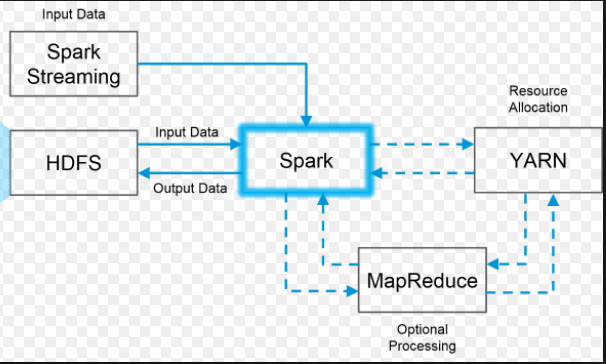
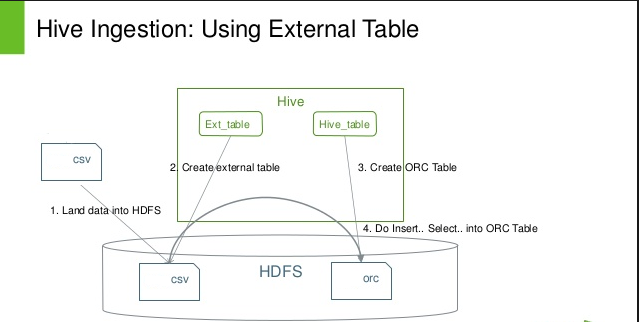


Chart 1:

🡨Flow of data from HDFS to Spark and the overall process flow.



Flow of data from HDFS to Hive and the overall process flow. 🡪

Chart 2:

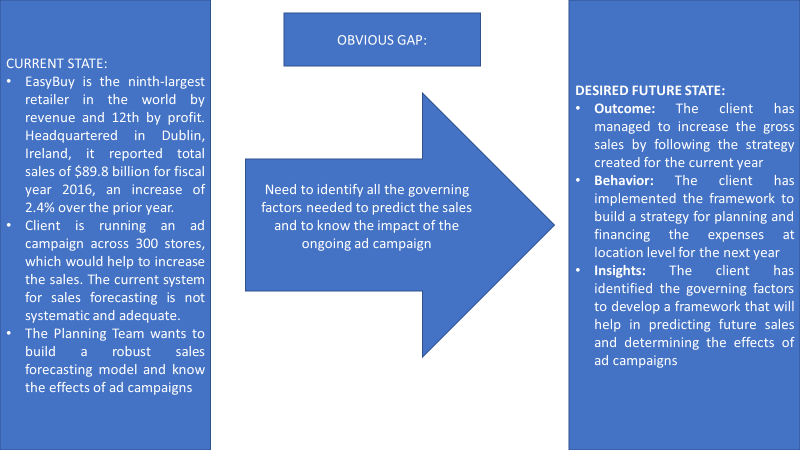
Source: <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>

**CHAPTER 6**

**ANALYSIS**

**6.1 TO PERFORM DATA PROCESSING BY EXTRACTING AND LOADING DATA THROUGH SPARK .**

This Analysis involves evaluating past sales, corresponding promotions offered, stores and the range of products being offered. Data extraction and loading is done using the Spark plug in within the HDFS ecosystem and further used to Analyze the data.

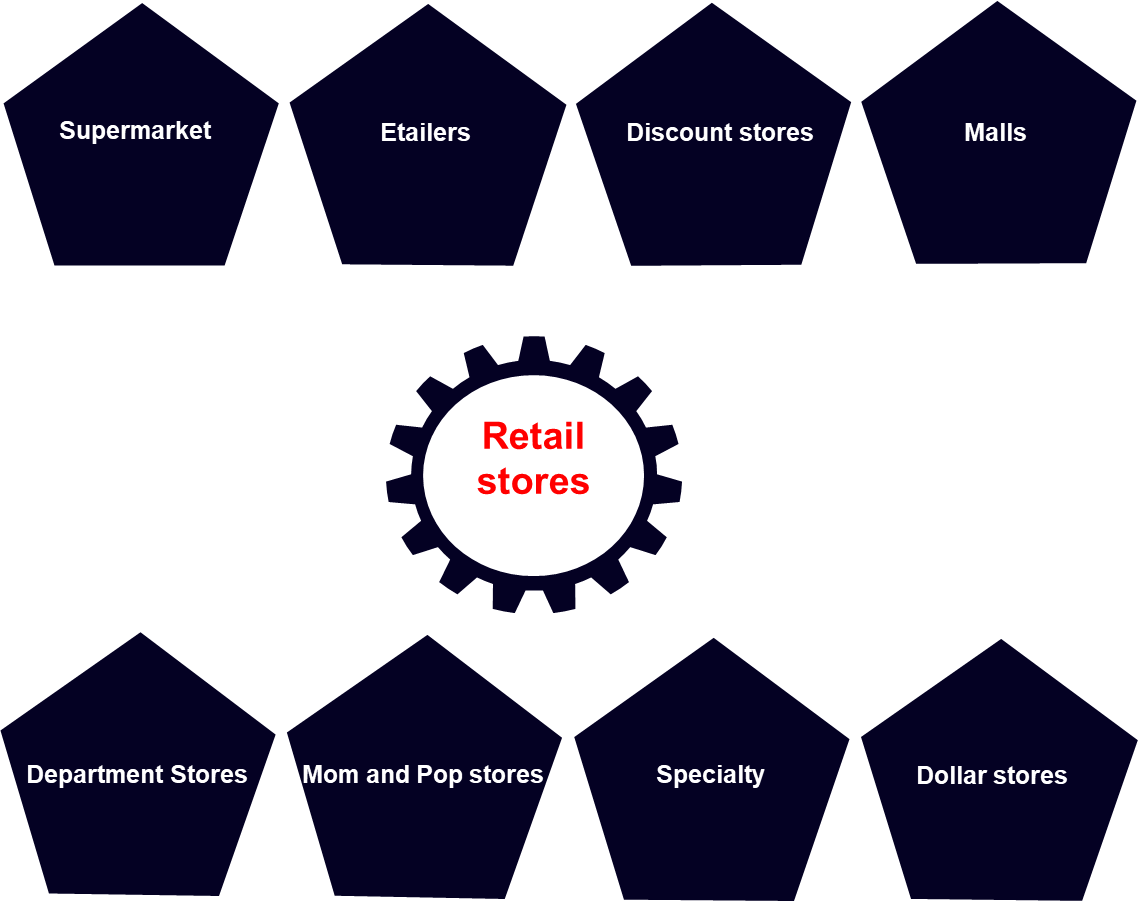
** MU-OBI** Fig .6.1: Mu-PDNA model

Source: Compiled by author

Retail involves the sale of goods from a single location (malls, markets, department stores etc.) without any middleman into the hands of the customer. Retailing is nothing but transaction of goods between the seller and the end user as a single unit (piece) or in small quantities to satisfy the needs of the individual and for his direct consumption. The manufactures and wholesales play a significant role. The retailers purchase goods in huge numbers (huge numbers) which are to be sold to the customers with the aid of the manufacturers or the wholesalers.

**TYPES OF RETAIL OUTLETS:**

Fig .6.2:Types of Retail stores

****

Source: <https://www.marketing91.com/10-types-of-retail-stores/>

**POPULAR BUSINESS MODELS:**

The most ubiquitous and most profitable business model is that of the traditional retailer. The standardized retailer profits obtained by selling products and services directly to buyers at a mark-up from the actual cost. Almost all bricks-and-mortar retailers employ this complementary business model online, and there are a huge number of online-only retailers that employ this business model. The retail strategies adopted by digital retailers and physical retailers are very heterogenous in nature .

**VARIATIONS ON TRADITIONAL BUSINESS MODELS:**

Not all traditional retailers operate in the same manner. There are three categories: low cost, cost plus, and premium retailers.

* **Low Cost Retailer**

Low cost retailers inadvertently sell to the mass market with a strong emphasis on price having a trade-off over quality or other premium product/service attributes. Low cost retailers typically offer a high number of Stock Keeping Units at the best possible price. By emphasizing on the price, the low-cost retailer operates on extremely low margins and must have the power to negate offers made by suppliers and wholesalers cheap rates. [Amazon](http://www.digitalbusinessmodelguru.com/2013/07/analysis-of-amazon-business-model.html) and Wal-Mart are prime paradigms of low cost digital retailers.

* **Cost Plus Retailer**

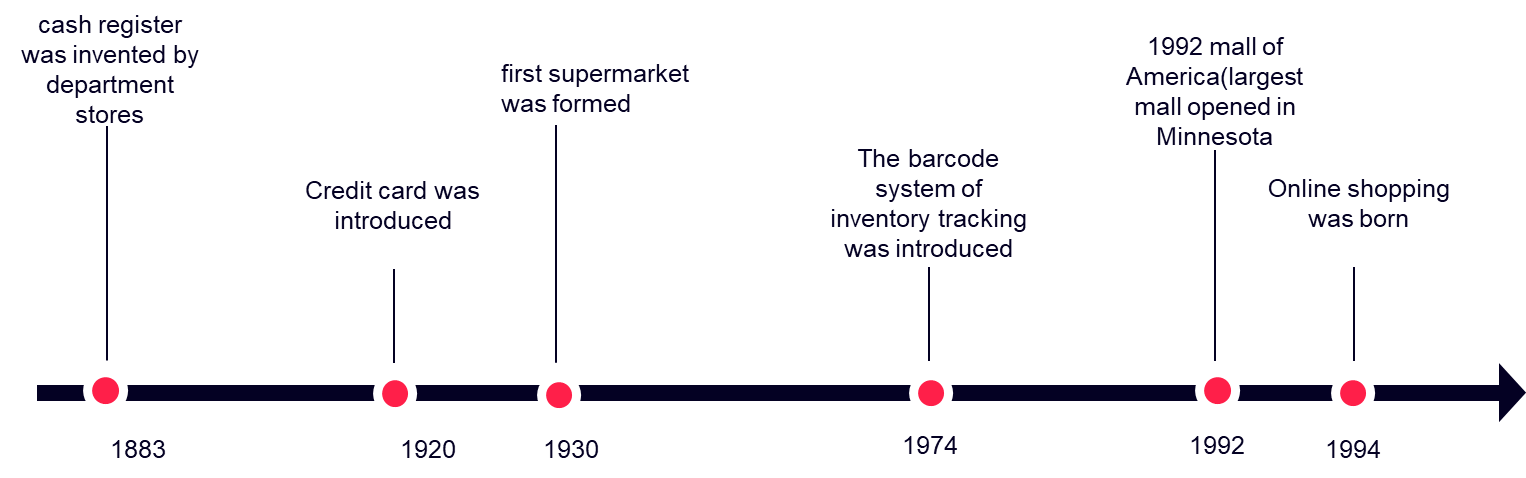
Cost plus retailers cater to a segmented mass market, trying to maintain mediocre margins balancing the importance of price and quality ,durability and other factors that go into ensuring customer satisfaction . Staples and Best Buy are common examples.

* **Premium Retailer**

Premium retailers focus on extremely niche markets and try to build their brand equity through improvement in prestige, quality, and performance much more than price.

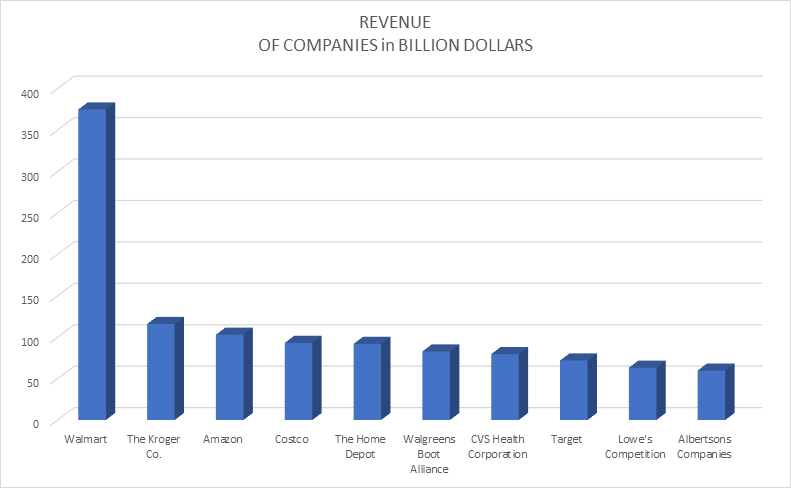
**HISTORY/PIVOTAL MOVEMENTS:**

Fig .6.3:Evolution Of Retail Stores

* 

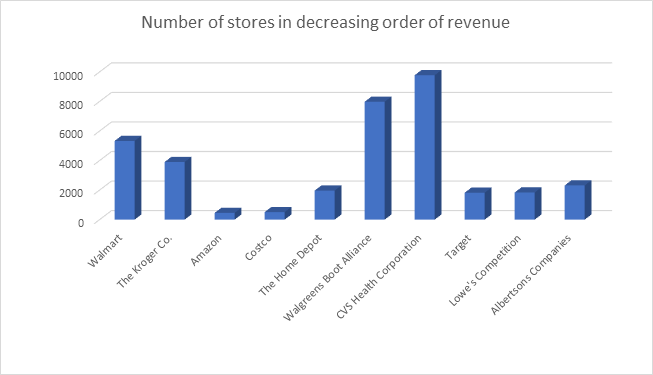
Source: <https://www.marketing91.com/10-types-of-retail-stores/>

* **KEY PLAYERS AND THEIR MARKET SHARE:**
* Fig .6.4:Revenue Of Retail Stores



Source: Revenue of companies(Billion Dollars. Source : https://stores.org/stores-top-retailers-2018

* Fig .6.5:Revenue Of Retail Stores



Source: Revenue of companies(Billion Dollars. Source : https://stores.org/stores-top-retailers-2018

**CHALLENGES AND OPPORTUNITIES:**

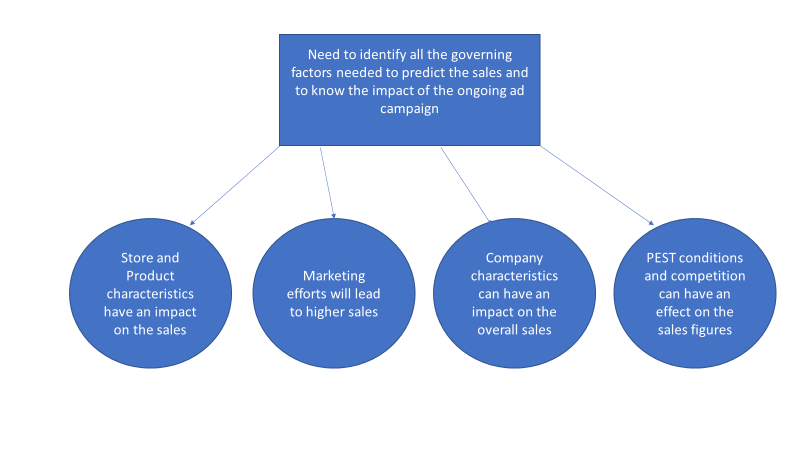
|  |  |  |
| --- | --- | --- |
|  | **Etailers** | **Brick and Mortar stores** |
| **Challenges** | **An absence of Online Identity Verification:** When a visitor visits an e-commerce website and signs up, the portal is unaware of the customer except for the information he/she entered. | **Engaging via Mobile**: The brick & mortar store needs to have a mobile presence not just to gain sales but to actively gain a lead customer stores |
| **Shopping cart abandonment:** Shopping cart abandonment is one of the biggest issues faced by eCommerce businesses today | **Reverse Showrooming:** Customers examine past customer feedback online and then head to the retail stores for purchase. |
| **Competitor Analysis:** In this competitive world, there will be too many competitors who will be offering the same products and services as you. Unless you have the best strategy that differentiates yourself from other competitors; it will become difficult to survive. | **Showrooming**: Customers walk into stores and examine the products and then cross check the products online and via their cellphones. |
| **Stuck at the old school way of selling:** Most of them lack the necessary insight into customer behavior and their buying patterns which can help them strive in the current e-commerce scenario. | **Customer Service and Checkout:** Capturing customer data in-store to retarget online, viewing the online profile of a client and targeting similar profiles. eReceipts that connect the offline experience back online. |
| **Opportunities** | **Inventory Cost Minimizing**: E-commerce can reduce inventory costs through just in time (JIT) system which increases firm’s ability to produce and forecast demand properly. | **Conveniance and speed emphasis**: Online shoppers have to wait a while for their purchases to arrive whereas brick-and-mortar merchant provide instant gratification to customers. |
| **Providing Better Customer Service:** E-commerce is an approach which is able to deliver proper sllengerservices to customer according their demand which increase customer satisfaction. | **The personal connection:** Getting help and advice in store from one of your pharmacy staff members can make it significantly more likely that a customer will make a purchase |

Table 1:Challengers and Opportunities table

**CREATION OF MECE(MUTUALLY EXCLUSIVE AND COLLECTIVELY EXHAUSTIVE TREE):**

Understanding of the problem can be done by creation of a MECE tree .The obvious gap is placed on the top most node .In an attempt to solve the problem by examining the list of all the exhaustive solutions , four factors have been identified on how to boost the sales for Easybuy.

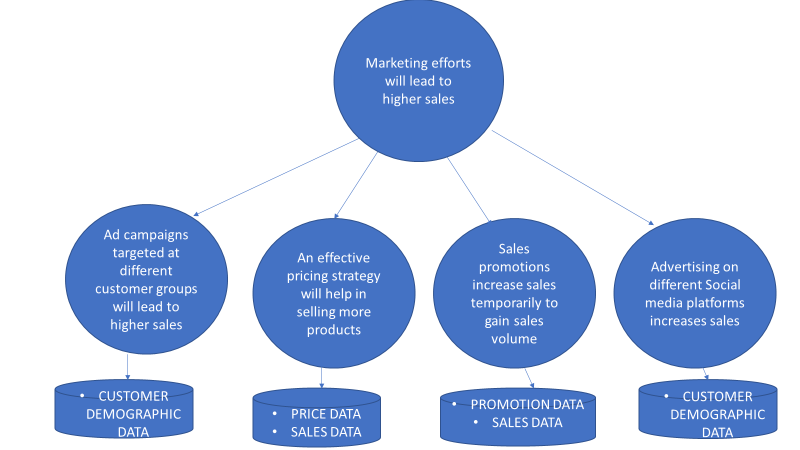
* Fig .6.6:MECE tree-Primary node

****

Source: Compiled by author

The node corresponding to marketing efforts is examined by enlisting the probable strategies . Marketing strategies form the foundations for boosting sales of Easybuy. Corresponding data that would be required to gather adequate insights are below.

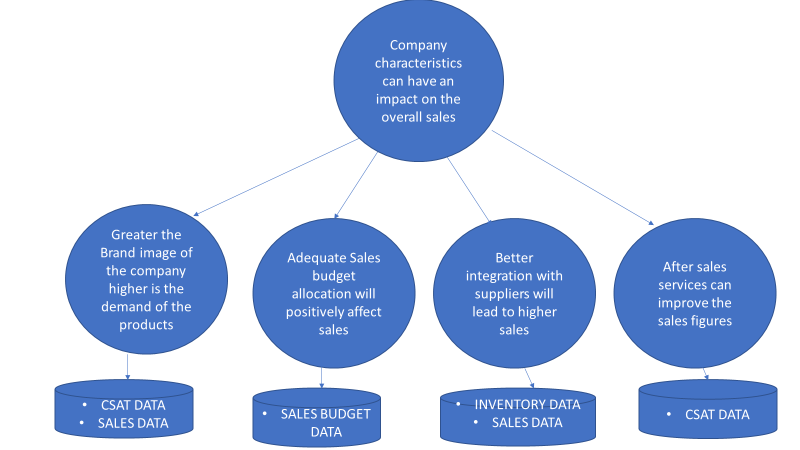
* Fig .6.7:MECE tree-Marketing node

****

Source: Compiled by author

The node corresponding Company characteristics is examined by enlisting the probable strategies .Decisions made by the board of the directors in accordance with the vision of the company must be analyzed .. Corresponding data that would be required to gather adequate insights are below.

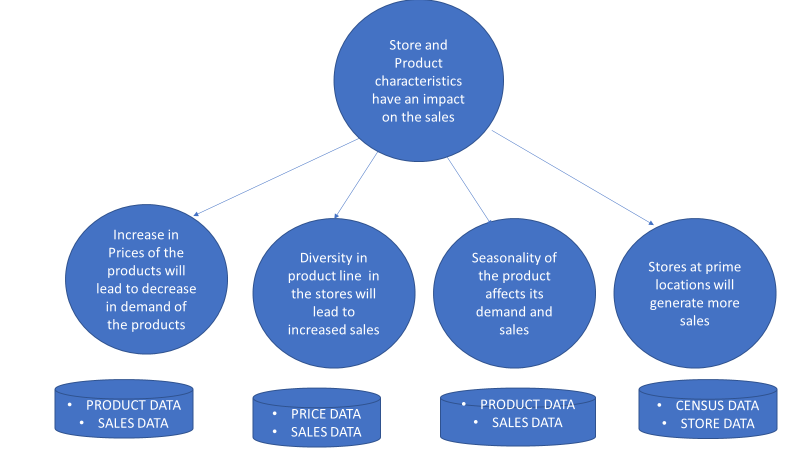
* Fig .6.8:MECE tree-Company charecteristics node

****

Source: Compiled by author

The node corresponding store and product characteristics is examined by enlisting the probable solutions .POS machines ,location , seasonality within the sale of products is listed below. Corresponding data that would be required to gather adequate insights are below.

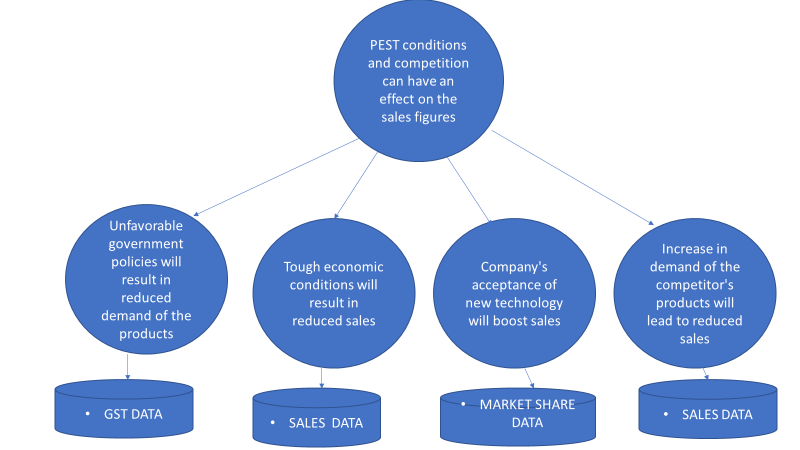
Fig .6.9:MECE tree-Store charecteristics node

****

Source: Compiled by author

The node corresponding PEST(POLITICAL ECONOMIC SOCIAL TECNOLOGICAL) conditions is examined by enlisting the probable strategies .Macro-Economic Conditions play a huge role in this . Corresponding data that would be required to gather adequate insights are below.

Fig .6.10:MECE tree-PEST charecteristics node

****

Source: Compiled by author

**Codes:**  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
sqoop import --connect jdbc:postgresql://10.1.2.60/testing --username ######## --password ######## --table PROMO\_DATA\_49 --target-dir /user/emanuel.thomas/problem\_space/PROMO\_DATA\_49 -m 1 -- --schema Demand\_Forecasting  
sqoop import --connect jdbc:postgresql://10.1.2.60/testing --username ########--password ########--table SALES\_DATA\_49 --target-dir /user/emanuel.thomas/problem\_space/SALES\_DATA\_49 -m 1 -- --schema Demand\_Forecasting  
sqoop import --connect jdbc:postgresql://10.1.2.60/testing --username ########--password ########--table STORE\_LOCATION\_49 --target-dir /user/ emanuel.thomas /problem\_space/STORE\_LOCATION\_49 -m 1 -- --schema Demand\_Forecasting  
sqoop import --connect jdbc:postgresql://10.1.2.60/testing --username ########--password ########--table STORE\_PRODUCT\_49 --target-dir /user/ emanuel.thomas /problem\_space/STORE\_PRODUCT\_49 -m 1 -- --schema Demand\_Forecasting  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
promo = spark.read.csv("/user/emanuel.thomas/problem\_space/PROMO\_DATA\_49")  
sales = spark.read.csv("/user/emanuel.thomas/problem\_space/SALES\_DATA\_49")  
store\_location = spark.read.csv("/user/ emanuel.thomas /problem\_space/STORE\_LOCATION\_49")  
store\_product = spark.read.csv("/user/ emanuel.thomas /problem\_space/STORE\_PRODUCT\_49")  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Checking for primary keys  
------------------------------  
promo.select("month").distinct().count()

sales.select("date").distinct().count()  
sales.select(concat(col('date'),lit(''),col('product'))).distinct().count()  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
promo\_columns = ['row\_names','month','promo\_1','promo\_2','promo\_3','promo\_4','promo\_5','promo\_6','promo\_7','promo\_8','promo\_9','promo\_10','promo\_11']  
promo=promo.toDF(\*promo\_columns)

sales\_columns = ['row\_names','date','product','sales']  
sales=sales.toDF(\*sales\_columns)

store\_location\_columns = ['row\_names','store','location']  
store\_location=store\_location.toDF(\*store\_location\_columns)

store\_product = store\_product.subtract(store\_product.limit(1))  
store\_product\_columns = ['row\_names','store','product\_1','product\_2','product\_3']  
store\_product=store\_product.toDF(\*store\_product\_columns)  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
promo.show()  
sales.show()  
store\_location.show()  
store\_product.show()  
------------------------  
store\_prod1= store\_product.select("row\_names","store","product\_1")  
store\_prod2= store\_product.select("row\_names","store","product\_2")  
store\_prod3= store\_product.select("row\_names","store","product\_3")

store\_prod = (store\_prod1.union(store\_prod2)).union(store\_prod3).drop("row\_names")  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
store\_loc\_prod = store\_location.join(store\_prod, store\_location.store == store\_prod.store).drop(store\_prod.store)  
store\_loc\_prod.show()  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
promo.show()  
sales.show()  
store\_loc\_prod.show()  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
from pyspark.sql.functions import \*  
sales = sales.withColumn("date", to\_date("date", "MM/dd/yyyy"))

sales= sales.select("row\_names","date","product","sales",year(sales.date).alias('dt\_year'), month(sales.date).alias('dt\_month'), dayofmonth(sales.date).alias('dt\_day'))  
sales.show()

sales =sales.select("row\_names","date","product","sales","dt\_year",format\_string("%02d", "dt\_month").alias('dt\_month'),"dt\_day")  
#sales =sales.select("row\_names","date","product","sales","dt\_year","dt\_month","dt\_day",format\_string("%02d", "dt\_month").alias('dt\_month'))

sales = sales.withColumn('month',concat(col('dt\_year'),lit(''),col('dt\_month')))  
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
sales\_promo = sales.join(promo, sales.month == promo.month).drop(promo.row\_names).drop(promo.month)

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------  
ps1\_data = sales\_promo.join(store\_loc\_prod, sales\_promo.product == store\_loc\_prod.product\_1).drop(store\_loc\_prod.product\_1).drop(store\_loc\_prod.row\_names)  
ps1\_data = ps1\_data.select("row\_names","date","location","store","product","dt\_year","dt\_month","dt\_day","month","promo\_1","promo\_2","promo\_3","promo\_4","promo\_5","promo\_6","promo\_7","promo\_8","promo\_9","promo\_10","promo\_11","sales")  
ps1\_data.show()

**6.2 TO ANALYZE AND INTERPRET CUSTOMER TRANSACTION DATA OF A CAB AGGREGATOR TO IDENTIFY THE LOCATION GENERATING MAXIMUM REVENUE WITH HIGHEST**.

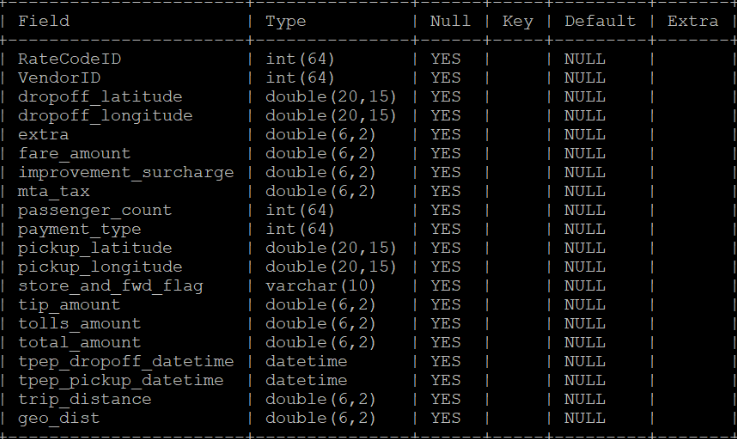
A Well-known cab Aggregator has approached mu-sigma and has brought forward their data dump towards the mu-sigma for an analysis of their data pertaining to their customer transactions. The data is semi-structured and has 28 lakh observations. Insights and suggestions pertaining to the data is of paramount importance

**6.2.1DATA DESCRIPTION**

The details of all the columns and their respective datatypes for the table nyc cab details.

The picture shows the details of the tables:-

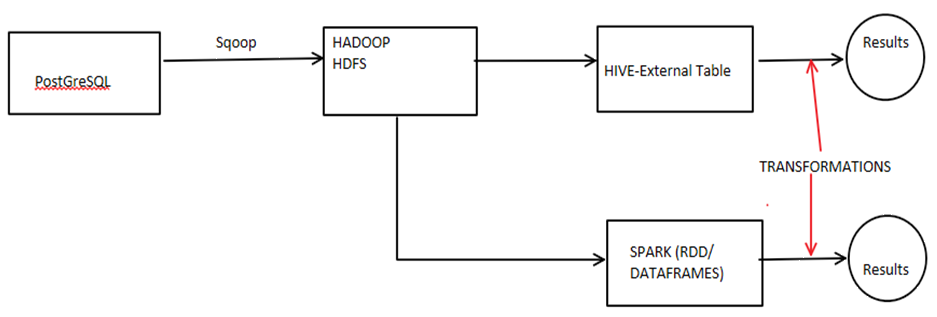
Fig .6.11:Self made Data Dictionary



Source: Compiled by author

**6.2.2 DATA ARCHITECTURE**

Fig .6.12: HDFS functioning



Source: Compiled by author

**CODE:**

IMPORTING DATA TO PYSPARK FROM Hadoop environment:

from pyspark.sql import SparkSession --- package for getting sparksession

The file is then imported in a format where in, every field is of a string data type.

------------------Importing the file ----------------------------------------------------------------------------------------------------------------------------------------------------------------

data=spark.read.csv("/user/emanuel.thomas/abcd/part-m-00000")

---importing the semistructured data file from hadoop.

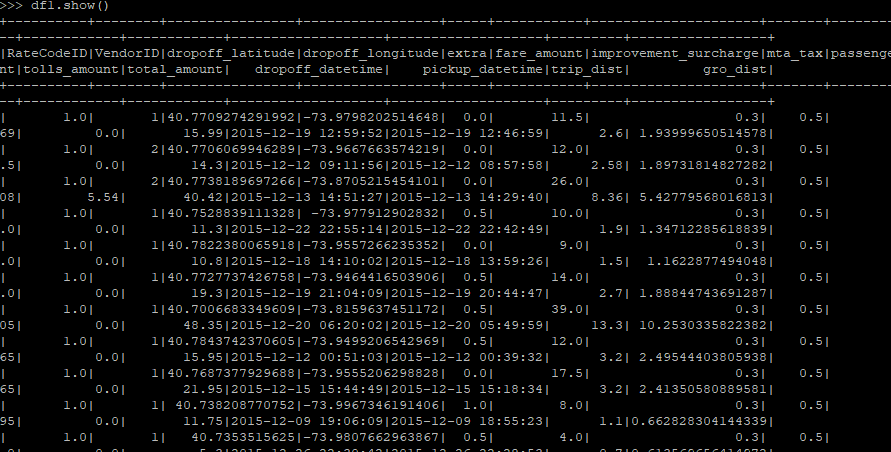
--- renaming the columns and creating a structure dataframe for values seperated by commas -----------------------------------------------------------------------------------------

df1 = data.selectExpr("\_c1 as RateCodeID","\_c2 as VendorID","\_c3 as dropoff\_latitude","\_c4 as dropoff\_longitude","\_c5 as extra","\_c6 as fare\_amount","\_c7 as improvement\_surcharge",

"\_c8 as mta\_tax","\_c9 as passenger\_count","\_c10 as passenger\_type","\_c11 as pickup\_latitude","\_c12 as pickup\_longitude","\_c13 as store\_and\_fwd\_flag","\_c14 as tip\_amount","\_c15 as tolls\_amount",

"\_c16 as total\_amount","\_c17 as dropoff\_datetime","\_c18 as pickup\_datetime","\_c19 as trip\_dist","\_c20 as gro\_dist")

Fig .6.13: HDFS functioning



Source: Compiled by author

df1.printSchema() --- returns the list of variables and their data types , introduction for summary statistics

df1.limit(20).show() ---show the top 20 variables

df1.dropna() ---drop missing values .

df1.dtypes

from pyspark.sql.types import FloatType

/\*conversion to float\*/

df1 = df1.withColumn("RateCodeID", df1["RateCodeID"].cast('float'))

df1 = df1.withColumn("VendorID", df1["VendorID"].cast('float'))

df1 = df1.withColumn("dropoff\_latitude", df1["dropoff\_latitude"].cast('float'))

df1 = df1.withColumn("dropoff\_longitude", df1["dropoff\_longitude"].cast('float'))

df1 = df1.withColumn("extra", df1["extra"].cast('float'))

df1 = df1.withColumn("fare\_amount", df1["fare\_amount"].cast('float'))

df1 = df1.withColumn("improvement\_surcharge", df1["improvement\_surcharge"].cast('float'))

df1 = df1.withColumn("mta\_tax", df1["mta\_tax"].cast('float'))

df1 = df1.withColumn("passenger\_count",df1["passenger\_count"].cast('float'))

df1 = df1.withColumn("passenger\_type",df1["passenger\_type"].cast('float'))

df1 = df1.withColumn("pickup\_latitude", df1["pickup\_latitude"].cast('float'))

df1 = df1.withColumn("store\_and\_fwd\_flag", df1["store\_and\_fwd\_flag"].cast('float'))

df1 = df1.withColumn("tip\_amount", df1["tip\_amount"].cast('float'))

df1 = df1.withColumn("tolls\_amount", df1["tolls\_amount"].cast('float'))

df1 = df1.withColumn("total\_amount", df1["total\_amount"].cast('float'))

df1 = df1.withColumn("dropoff\_datetime", df1["dropoff\_datetime"].cast('float'))

df1 = df1.withColumn("pickup\_datetime", df1["pickup\_datetime"].cast('float'))

df1 = df1.withColumn("gro\_dist", df1["gro\_dist"].cast('float'))

df1 = df1.withColumn("trip\_dist", df1["trip\_dist"].cast('float'))

df1 = df1.withColumn("pickup\_longitude", df1["pickup\_longitude"].cast('float'))

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

from pyspark.sql.functions import \* --- imports the package to import round functions

df1 = df1.withColumn("dropoff\_latitude", round(df1["dropoff\_latitude"], 3)) ---roudning the dropoff latitude to 3 points

df1 = df1.withColumn("dropoff\_longitude", round(df1["dropoff\_longitude"], 3)) ---roudning the dropoff longitude to 3 points

df1 = df1.withColumn("pickup\_longitude", round(df1["pickup\_longitude"], 3)) ---roudning the pickup longitude to 3 points

df1 = df1.withColumn("pickup\_latitude", round(df1["pickup\_latitude"], 3)) ---roudning the pickup latitude to 3 points

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

import pyspark.sql.functions as func ---imports the library that is used to perform aggregate functions to arrive at the problem statements

df2=df1.groupby('pickup\_latitude','pickup\_longitude').agg(max("total\_amount").alias("revenue")) ----return the location that generates maximum revenue

df3=df1.groupby('pickup\_latitude','pickup\_longitude').agg(sum("passenger\_count"))

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

df4=df1.groupby('dropoff\_latitude','dropoff\_longitude').agg(sum("total\_amount")

df5=df1.groupby('dropoff\_latitude','dropoff\_longitude').agg(sum("passenger\_count")).filter("count>= 10").sort('count', ascending=False)

----------------------------------------------------------------------------------------------------

ydata2=df1.groupby('pickup\_latitude','pickup\_longitude').agg(sum('total\_amount').alias("Revenue"),sum("passenger\_count").alias("Frequency")) ----returns the location that generates maximum frequency in ascending order

data3=data2.orderBy(desc("Revenue"))

----------------------------------------------------------------------------------------------------

data2=df1.groupby('dropoff\_latitude','dropoff\_longitude').agg(sum("passenger\_count").alias("Frequency")) ---returns the location that gives the least frequency in descending order

data4=data2.orderBy(desc("Frequency"))

----------------------------------------------------------------------------------------------------

**6.3 TO IDENTIFY AND ASSIGN POLARITY TO DATA BASED ON CUSTOMER REVIEWS ACROSS A E-RETAIL WEBSITE.**

**SENTIMENT ANALYSIS:**

This involves monitoring the comments posted on social media sites. Certain comments made by the consumers are excessively long.Breaking these comments into fragments separated by delimiters is crucial.

Fig .6.14: Sentiment Analysis

Source: Pluralsight.com



1. Once separated into fragments , each fragment is categorized into a particular categories.
2. Polarity is assigned based on positive or negative feedback. The words are classified into certain categories .Assigning polarity to the words is important.
3. The lexicon comprises of

* Operational categories : Comprises of certain categories relating to the product specificities and service.
* Non-operational categories comprising of queries, feedback and generic brand.

In [101]:

##importing the necessary libraries for performance of necessary funtions

import pandas as pd

import numpy as np

import pandas.io.sql as psql

import nbconvert

import psycopg2

from wordcloud import WordCloud, STOPWORDS

import matplotlib.pyplot as plt

%matplotlib inline

conn=psycopg2.connect("dbname='postgres'user='msu\_user'host='10.1.2.60'password='Password@1234'")

df1=psql.read\_sql("""SELECT \* FROM "msu\_greco"."product\_reviews" limit10""",conn)

In [87]:

## Checking out the values of the entire dataframe

df1.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 7980 entries, 0 to 7979

Data columns (total 20 columns):

row.names 7980 non-null object

ID 7980 non-null object

Title 7980 non-null object

Published 7980 non-null object

Inserted 7980 non-null object

Author Name 7980 non-null object

Author ID 7980 non-null int64

Author Age 7980 non-null object

Author Location 7980 non-null object

Author Sex 7980 non-null object

URL 7980 non-null object

Site Info Name 7980 non-null object

Site Info URL 7980 non-null object

Country 7980 non-null object

Search Term 7980 non-null object

Item name 7980 non-null object

Review\_rating 7980 non-null int64

Item URL 7980 non-null object

Review\_Content 7980 non-null object

Market Segment 7980 non-null object

dtypes: int64(2), object(18)

memory usage: 1.2+ MB

## filtering the dataset for only value of amazon.in returns a single value

df2=df1[df1['Site Info Name']=='amazon.in'] ## filtering for only amazon.in

df2.head(2)

Out[88]:

TABLE 3

filter for only walmart.com

df3=df1[df1['Site Info Name']=='walmart.com']

df3.head(40)

TABLE 4

checking for influenster.com

df4=df1[df1['Site Info Name']=='influenster.com']

df4.head(2)

Out[90]:

##As per the problem statement it is required to filter as per the 4 websites alone

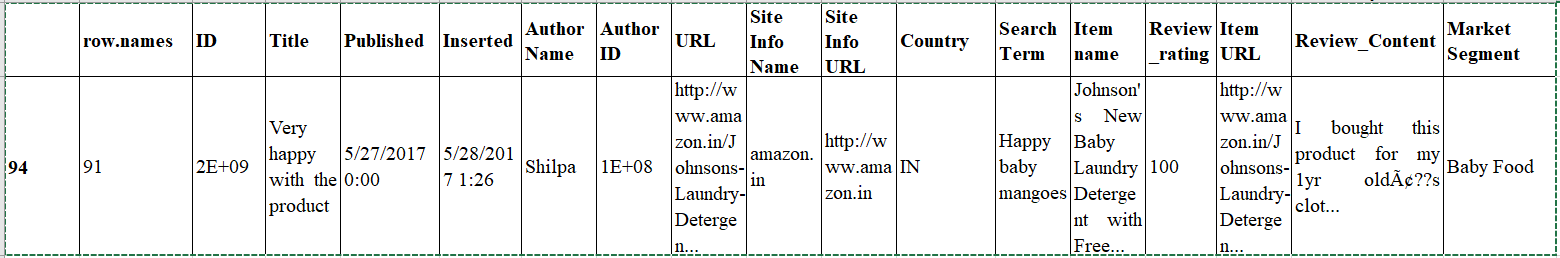
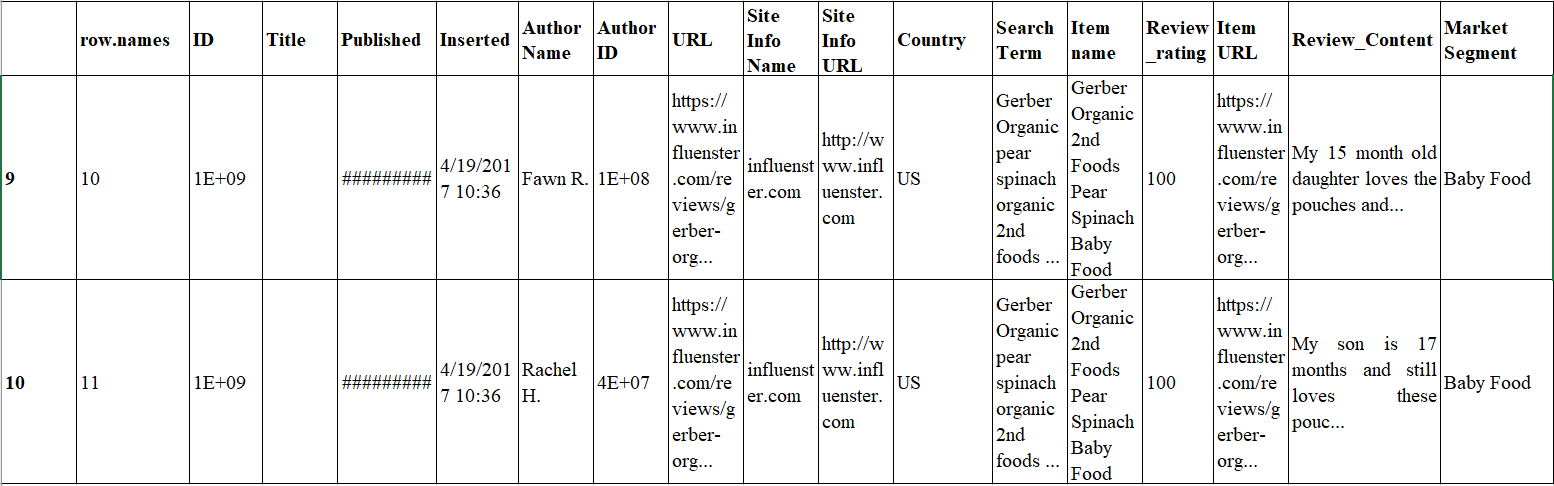
a=['walmart.com','amazon.in','Target','influenster.com']

df5=df1.loc[df1['Site Info Name'].isin(a),:]

df5.info()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 4769 entries, 0 to 7976

Data columns (total 20 columns):

row.names 4769 non-null object

ID 4769 non-null object

Title 4769 non-null object

Published 4769 non-null object

Inserted 4769 non-null object

Author Name 4769 non-null object

Author ID 4769 non-null int64

Author Age 4769 non-null object

Author Location 4769 non-null object

Author Sex 4769 non-null object

URL 4769 non-null object

Site Info Name 4769 non-null object

Site Info URL 4769 non-null object

Country 4769 non-null object

Search Term 4769 non-null object

Item name 4769 non-null object

Review\_rating 4769 non-null int64

Item URL 4769 non-null object

Review\_Content 4769 non-null object

Market Segment 4769 non-null object

dtypes: int64(2), object(18)

memory usage: 782.4+ KB

## Further it is required to segment as per the market segment

b=['Baby Food','Beverages and Alcohols','Fruit Purees and Snacks','Sauces and Dressings']

df\_3=df5.loc[df5['Market Segment'].isin(b),:] ## isin is used to see if the values are within a list or not

## Keep columns only for review\_rating and review content

df\_4=df\_3.loc[:,('Review\_rating','Review\_Content')]

df\_4.info()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 4769 entries, 0 to 7976

Data columns (total 2 columns):

Review\_rating 4769 non-null int64

Review\_Content 4769 non-null object

dtypes: int64(1), object(1)

memory usage: 111.8+ KB

## Remove the duplicate values

df\_4[df\_4.duplicated()].info()

df\_4.drop\_duplicates(inplace=True)

df\_4.head()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 956 entries, 2 to 7835

Data columns (total 2 columns):

Review\_rating 956 non-null int64

Review\_Content 956 non-null object

dtypes: int64(1), object(1)

memory usage: 22.4+ KB

Out[94]:

|  | **Review\_rating** | **Review\_Content** |
| --- | --- | --- |
| **0** | 100 | I received the Gerber Organic Apple Wild Blueb... |
| **1** | 100 | I received a sample of Gerbers new organic app... |
| **3** | 80 | just recently started giving my daughter somew... |
| **4** | 100 | My daughter hates veggies but she absolutely l... |
| **5** | 100 | I received a free sample from PinchMe. I was s... |

Table 3:

## removing the lesser than

df\_4=df\_4[df\_4['Review\_Content'].str.len()<=800]

In [96]:

##7. Giving polarity to the words for sentiment analysis

df\_4['rating']=np.where(df\_4['Review\_rating']>=50, 'positive', 'negative') #using the numpy library we segregate the values into positive or negative values

df\_5=df\_4[['rating','Review\_Content']]

df\_5

Out[96]:

|  | **rating** | **Review\_Content** |
| --- | --- | --- |
| **0** | positive | I received the Gerber Organic Apple Wild Blueb... |
| **1** | positive | I received a sample of Gerbers new organic app... |
| **3** | positive | just recently started giving my daughter somew... |
| **4** | positive | My daughter hates veggies but she absolutely l... |
| **5** | positive | I received a free sample from PinchMe. I was s... |
| **...** | ... | ... |
| **7973** | positive | I purchase fruit from Dole on a weekly basis. ... |
| **7974** | positive | My kids absolutely love these. And they are de... |
| **7975** | positive | I have found these Welch's Fruit Snacks made w... |
| **7976** | positive | I LOVE fruit snacks! But they can be so unheal... |

Table 4:

3727 rows × 2 columns

In [98]:

##8.Creating a word cloud

comment\_words = ' '

stopwords = set(STOPWORDS)

# iterate through the csv file

for val in df\_5.Review\_Content:

# typecaste each val to string

val = str(val)

# split the value

tokens = val.split()

# Converts each token into lowercase

for i in range(len(tokens)):

tokens[i] = tokens[i].lower()

for words in tokens:

comment\_words=comment\_words+words+' '

wordcloud = WordCloud(width = 800, height = 800,

background\_color ='white',

stopwords = stopwords,

min\_font\_size = 10).generate(comment\_words)

plt.figure(figsize = (8, 8), facecolor = None)

plt.imshow(wordcloud)

plt.axis("off")

plt.tight\_layout(pad = 0)

plt.show()

Fig .6.15: Word Cloud



Source: Compiled by Author

**CHAPTER 7**

**FINDINGS AND RECOMMENDATIONS**

**7.1. DATA PREPERATION:**

Data is obtained in a single table which is a 3.5 normalized form . This is necessary for proper demand forecasting .The table has four columns :-

|  |  |  |  |
| --- | --- | --- | --- |
| **DATE** | **STORE** | **PROMO** | **SALES** |
| 22-04-2014 | STORE A | PROMO A | 1417 |
| 23-04-2014 | STORE A | PROMO A | 1000 |
| 24-04-2014 | STORE A | PROMO A | 895 |
| 25-04-2014 | STORE A | PROMO A | 850 |
| 26-04-2014 | STORE A | PROMO A | 944 |
| 27-04-2014 | STORE A | PROMO A | 1133 |
| 28-04-2014 | STORE A | PROMO A | 1063 |
| 29-04-2014 | STORE A | PROMO A | 1000 |

Table 5.

Source: Compiled by author

There are 4 types of stores and 6 types of promos. Models can be prepared for all the stores or all the promos by including them in a linear model with sales. The objective of bringing a table in this format is to be able to check for violations of the sales column in a single attempt for multifarious combinations of stores ,promos and sales . Initially there existed 4 tables, however by using spark and efficient joining techniques is performed on the tables to obtain a singular table .As shown in the diagram, the resultant table is the best possible arrangement for our analysis . Checking for violations can be performed smoothly on the data without any interruptions. The violation tests that can be performed are Durbin Watson(autocorrelation), Breusch pagan test( hetoeroskedascity) and shapiro wilk test ( for normality).

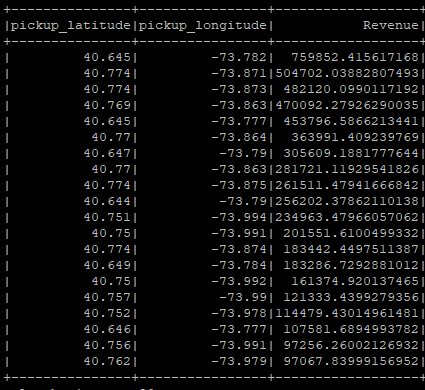
Checking for seasonality and cyclicity in the corresponding time series data is essential and the obtained model is devoid of cyclicity. The obtained residuals are independent of each other and this is necessary. It is important to convert the data to a stationary data and then perform analysis . This is necessary to satisfy one of the assumptions of conducting a linear regression.

**7.2INSIGHTS PERTAINING TO CAB AGGREGATOR SERVICE:**

* + 1. **INSIGHTS PERTAINING TO REVENUE IN PICKUP :**

**Locations that are arranged in decreasing order of revenue collection:**

Fig .7.1:Locations in decreasing order of revenue

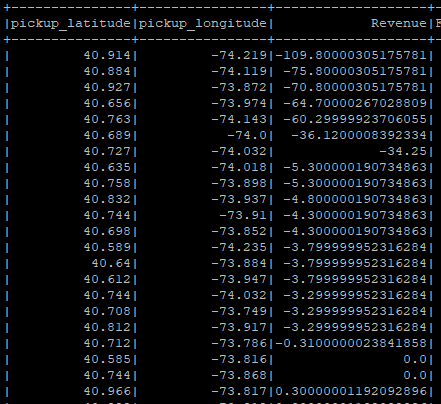


Source: Compiled by author

* After ordering the top pickup spots in descending order of revenue generated, maximum revenue is generated from **John F Kennedy airport** at Pennsylvania, which is obvious, closely followed by the **la Guardia airport.**
* Madam Tussauds
* The iconic museum of modern art is amongst the top 10 along with Pennsylvania station

**Locations arranged in increasing order of revenue :**

Fig .7.2:Locations in increasing order of revenue



Source: Compiled by author

* The areas that incur maximum lost are in the order

1. Barnyard carriage house in totowa district.
2. Artic ice manufacturing in Garfield district adjacent to plauderville.
3. Lincoln park
4. Windsor terrace
5. Columbia street waterfort district

**Rectification advised :**

Losses are in the range of 30 to 110 dollars.It is adviced that service be stopped in these areas.

* + 1. **INSIGHTS PERTAINING TO FREQUENCY IN PICKUP :**

**LOCATIONS THAT ARE MOST FREQUENT :**

Fig .7.3:Decreasing order of locations in frequency



Source: Compiled by author

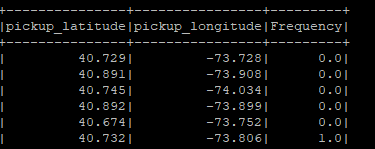
* The most frequent locations in increasing frequency and most crowded locations:

1. Pennsylvania station
2. John F Kennedy airport
3. Manhattan Hall
4. East Elmhurst near the la Guardia airport .

Scaling up operations to meet the existing demand of passengers and fending off competition is a probable solution.

**PLACES WHICH ARE LEAST IN FREQUENCY:**

Fig .7.4:Increasing order of locations in frequency



Source: Compiled by author

The place where the frequency of pick-up is almost none or one is given in the list below:

1. Bellerose manor
2. Riverdale
3. Hoboken
4. Manhattan college adjacent to Riverdale
5. Pomonok

Cost of operations in such areas might be unable to break even as the frequency is very less.The decimal place within the latitude and longitude component is rounded off to a precision of 3 places as this precision corresponds to a perimeter of a mile .The interest of the client is to evaluate which are the possible locations that generate maximum and minimum revenue as well the determining the frequency of pick-up and drop off. This arrangement of data is a complex data , when the data is stored across 18 lakh rows in a 400gb ram of dynamic memory . The technique was to use concepts from data engineering , such as parallel processing .The data is critical for the client and is maintained by Mu-sigma servers. The data is the confidential property of the client . It is Mu-Sigma’s responsibility to look after such data servers and maintain the data in a structured format . Maintenance of this data is necessary as it enables mu-sigma to gain insights .

The areas that would not have performed well would be a sunk cost and services could be stopped in such areas.

Transformation of the data is vital , as the data is first received in a raw format . The data is then converted to a semi structured format where operations pertaining to sorting and arranging the data can be done . Transformation of the data is not done in the statistical sense such as applying logarithmic transformations etc. but is done at a very grassroot level of making the data fit for analysis and this is the basic cleaning the data .Being a consulting company, Mu-sigma advises its clients on strategies that can boost revenue and minimize lost cost by stopping services in areas that incurr losses.

**CHAPTER 8**

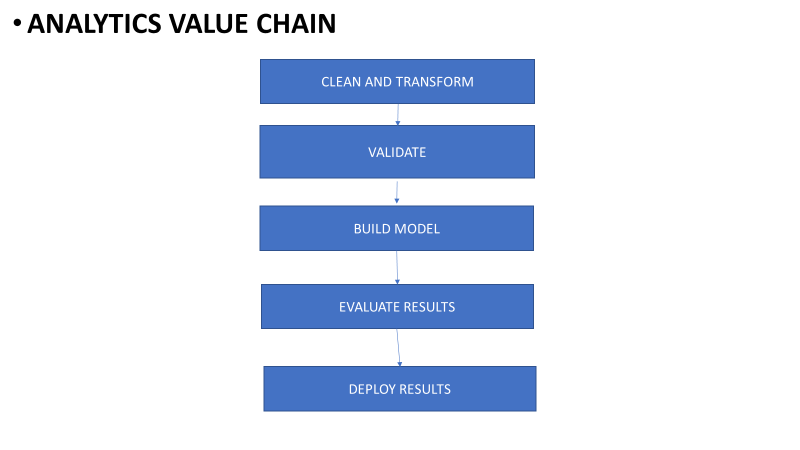
**LEARNINGS**

There exist 4 types of analytics :-

1. Predictive
2. Descriptive
3. Inquisitive
4. Prescriptive

Whatever be the type of analytics , all follow the similar process flow as shown in the diagram.

Fig .8.1:Analytics Value Chain

**** Source: Compiled by author

However none of the above methodologies can be adopted at a commercial level without data engineering. Data engineering as a subject was created due to lackluster database maintenance. Real world decision making comprises a combination of knowledge, information and wisdom. Feedback comprises knowledge, information and data.Three categories of databases exist namely unstructured, semi-structured and structured databases.

Unstructured databases comprise of PDF, JPEG, MP3 and movies, while semi structured databases consist of CSV files, mongo DB etc.

Structured database consists strictly of 2-dimensional data.

Semi structured data requires the attributes be mentioned again and again. Based on problem complexity we can opt for structured or semi-structured databases. Structured data can be converted to unstructured data, however the vice versa cannot be done.

Trade-offs exist even in this domain, if we want a database that gives faster retrieval than we opt for an unstructured database. If we want a database that is meant for analytics, then we use the relational model. Structured data sets are meant for analytical data sets.

CRISP model:

The data understanding is dependent on the business understanding. The steps following data understanding are: -

1.data understanding

2.data preparation

3.Modelling

4.Evaluation

5.Deployment.

Data ingestion is getting data, transformation and loading. Data lake is a dump of data which could be structured and unstructured. Data catalog is useful for determining the source and it gives an overview of basic information of the database.

Big data involves data that is costly and extremely important to an organization. The limit is decided by the technology. If python can't handle it, then some other technology is used. Majority of the data comes in the form of photos. The 4V s of data comprise of Volume, Variety, Velocity and veracity. The train explains that veracity is the most critical as unauthentic information is available across the internet. Concepts of Hadoop includes effective usage of unused space and distributed processing. Fault tolerance is improved due to replication of data.

There exist 3 components of distributed processing:

1.HDFS (Hadoop Distributed File System) used for storage: - placed in a master node i.e. the main computer.

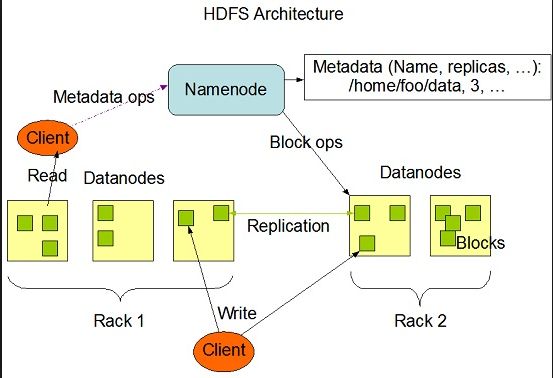
2.Mapreducer which defines the logic of data processing. It is used for mapping the nodes across which the data would be stored and for returning the values to the user, it zeroes down on the

3YARN: Framework to run the processing task (resource allocator)

Hadoop ecosystem comprises of HIVE (SQL like query interface), HBASE (a different kind of a database), PIG (Unstructured to structured database)

It uses the concept of distributed processing where data is stored in parallel across several nodes. Parallelism is promoted as it improves efficiency and faster processing. At Mu-Sigma, each node allows a storage of 128 MB, the data is split into chunks and stored across the nodes. Mu-sigma believes that the MapReduce slows down the speed of retrieval and hence this induction of apache spark which operates on python.

Essentially apache SPARK works in collaboration with HADOOP. HADOOP runs on commodity hardware, which involves the slave nodes.



Source: https://www.dezyre.com/article/hadoop-architecture-explained-what-it-is-and-why-it-matters/317

Data redundancy is the repetition of observations within certain mentioned attributes.

Insertion, updating and deletion cannot be performed when data redundancy exists.

Normalization helps in reducing data redundancy.

Criterion for a table being in 1st normalized form:

1.Each column must hold a single value.

2.Each attribute should hold only particular values.

3.Each attribute must have a unique name.

2nd normalized form:

1.Should have a primary key.

2.Should not have partial dependency

3rd normalized form:

1.Should not have transitive dependency.

D3 stack components:

1.Data engineering

2.Data Science

3.Decision Science.

Types of database: -

1.hierarchical: each node has a single parent node.

2.Network: each node might have one or more parent nodes.

3.Entity

4.Relational: The level of data is a primary key.

**CHAPTER 8**

**REFERENCES**

**REFERENCES:**

1. <https://craft.co/mu-sigma>
2. <https://yourstory.com/2013/04/the-billion-dollar-mu-sigma-story-with-founder-dhiraj-c-rajaram>
3. <https://en.wikipedia.org/wiki/Mu_Sigma#History>
4. “Service innovation and smart analytics for Industry 4.0 and big data environment” by Jay Lee, Hung-An Kao, Shanhu Yang
5. <https://www.analyticsindiamag.com/analytics-data-science-industry-in-india-study-2018-by-analytixlabs-aim/>
6. “Business Intelligence and analytics :From big data to big impact “ by Veda C Storey, Roger H.L Chiang and Hsinchun Chen

**CHAPTER 9**

**APPENDIX**