**Name :- Ankit Senjaliya**

**Enrollment No. :- 19BT04046**

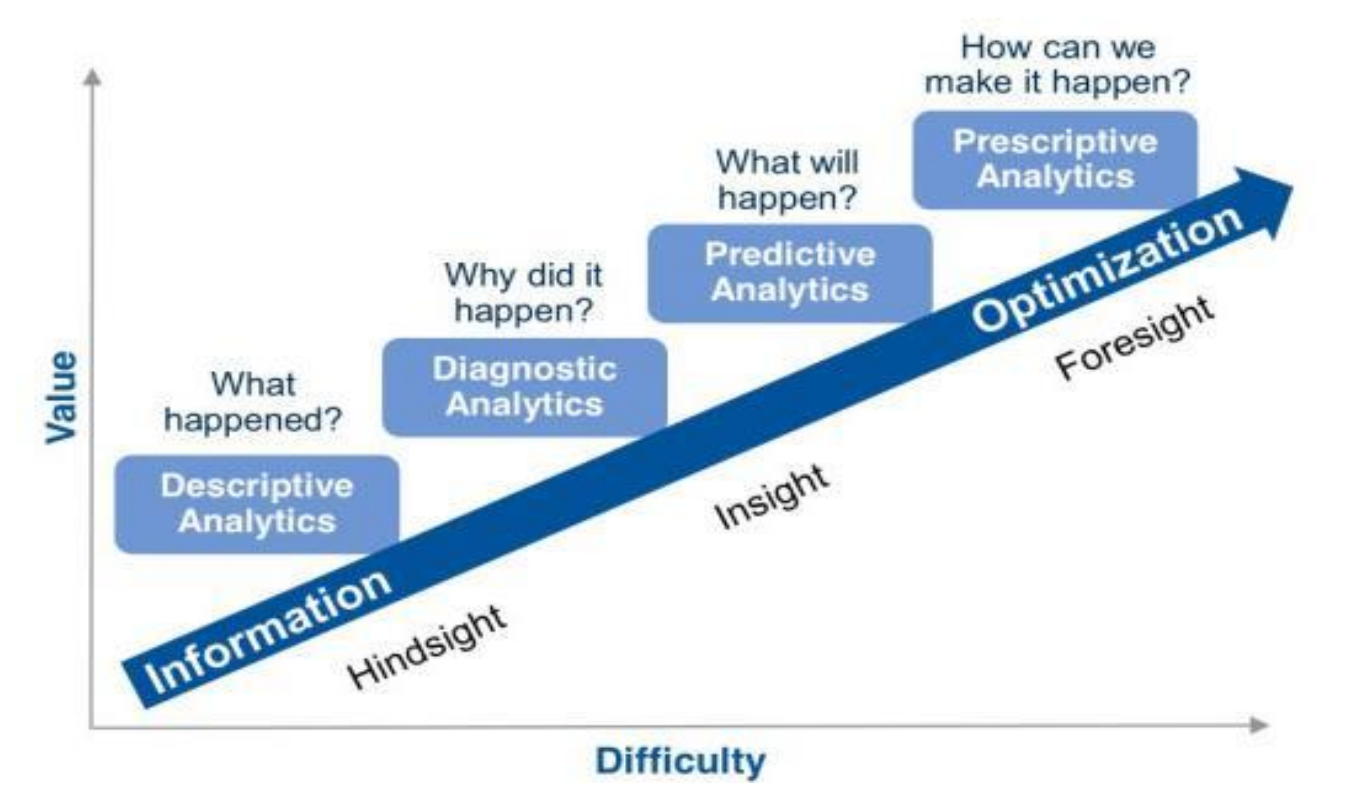
**Big Data Assignment – 1**

1. What is Business Analytics? What are the various types of Analytics?

* **Business Analytics :-**
* It is the use of data, information technology, statistical analysis, quantitative methods and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions.
* Let’s start by differentiating between data analytics and traditional analytics. The terms are often used interchangeably, but a distinction does exist. Traditional data analytics refers to the process of analyzing massive amounts of collected data to get insights and predictions. Business data analytics (sometimes called business analytics) takes that idea, but puts it in the context of business insight, often with prebuilt business content and tools that expedite the analysis process.



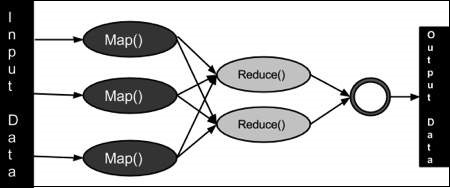
* Specifically, business analytics refers to:
* Taking in and processing historical business data
* Analyzing that data to identify trends, patterns, and root causes
* Making data-driven business decisions based on those insights
* In other words, data analytics is more of a general description of the modern analytics process. Business analytics implies a narrower focus and has functionally become more prevalent and more important for organizations around the globe as the overall volume of data has increased.
* **Various Types Of Analytics :-**



* **What is Descriptive Analytics?**
* Descriptive analytics answer the question, “What happened?” This type of analytics is by far the most commonly used by customers, providing reporting and analysis cantered on past events. It helps companies understand things such as:
* How much did we sell as a company?
* What was our overall productivity?
* How many customers churned in the last quarter?
* Descriptive analytics is used to understand the overall performance at an aggregate level and is by far the easiest place for a company to start as data tends to be readily available to build reports and applications.
* **What is Diagnostic Analytics?**
* Diagnostic analytics, just like descriptive analytics, uses historical data to answer a question. But instead of focusing on “the what”, diagnostic analytics addresses the critical question of why an occurrence or anomaly occurred within your data. Diagnostic analytics also happen to be the most overlooked and skipped step within the analytics maturity model. Anecdotally, I see most customers attempting to go from “what happened” to “what will happen” without ever taking the time to address the “why did it happen” step. This type of analytics helps companies answer questions such as:
* Why did our company sales decrease in the previous quarter?
* Why are we seeing an increase in customer churn?
* Why are a specific basket of products vastly outperforming their prior year sales figures?
* Diagnostic analytics tends to be more accessible and fit a wider range of use cases than machine learning/predictive analytics. You might even find that it solves some business problems you earmarked for predictive analytics use cases.
* **What is Predictive Analytics?**
* Predictive Analytics is a form of advanced analytics that determines what is likely to happen based on historical data using machine learning. Historical data that comprises the bulk of descriptive and diagnostic analytics is used as the basis of building predictive analytics models. Predictive analytics helps companies address use cases such as:
* Predicting maintenance issues and part breakdown in machines.
* Determining credit risk and identifying potential fraud.
* Predict and avoid customer churn by identifying signs of customer dissatisfaction.
* **What is Prescriptive Analytics?**
* Prescriptive analytics is the fourth and final pillar of modern analytics. Prescriptive analytics pertains to true guided analytics where your analytics is prescribing or guiding you toward a specific action to take. It is effectively the merging of descriptive and predictive analytics to drive decision making. Existing scenarios or conditions (think your current fleet of freight trains) and the ramifications of a decision or occurrence (parts breakdown on the freight trains) are applied to create a guided decision or action for the user to take (proactively buy more parts for preventative maintenance).
* Prescriptive analytics requires strong competencies in descriptive, diagnostic, and predictive analytics which is why it tends to be found in highly specialized industries (oil and gas, clinical healthcare, finance, and insurance to name a few) where use cases are well defined. Prescriptive analytics help to address use cases such as:
* Automatic adjustment of product pricing based on anticipated customer demand and external factors.
* Flagging selects employees for additional training based on incident reports in the field.
* Prescriptive analytics primary aim is to take the educated guess or assessment out of data analytics and streamline the decision-making process.

1. Define HDFS & MapReduce. Explain the process of MapReduce.

* **HDFS :-**
* HDFS is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes. HDFS is one of the major components of Apache Hadoop, the others being MapReduce and YARN.
* HDFS should not be confused with or replaced by Apache HBase, which is a column-oriented non-relational database management system that sits on top of HDFS and can better support real-time data needs with its in-memory processing engine.
* **MapReduce :-**
* MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.
* The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into *mappers* and *reducers* is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.
* **Explain the process of MapReduce.**
* Generally MapReduce paradigm is based on sending the computer to where the data resides!
* MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
* **Map Stage** :- The map or mapper’s job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
* **Reduce Stage** :-This stage is the combination of the Shufflestage and the Reduce stage. The Reducer’s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.



* During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
* The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
* Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
* After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.