BIG DATA – AN ANALYSIS

Abstract

The data retrieved from various sources such as Facebook, Twitter, YouTube, messages, call logs, history in various forms like structured, unstructured and semi-structured are simply known as big data. Big data plays a vital role in business for making good decisions and making effective strategies. This chapter discusses data, complex data, big data, objectives and tolls for data analysis. This chapter also discusses data pre-processing with proper examples. This is also a highlight challenge. There are many challenges to be addressed. The new algorithms are to be designed to handle diverse types of data. The new algorithms or methodologies should support retrieving hidden patterns efficiently. Business and organizations or governments have to address methodological developments in knowledge discovery and systems and application with regard to using and integrating large data sets.

Researchers from numerous disciplines, including computer science, health, data science and social and policy issues must also collaborate on big data analytics across education institutions, government and society.

This chapter discusses the methods and tools used for big data analysis as well as applications with real-time examples.

Keywords: Big data, Structured, Semi-Structured, Strategies, algorithms, hidden pattern

Authors

Mrs. Viji Parthasarathy
Assistant Professor
Department of Computer Science
Shrimati Indira Gandhi College
Trichy, Tamil Nadu, India

Mrs. R. Indra
Assistant Professor
Department of Computer Science
Shrimati Indira Gandhi College
Trichy, Tamil Nadu, India

I. INTRODUCTION

Data is simply known as raw details. We may define data as uncooked or unprocessed. Why do we need to discuss the data now? Once upon a time, many people in the world in remote places searched for food, shelter and dress. Nowadays, just by clicking the buttons, data is showered. Even for food to eat. People search through Google, analyse and pick the suitable shops and items according to their needs. "No Mobile -No people" changed to "No data -No people". Data become a basic need of the people. As the population grows, data is also growing exponentially.

1. **Size:** The size of the data starts with a bit which has the value 0 or 1. Next is a nibble, which has 4 bits and a sequence of 8 bits, which is known as a byte. In this way, the following table shows the size of the data in detail.

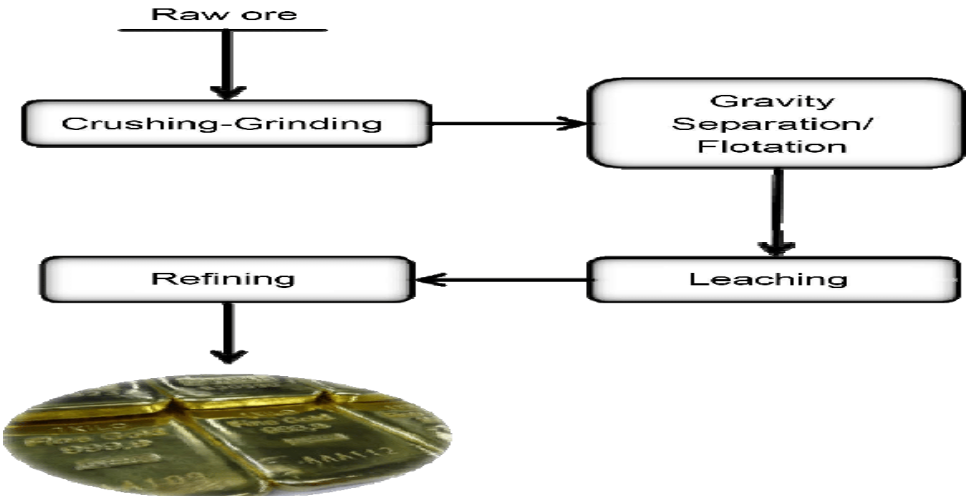
Name	Equal to	Size in Bytes
Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8 bits	1
Kilobyte	1,024 bytes	1,024
Megabyte	1,024 kilobytes	1,048,576
Gigabyte	1,024 megabytes	1,073,741,824
Terrabyte	1,024 gigabytes	1,099,511,627,776
Petabyte	1,024 terrabytes	1,125,899,906,842,624
Exabyte	1,024 petabytes	1,152,921,504,606,846,976
Zettabyte	1,024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1,024 zettabytes	1,208,925,819,614,629,174,706,176

See the value of YB, it is huge. People in the world are using huge data for a variety of purposes. People need a variety of high volume of data in various formats with greater velocity. This data is known as Big Data, which has these three v's - variety, volume, velocity.

2. **Big Data Analysis:** Nowadays, all people analyse the data for picking clothes, electrical and electronic products, retails, Petty shops, Books, online classes, Netflix, etc, list cannot be ended.

As technology comes into our hands, data is much more important. People prefer onlinemode. They need to analyse the data for their needs.

Data analysis needs a number of steps. For example, we could extract gold by mining. But we cannot use it directly; it has to go through the process to a gold.



The above picture says the steps involved in getting gold. Similarly, data processing involves several steps to getting the necessary data.

3. **Why?:** It is the big question. People can understand the product/classes/movies, whatever they want, by analysing. On another side of business, people can understand the difficulties or issues by comparison, and improve their strategies. Business people need to attract customers, they need to retain the customers, should understand the customers by analysing the habits of site visiting and purchasing and they may predict the trends.

   For business people, it is a big boon. Because various tools and methodologies are available online. They may update their skills online. People get the sense of information and make better decisions when analysing the data.

II. OBJECTIVES AND TOOLS

Objectives are to be framed before entering the analysis. It could be anyone of the domains like education, engineering, finance and some other domain. These objectives help us to decide what data we need, what data to collect, what insights we get from it.

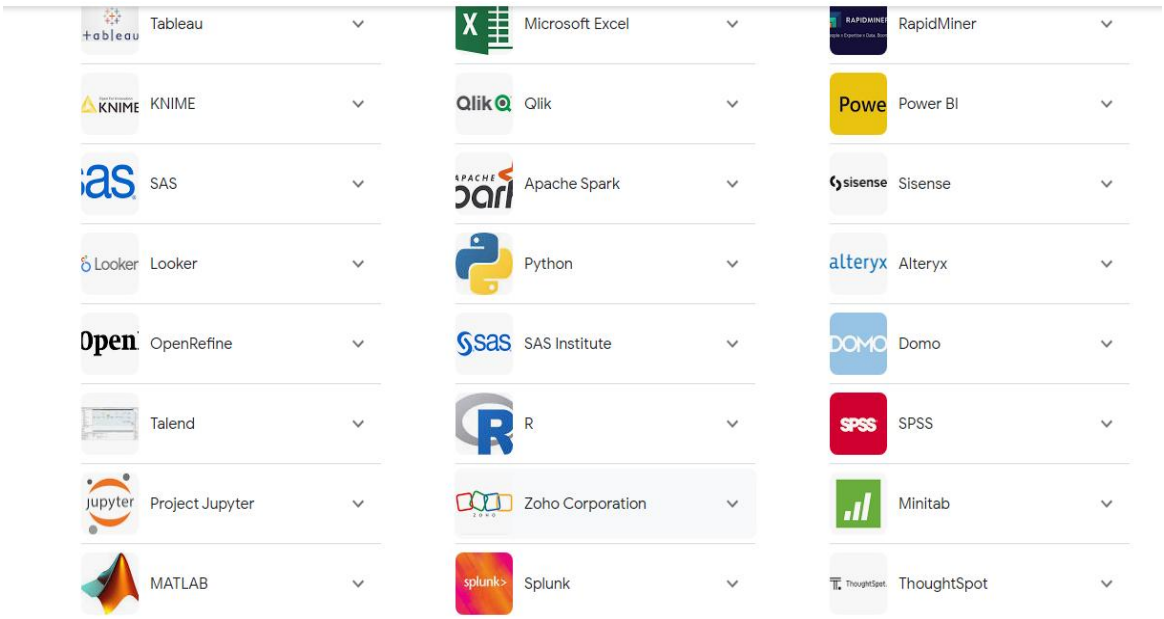
We should have the assurance that our data is good. We may get inaccurate or duplicate data which mislead and analysis will be incomplete.

After the collection of data, this data has to be cleaned and confirmed. The data is accurate and with no duplication.

Data is to be stored as a table. It could be stored spread sheets (Excel) Column names can be renamed. Any unwanted columns could be removed. Data is to be normalized, that is, data can be transformed into the form which we need. There are many tools available for data cleaning.

We have to adopt some standardised process to collect data and transform data. The data will be under new standards after refining or modifying it.

Next, we have to fix the data tool for analysing-which depends on quantity of data, types of data (structured data, unstructured data and semi-structured data). The data are qualitative (question answers, response forms.) or quantitative (in numbers).



III. DATA ANALYSIS TOOLS

Excel sheets and Weka tools can be used for a small scale of data, while BI, TABLEAU can be used for analysing a large scale of data. There are many tools, like Predictive tools, Data Modelling tools. Domain specific tools and data Visualization tools are available.

IV. STURCTURED/ UNSTRUCTURED DATA

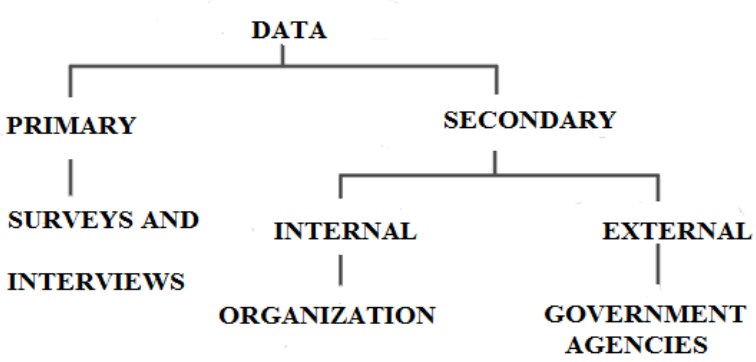
We have to look for trends or types of data. If the data is in numerical, we can go for charts or other visualisation techniques. Suppose data is unstructured like email, we have to go some other approaches. From unstructured data, structured data can be extracted by using text analysis tools.

1. Data Collection:

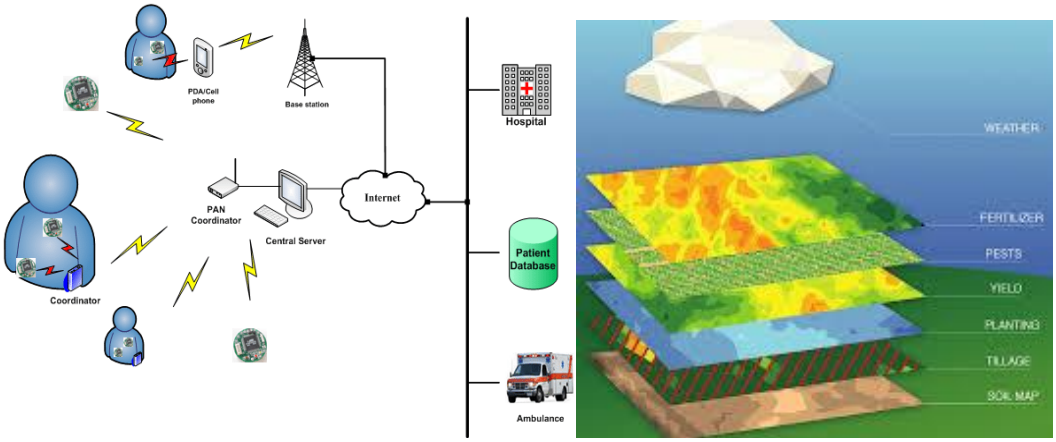
When the users start thinking about analysis, the following questions should be answered:

- Purpose of data collection
- Source of data
- Type of data

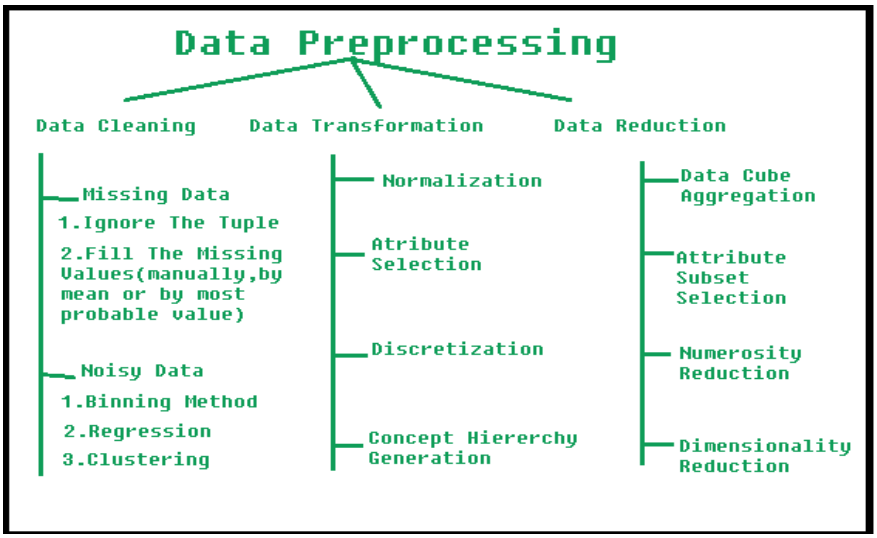
Types of data may be either qualitative or quantitative. Qualitative data refers to textual data like words and sentences. Quantitative refers to numerical data. The data can be further classified into primary and secondary data.



From the above diagram, the users can understand the types of data with examples. Primary data can be collected through surveys, interviews and observation methods, while the secondary data can be drawn from some internal and external sources and some other sources like sensor data, satellite data, and web traffic data.



2. **Pre-Processing:** The most important action before the actual usage of data is pre-processing. Pre-processing is the most important step for removing garbage for the analysis. Cleaning, transforming and integration are the steps followed in pre – processing of data.



3. Data Cleaning:

The above diagram represents the cleaning process.

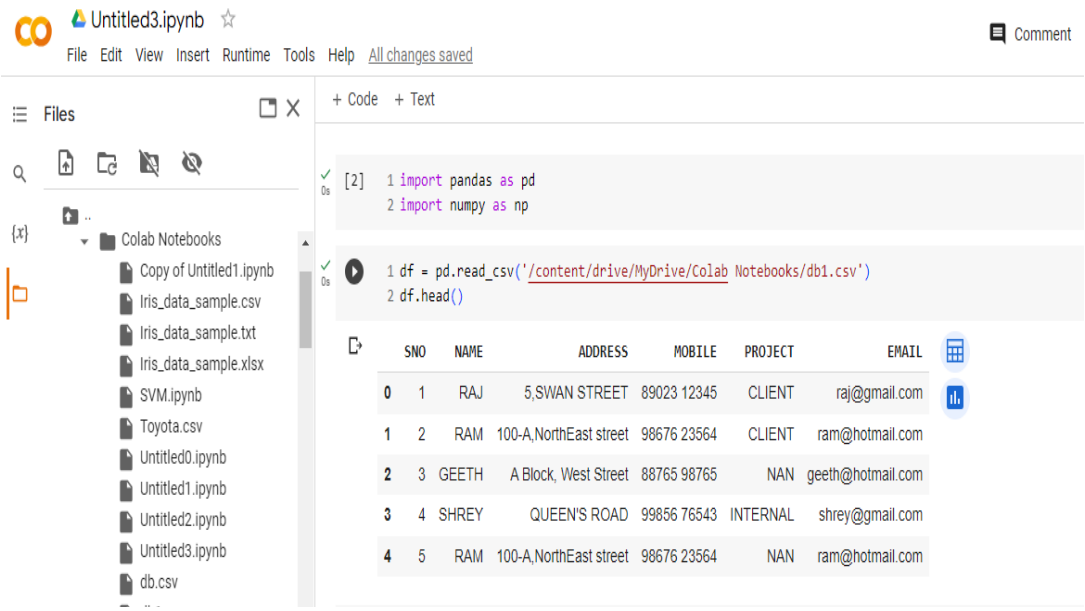
- Unwanted observations should be removed from the dataset; it could be duplicate or irrelevant data items.
- Naming conventions are to be done. It means fixing of structural errors.
- Grouping the similar data in to same categories or classes. For "Not Applicable" or NA
- Missing data are to be handled by applying methods like mean values and medians.

The following coding and outputs show the sample data cleaning process using Python

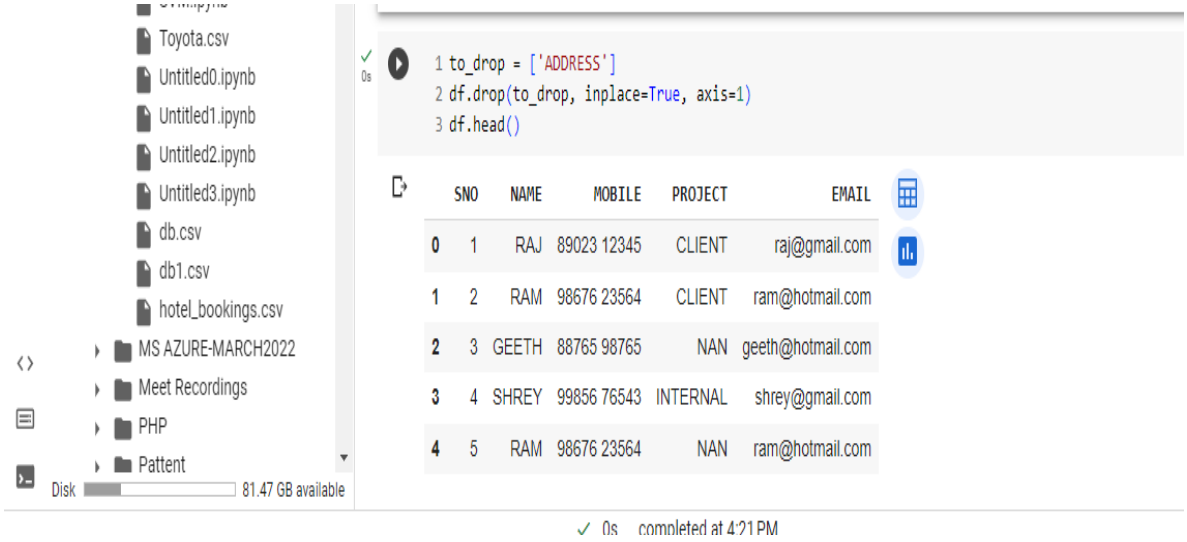
```
import pandas as pd
import numpy as np

df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/db1.csv')
df.head()
```

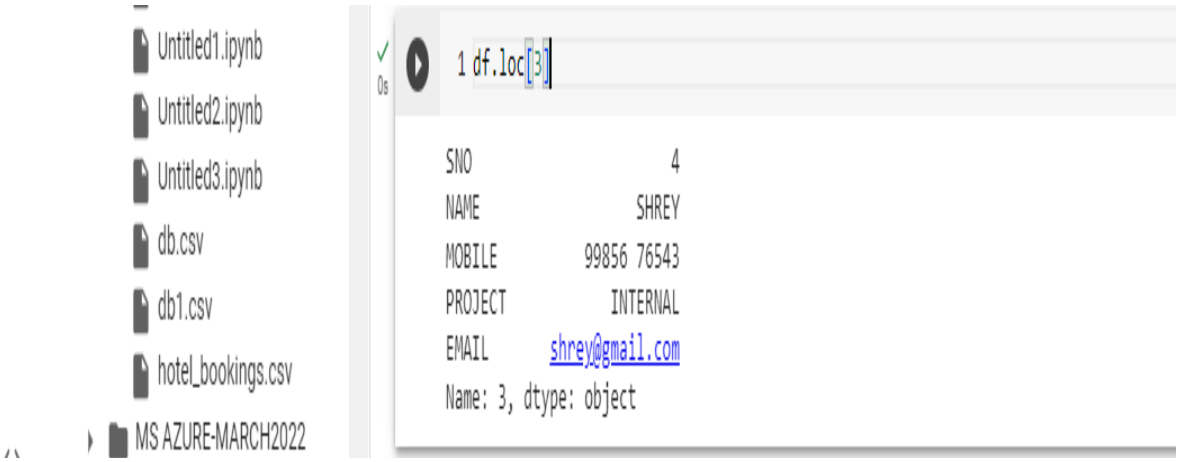
Table DB1 has been read, and shows the first 5 rows.



The particular attribute “ADDRESS” has been removed

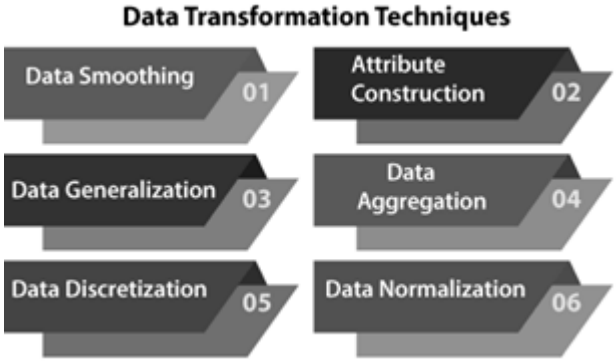


Shows the particular location

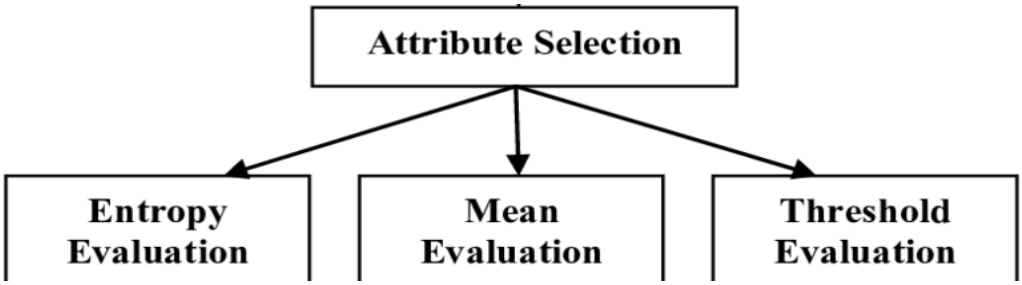


Removal of duplicates, in the below table “RAM” duplicates have been removed

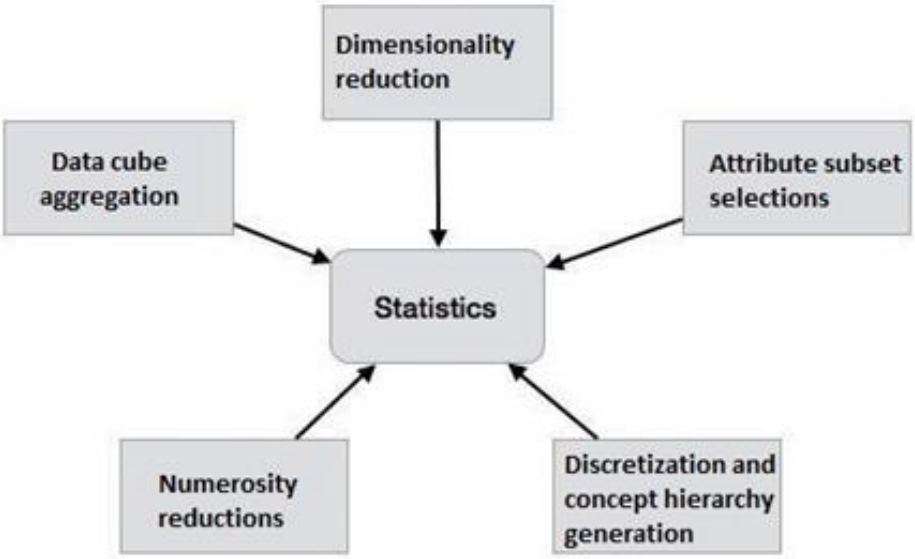
4. **Data Transformation:** Data Transformation means conversion of data into the required format. This involves the following actions.

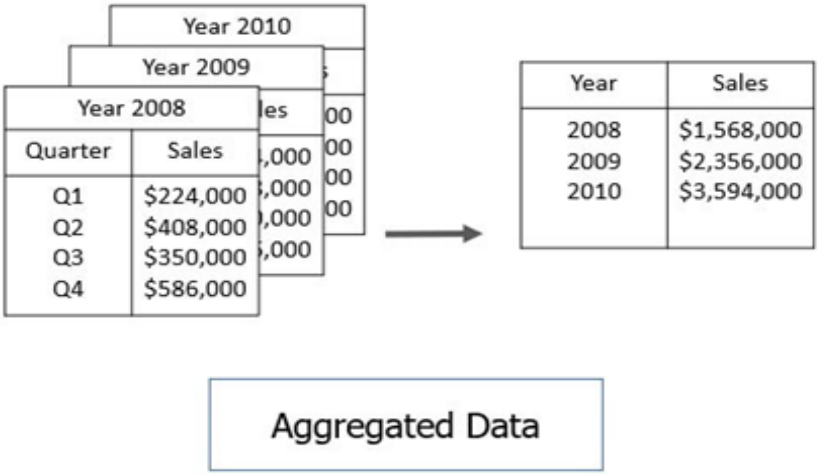


- **Normalization:** Convert the data into different scales and units.
- **Attribute Selection:** New features can be developed from the existing attributes to support the mining process



5. **Data Reduction:** This means reduction of the size of the data set without losing information. So that over fitting can be avoided.





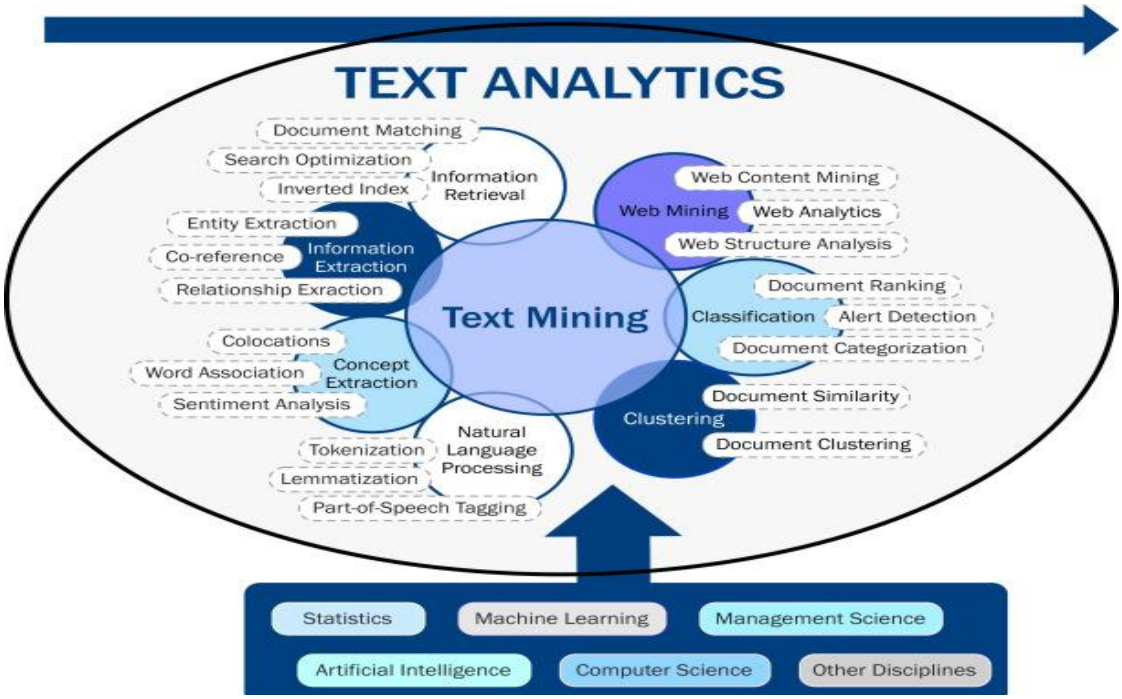
6. **Data Aggregation:** Aggregation stands for summarization of data from various resources. It is actually a statistics calculation on the purchasing habits, transactions of various age groups of people or customers. The company will recognize high profit-yielding products and low. The decision can also be made about the allocation of budget for marketing and development.

V. TEXT ANALYSIS

Text analysis using machine learning and parsing the text from the given text document. This is the process by which machines can understand human written text for business insights.

1. **Analysis using Monkey Learn:** Open-ended questions support analyzing customer's thoughts. Text analytics support understanding the content of different languages from different media, such as You Tube Video, Hash Tags, Twitter. It is a big challenge

because social media are a scalable. Even though it is faster to make decisions accurately compared to manual tasks. Very effective and very supportive to increase the revenue.



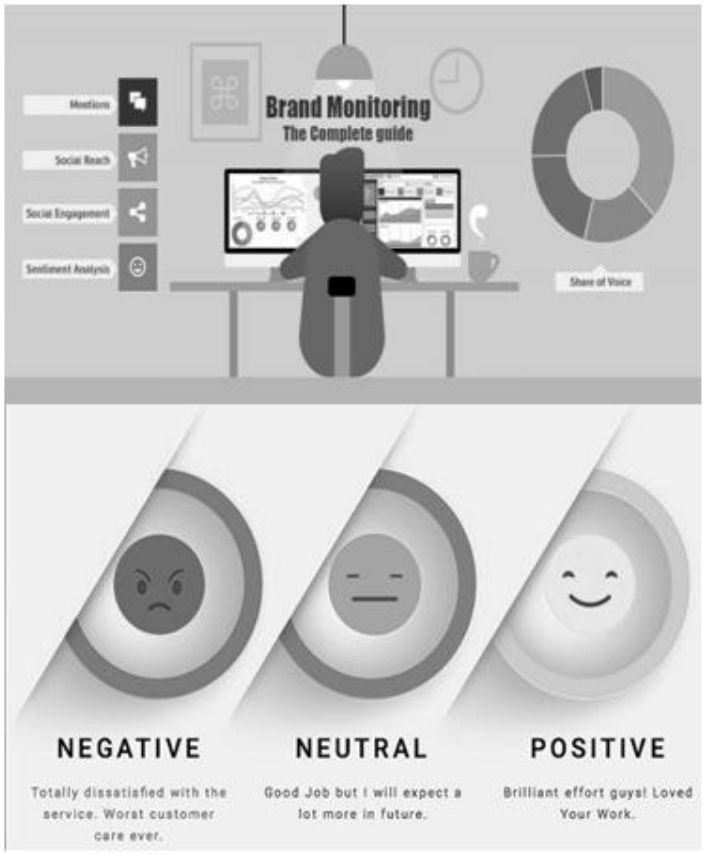
Example tools: MonkeyLearn, Google Cloud NLP

2. **Visualization of Data:** It is very supportive and useful. Charts inform data clearly. TABLEAU AND BI are wonderful tools.

Transformation of Health Care Data using TABLEAU

VI. SENTIMENT ANALYSIS

It classifies the sentiment in each text. It is very useful for data of social media, where business people may predict people's view or opinion on their product or brand. Business people gather data through what people like, dislike, comments, and tweets and analyse the data, change their business strategies to retain and attract their customers and people. Business people, for example, customer service frequently monitor the feedback of the customer to upgrade the brand. If business people find any unfavourable reactions, they address the issues immediately.



The feedback or responses are connected with human emotions like anger, happiness, irritation, unhappiness and more. Sentiment Analysis has also become part of people's lives, because sites are monitored while people are surfing. Brand Monitoring tools are used by several organizations.

Example: Brand24, Brandwatch.

Social Media Analytics Using Brand Watch

VII. POINTS TO REMEMBER WHILE ANALYSING

In general, in order to put your findings into perspective, compare your current data with previous performance.

If this isn't feasible, you may find it useful to look at industry benchmarks instead, since they may help you analyse your support performance or learn about a completely new product feature.

Make sure you remain open-minded when it comes to trends or data points that go against your expectations. In addition to looking at the raw data, you should also look for outliers.

As a result, you will be able to avoid cherry-picking findings that support your preexisting views. If you find anomalies in your data, you should investigate them further, as there may be a simple explanation.

Key Challenges with Big Data:

1. As the population grows, the data grows equally. The challenge enters here: how different types of data whether structured data , or unstructured data have to be stored.
2. Data has to be of high quality.
3. As big data grows. Trending technologies are to be updated.
4. An unpredictable format would be a challenge in analysing.

VIII. CONCLUSION

Even large organizations with well-established businesses should incorporate big data into their architecture. All the IT companies in various countries have been working on big data. Companies are eyeing Google, LinkedIn, Facebook and Twitter to upgrade their business and to understand their people. Many companies benefit by the use of big data for their business growth, and for tuning their strategies. Large volumes of data (big data) can be analysed with various tools like sentiment and text analytics tools. The important challenge is designing samples for big data and prediction models with security. Because Data or information is power and Data is oil. In the future, unpredictable data formats could be produced by people and nature. It is another challenge in designing.

REFERENCES

[1] J. P. Dijcks, "Oracle: Big data for the enterprise," *Oracle White Paper*, 2012.
[2] "Big Data Survey Research Brief," *Sas White Paper*, 2013.
[3] W.-H. Weng and W.-T. Lin, "A Scenario Analysis of Big Data Technology Portfolio Planning," in *International Journal of Engineering Research and Technology*, 2013.
[4] Big Data, Data Mining and Machine Learning, John Wiley & sons, 07_May-2014.
[5] Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda, O'Reilly Media, Inc., ISBN: 9781491963043, June 2018.
[6] https://www.javatpoint.com/data-transformation-in-data-mining
[7] https://www.electronicsmedia.info/2017/12/26/data-reduction-strategies-in-data-mining/
[8] Text Mining and Analytics using Natural Language Processing. Isha Gupta, Medium Digest

Image Source:

[9] https://www.javatpoint.com/data-transformation-in-data-mining
[10] https://www.electronicsmedia.info/2017/12/26/data-reduction-strategies-in-data-mining/
[11] Text Mining and Analytics using Natural Language Processing.Isha Gupta, Medium Digest

