**BIG DATA ANALYSIS: A STUDY**

**1Ms. VIJI PARTHASARATHY**

**2 Dr. S. REGHA**

**Asst. Professor, Department of Computer Science**

**Shrimati Indira Gandhi College, Trichy, Tamilnadu, India**

**ABSTRACT**

The data getting from various resources such as face book, twitter, you tube, messages, call logs, history in the 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 make effective strategies. This chapter discusses about data, complex data, big data, objective and tolls for data analysis. This chapter also discusses data pre-processing with proper examples. This also high lights challenge. There are many challenges to be addressed. The new algorithms are to be designed to handle diverse types of data. The new algorithm or the methodologies should support to retrieve the hidden patterns efficiently. Business, organization or government have to address to methodological developments in knowledge discovery and systems and application with regard to using and integrating large data set.

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.

Key Word: Big data, Structured, Semi structured, Strategies , algorithms, hidden pattern

**INTRODUCTION:**

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

**SIZE**

The size of the data starts from bit which has the 0 or 1. Next is Nibble which has 4 bits and sequence of 8 bits form a byte. In this way, the following table shows the size of the data in detail

|  |  |  |  |
| --- | --- | --- | --- |
| Abbreviation  B | Unit  Bit | Value  O or1 | Size(inbytes)  1/8ofabyte |
| B | byte | 8bits | 1byte |
| KB | Kilo Byte | 1,000bytes | 1,000bytes |
| MB | Mega Byte | 1,0002bytes | 1,000,000bytes |
| GB | Giga Byte | 1,000'bytes | 1,000,000,000bytes |
| TB | TeraByte | 1,000•byte | 1,000,000,000,000 bytes |
| PB | Peta Byte | s | l,000,000,000,000,000 bytes |
| EB | Exa Byte | 1,000'bytes | l,000,000,000,000,000,000bytes |
| ZB  YB | Zetta Byte  YottaByte | 1,000'bytes  1,000 Bytes | 1,000,000,000,000,000,000,000bytes  l,000,000,000,000,000,000,000,000  Bytes |

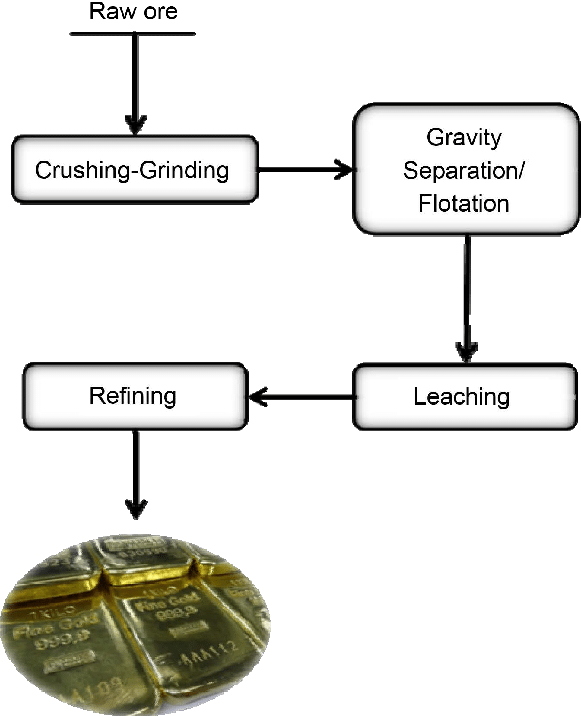
See the value of YB, it is huge. People in the world are using huge data for variety of purposes. People need variety of high volume of data in various formats with greater velocity. This data is known as Big Data which has these three v’s variety, volume, velocity.

**BIG DATA ANALYSIS**

Now a days, every 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 on our hand, data are much important. People prefer online mode very thing. They need to analyse the data for their need.

Data analysis needs a number of steps. For an example we may extract the gold by mining. But we cannot use 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 get necessary data.

**WHY**?

It is the big question. People can understand the product/classes/movies whatever they want if they analyses. In another side business people can understand the difficulties by comparison, and improve their strategies. Business people need to attract the customers, they need to retain the customers, should understand the customers by analysing the habits in site visiting, purchasing and they may predict the trends.

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

**OBJECTYIVES AND TOOLS**

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

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

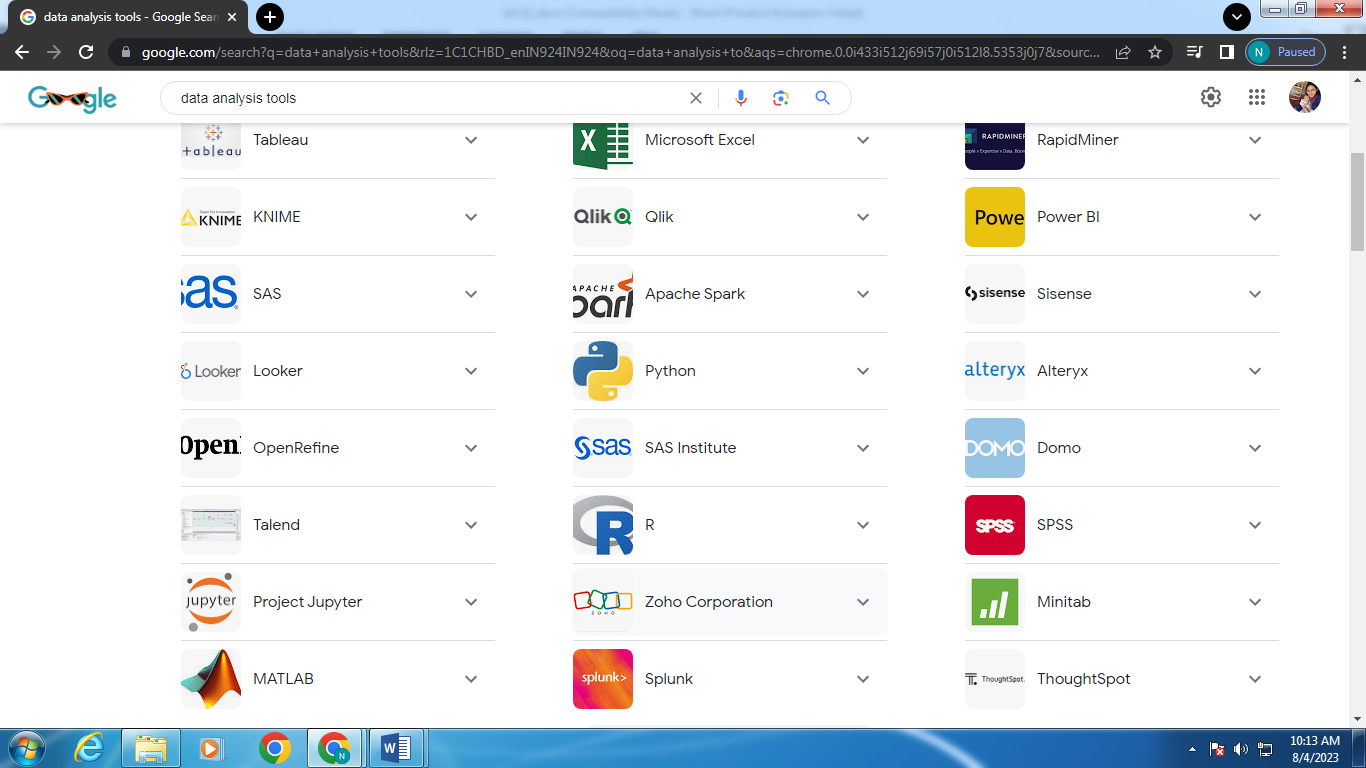
After the collection of data, these data to be cleaned and confirm the data are with accuracy and with no duplication.

Data are to be stored as a table. It could be stored in Spread sheets (Excel)

Column names can be renamed. Any unwanted columns could be removed. Data are to BE normalized that is data can be transformed to the form which we need. There are many tools are available for data cleaning.

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

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



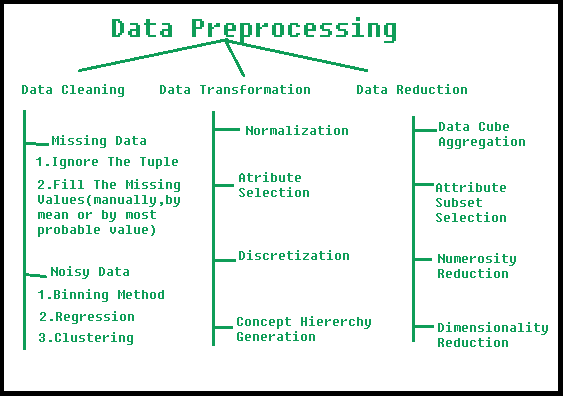
**DATA ANALYSIS TOOLS**

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

**STURCTURED/ UNSTRUCTURED DATA**

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

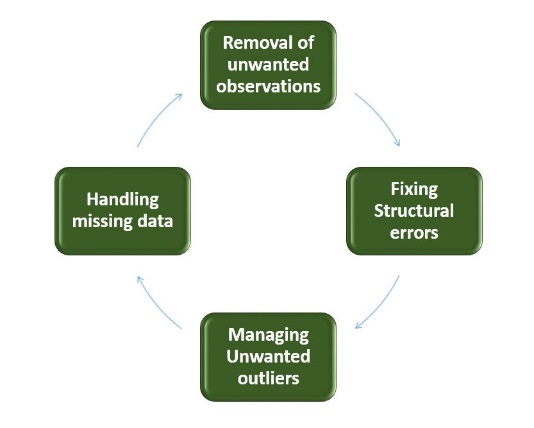
**PRE-PROCESSING**



Pre-processing is the most important step to remove garbage for the analysis.

Cleaning, transforming and integration are the steps followed in pre – processing of data.

**Data cleaning**:



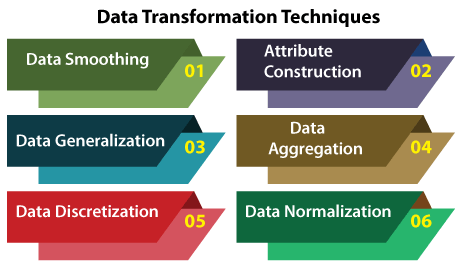
**T**he above diagram represents the cleaning process.

\* 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.

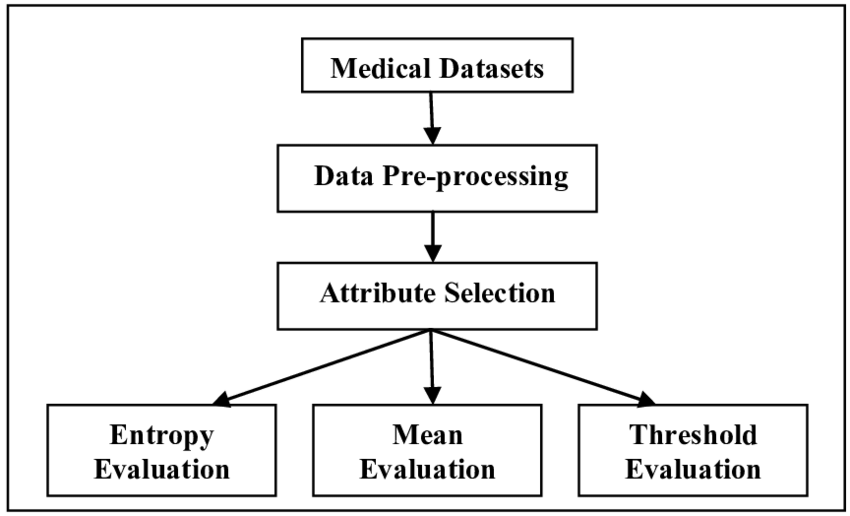
**Data Transformation**:



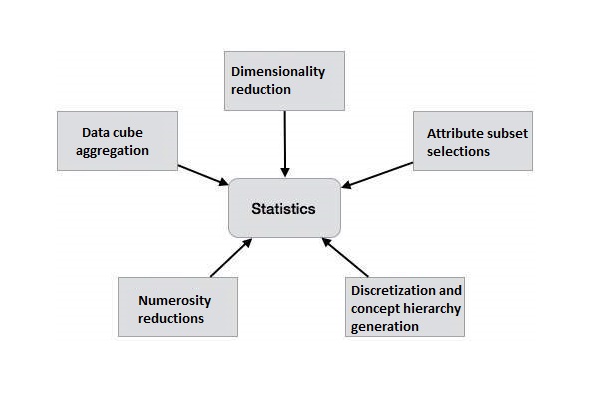
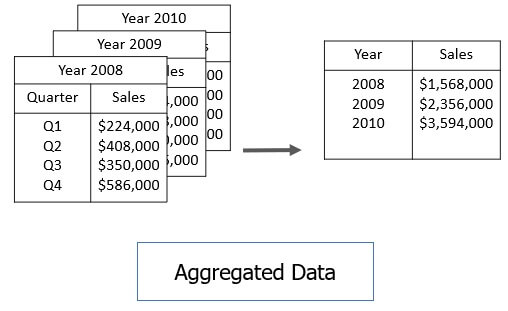
Data Transformation means conversion of data into required format. This involves the following actions.

**Normalization:** Convert the data into different scales and units.

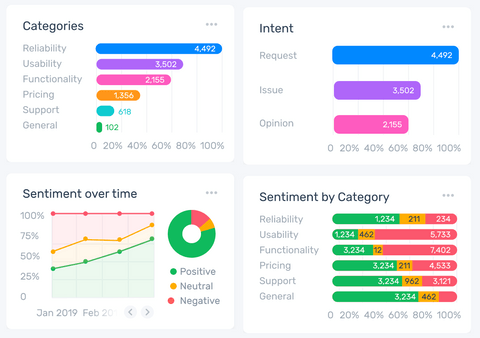
**Attribute Selection:** New features can be developed from the existing attributes to support mining process



**Data Reduction:**

This means, reduction of the size of the data set without losing information. So that over fitting can be avoided.

**Data Aggregation**:

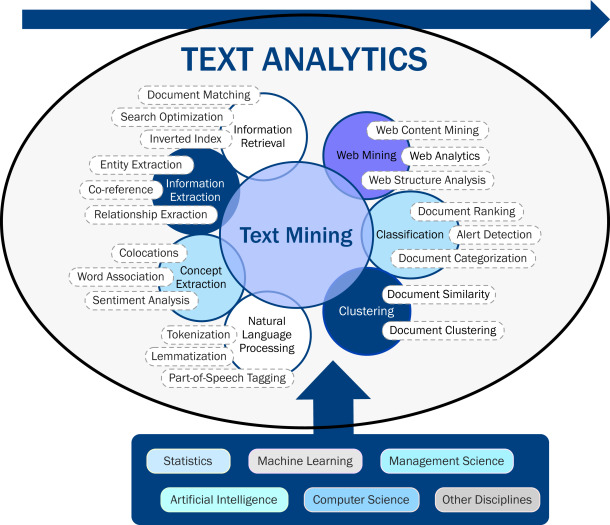
Aggregation stands for summarization of data from various resources. It is actually a statistics calculation on the purchasing habits, transactions of various age group of people or customers. The business will recognize the high profit yielding products and low. The decision also can be made for allocation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

**TEXT ANALYSIS**

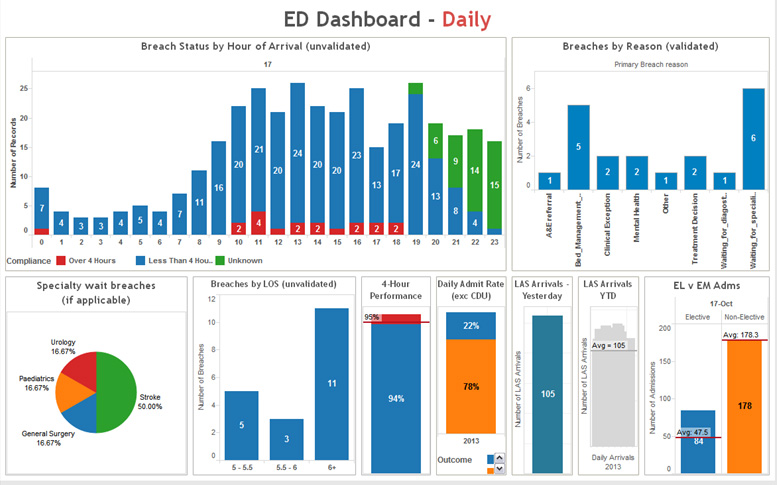
Text analysis using machine learning and parse the text from the given text document. This is the process by which machine can understand human written text for Business insights.

**Analysis in MonkeyLearn**

Open ended questions support to analyses customer’s thoughts. Text analytics support to understand the content of different languages from different media such as You Tube Video, Hash Tags, Twitter. It is a big challenge because social media is 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

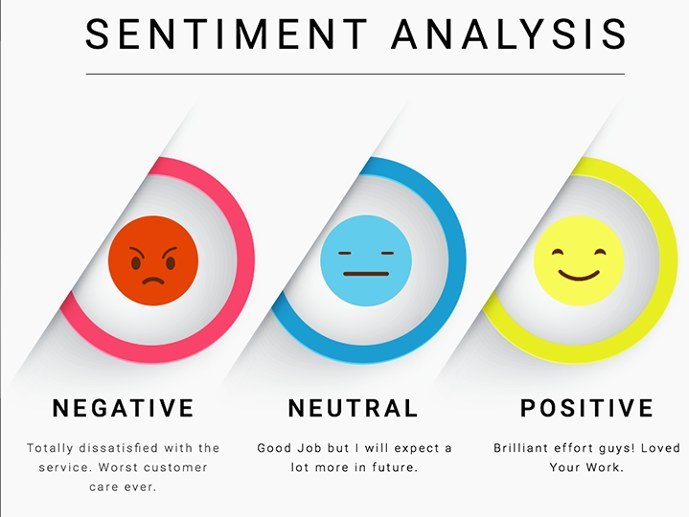
**Visualization of data**: It is very supportive and useful. Charts inform data clearly. TABLEAU AND BI are wonderful tools.

**Transformation of Health Care Data using TABLEAU**

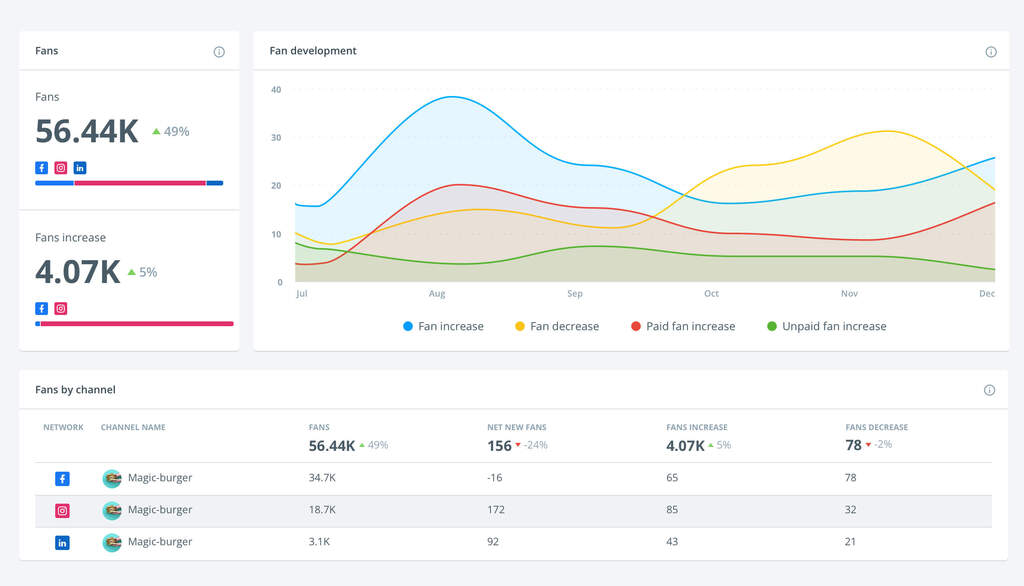
**SENTIMENT ANALYSIS:**

It classifies the sentiment each text.

It is very useful in social media where business people may predict the people’s view or opinion on their product or brand. Business people gather data through the 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 found 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 also become part of people life, as it is monitoring while people surfing. Brand Monitoring tools are used by the several organization.Example: Brand24, Brandwatch.



**Social Media Analytics Using BrandWatch**

**POINTS TO REMEMBER WHILE ANALYSING:**

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

When 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 population grows, data grows equally the data. Here the challenge, how different types of data structured, or unstructured have to stored.
2. Data are to be in high quality.
3. As big data grows. Trending technologies are to be updated.
4. Unpredictable format would be challenge in analysing.

**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, Face book and Twitter to upgrade their business and to understand their people. Many companies are getting benefitted with the use of big data for their business growth, and for tuning their strategies. Large volume of data (Big data) can be analysed with various tools like sentiment and text analytics tools. The important challenge is designing sample for big data and prediction models with security. Because Data or information is power and Data is Oil. In future, unpredictable data format can be produced by the people and the 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.