**Introduction to big data platform**

Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.

In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Following are some the examples of Big Data-

* The **New York Stock Exchange** generates about ***one terabyte*** of new trade data per day.
* **Social Media**: The statistic shows that ***500+terabytes*** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.
* A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time. With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

## Types Of Big Data

1. **Structured**
2. **Unstructured**
3. **Semi-structured**

**Structured**

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

Eg. An 'Employee' table in a database is an example of Structured Data

**Unstructured**

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format. Eg. The output returned by 'Google Search'.

**Semi-structured**

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

Examples Of Semi-structured Data Personal data stored in an XML file.

**Characteristics or Traits of Big data**

**(i) Volume –** The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data.

**(ii) Variety –** The next aspect of Big Data is its **variety**.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

**(iii) Velocity –** The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

**(iv) Variability –** This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

**Benefits**

1. Ability to process Big Data brings in multiple benefits, such as-
   1. Businesses can utilize outside intelligence while taking decisions
2. Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.
   1. Improved customer service
3. Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.
   1. Early identification of risk to the product/services, if any
   2. Better operational efficiency
4. Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

|  |  |
| --- | --- |
| Reporting | Analysis |
| Reporting is “the process of organizing data into informational summaries in order to monitor how different areas of a business are performing.” | Analytics is “the process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.” |
| Reporting provides you with information | Analytics give you insights |
| Reporting raises questions | Analytics attempts to answer them. |
|  |  |

**Sampling and Resampling**

**Sampling** is a statistical procedure that is concerned with the selection of the individual observation; it helps us to make statistical inferences about the population.

**Random sampling:**

In data collection, every individual observation has equal probability to be selected into a sample. In random sampling, there should be no pattern when drawing a sample. Significance: [Significance](http://www.statisticssolutions.com/resources/directory-of-statistical-analyses/significance) is the percent of chance that a relationship may be found in sample data due to luck. Researchers often use the 0.05% significance level.

**Probability and non-probability sampling:**

Probability sampling is the sampling technique in which every individual unit of the population has greater than zero probability of getting selected into a sample.

Non-probability sampling is the sampling technique in which some elements of the population have no probability of getting selected into a sample.

**Types of random sampling:**

With the random sample, the types of random sampling are:

* **Simple random sampling:** By using the random number generator technique, the researcher **draws a sample from the population** called simple random sampling. Simple random samplings are of two types. One is when samples are drawn with replacements, and the second is when samples are drawn without replacements.
* **Equal probability systematic sampling:**In this type of sampling method, a researcher starts from a random point and selects **every nth subject** in the sampling frame. In this method, there is a danger of order bias.
* **Stratified simple random sampling:**In stratified simple random sampling, a proportion from **strata of the population** is selected using simple random sampling. For example, a fixed proportion is taken from every class from a school.
* **Multistage stratified random sampling:** In multistage stratified random sampling, a proportion of strata is selected from a **homogeneous group using simple** random sampling. For example, from the nth class and nth stream, a sample is drawn called the multistage stratified random sampling.
* **Cluster sampling:**Cluster sampling occurs when a random sample is drawn from certain **aggregational geographical groups**.

**Types of non-random sampling:** Non-random sampling is widely used in qualitative research.  Random sampling is too costly in qualitative research. The following are non-random sampling methods:

* **Availability sampling:** Availability sampling occurs when the researcher selects the sample based on the availability of a sample. This method is also called haphazard sampling. E-mail surveys are an example of availability sampling.
* **Quota sampling**: This method is similar to the availability sampling method, but with the constraint that the sample is drawn proportionally by strata.

Once generated, a sample can be used for predictive analytics. For example, a retail business might use data sampling to uncover patterns about customer behavior and predictive modeling to create more effective sales strategies.

**Resampling**

* Resampling is the method that consists of **drawing repeated samples from the original data** samples.
* The method of Resampling is a nonparametric method of statistical inference. In other words, the method of resampling does not involve the utilization of the generic distribution tables (for example, normal distribution tables) in order to compute approximate p probability values.
* Resampling involves the selection of randomized cases with replacement from the original data sample in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample.
* Due to replacement, the drawn number of samples that are used by the method of resampling consists of repetitive cases.
* Resampling generates a unique sampling distribution on the basis of the actual data.
* The method of resampling uses experimental methods, rather than analytical methods, to generate the unique sampling distribution. The method of resampling yields unbiased estimates as it is based on the unbiased samples of all the possible results of the data studied by the researcher.
* Resampling is also known as Bootstrapping or Monte Carlo Estimation.
* The method of bootstrapping, which is equivalent to the method of resampling, utilizes repeated samples from the original data sample in order to calculate the test statistic.
* Monte Carlo estimation, which is also equivalent to the bootstrapping method, is used by the researcher to obtain the resampling results.

**Regression Modelling**

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

There are multiple benefits of using regression analysis. They are as follows:

* It indicates the **significant relationships** between dependent variable and independent variable.
* It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured **on different scales, such as the effect of price** changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

**Types of regression Technique:**

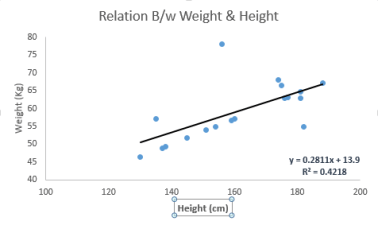
There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).

**Linear Regression**

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be [continuous or discrete](https://en.wikipedia.org/wiki/Continuous_and_discrete_variables), and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b\*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

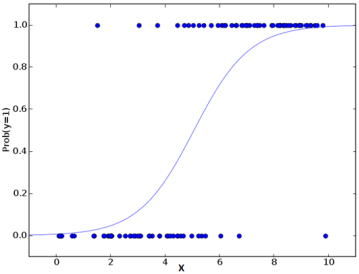
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Linear_Regression1.png)

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

**Logistic Regression**

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is [**logit**](https://en.wikipedia.org/wiki/Logistic_function) function.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Logistic_Regression.png)

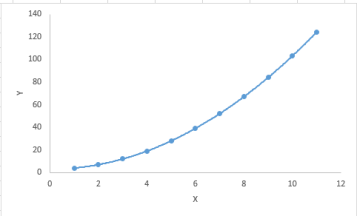
* It is widely used for **classification problems**
* Logistic regression **doesn’t require linear relationship** between dependent and independent variables.  It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
* To avoid **over fitting and under fitting, we should include all significant variables**. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
* It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
* The independent variables should not be correlated with each other i.e. **no multi collinearity**.  However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
* If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**
* If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

**Polynomial Regression**

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

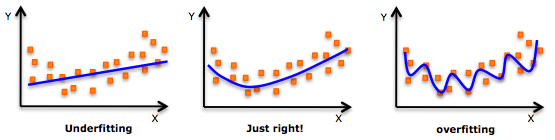
y=a+b\*x^2

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Polynomial.png)

#### Important Points:

* While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/underfitting-overfitting.png)

* Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation.

**Rule Induction**

1. Rule induction is a data science process of deducing if-then rules from a dataset.
2. These symbolic decision rules explain an inherent relationship between the attributes and class labels in a dataset.
3. Many real-life experiences are based on intuitive rule induction. For example, one could come up with a rule that states that “if it is 8:00 a.m. on a weekday, then highway traffic will be heavy” and “if it is 8:00 p.m. on a Sunday, then the traffic will be light.”
4. These rules are not necessarily right all the time. The 8:00 a.m. weekday traffic may be light during a holiday season.
5. But, in general, these rules hold true and are deduced from real-life experience based on our everyday observations.
6. The rule induction provides a **powerful classification approach** that can be easily understood by the general audience.
7. Apart from its use in data science by classification of unknown data, rule induction is also used to describe the patterns in the data.
8. The description is in the **form of simple if-then rules** that can be easily understood by general users.
9. The easiest way to extract rules from a dataset is **from a decision tree** that is developed on the same dataset.
10. A decision tree splits data on every node and leads to the leaf where the class is identified.
11. If one traces back from the leaf to the root node, they can combine all the split conditions to form a distinct rule.
12. For example, in the Golf dataset (Table 4.1), based on four weather conditions, a rule set can be generalized to determine when a player prefers to play golf or not. Fig. 4.19 shows the decision tree developed from the Golf data with five leaf nodes and two levels.
13. A rule can be extracted if one traces back the first leaf from the left: If Outlook is overcast, then Play 5 yes.

Similarly, rules can be extracted from all the five leaves:

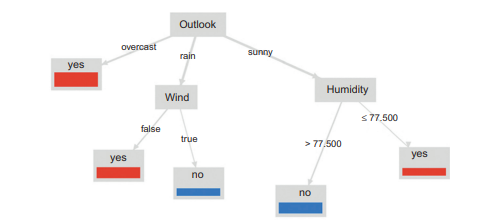
Rule 1: if (Outlook 5 overcast) then Play 5 yes

Rule 2: if (Outlook 5 rain) and (Wind 5 false) then Play 5 yes

Rule 3: if (Outlook 5 rain) and (Wind 5 true) then Play 5 no

Rule 4: if (Outlook 5 sunny) and (Humidity > 77.5) then Play 5 no

Rule 5: if (Outlook 5 sunny) and (Humidity <77.5) then Play 5 yes



**How knowledge data discovery is an aid to big data Analytics (Question paper)**

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.

Major KDD application areas include marketing, fraud detection, telecommunication and manufacturing.

Traditionally, data mining and knowledge discovery was performed manually. As time passed, the amount of data in many systems grew to larger than terabyte size, and could no longer be maintained manually. Moreover, for the successful existence of any business, discovering underlying patterns in data is considered essential. As a result, several software tools were developed to discover hidden data and make assumptions, which formed a part of artificial intelligence.

The KDD process has reached its peak in the last 10 years. It now houses many different approaches to discovery, which includes inductive learning, Bayesian statistics, semantic query optimization, knowledge acquisition for expert systems and information theory. The ultimate goal is to extract high-level knowledge from low-level data.

KDD includes multidisciplinary activities. This encompasses data storage and access, scaling algorithms to massive data sets and interpreting results. The data cleansing and data access process included in data warehousing facilitate the KDD process. Artificial intelligence also supports KDD by discovering empirical laws from experimentation and observations. The **patterns recognized** in the data must be valid on new data, and possess some degree of certainty. **These patterns** are considered new knowledge. Steps involved in the entire KDD process are:

1. Identify the goal of the KDD process from the customer’s perspective.
2. Understand application domains involved and the knowledge that's required
3. Select a target data set or subset of data samples on which discovery is be performed.
4. Cleanse and preprocess data by deciding strategies to handle missing fields and alter the data as per the requirements.
5. Simplify the data sets by removing unwanted variables. Then, analyze useful features that can be used to represent the data, depending on the goal or task.
6. Match KDD goals with data mining methods to suggest hidden patterns.
7. Choose data mining algorithms to discover hidden patterns. This process includes deciding which models and parameters might be appropriate for the overall KDD process.
8. Search for patterns of interest in a particular representational form, which include classification rules or trees, regression and clustering.
9. Interpret essential knowledge from the mined patterns.
10. Use the knowledge and incorporate it into another system for further action.
11. Document it and make reports for interested parties