Data Analytics
UNIT 1

What is data, and why is it important?

Data – a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process

Data is of much importance now a days

1) Data helps to make better decisions

- Finding new customers
- Increasing customer retention
- Improving customer service
- Better managing marketing efforts
- Tracking social media interaction
- Predicting sales trends

2) Data helps you solve problems

Ex After experiencing a slow sales month or watching a poor-performing marketing campaign, how do you pinpoint what went wrong? Tracking and reviewing data from business processes helps you uncover performance breakdowns so you can better understand each part of the process and know which steps need to be fixed and which are performing well.

3) Data helps you understand performance

One example: Say you have a top-performing sales rep who you send most leads to. However, when you delve into the data it shows that she closes deals at a lower rate than one of your other sales reps who receives fewer leads but closes deals at a higher percentage. (**In fact, here's how you can easily track sales rep performance**.) Certainly some performance data that can affect how you portion out leads--which can lead to revenue increase. Performance data provides the clarity needed for better results.

4) Data helps you improve processes

Data helps you understand and improve business processes so you can reduce wasted money and time. Every company feels the effects of waste. It depletes resources, squanders time, and ultimately impacts the bottom line.

For example, bad advertising decisions can be one of the greatest wastes of resources in a company. With data showing how different marketing channels are performing, however, you can see which ones offer the greatest ROI and focus on those. Or you could dig into why other channels are not performing as well and work to improve their performance. This would allow you budget to generate more leads without having to increase the advertising spend.

5) Data helps you understand consumers

TYPES OF DATA

1. **Structured data –**
   Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. Example: Relational data.
2. **Semi-Structured data –**
   Semi-structured data is information that does not reside in a relational database but that have some organizational properties that make it easier to analyze. With some process, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. Example: XML data.
3. **Unstructured data –**
   Unstructured data is a data that is which is not organized in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs.

U N S T R U C T U R E D D A T A I S C R I T I C A L

Notable fact: almost all information we used to operate with is unstructured: emails, articles, or business-related data like customer interactions. Unstructured data can be extremely different: extracted from a human language with NLP (Natural Language Processing), gained thru various sensors, scrapped from the Internet, acquired from NoSQL databases, etc.

As the majority of information we can access is unstructured, the benefits of unstructured data analysis are obvious. It can bring many useful insights and ideas on how to improve the performance of the company or a specific service.

If we want a machine to process the data, so the first step is to make it “understandable” for computers. We should build a bridge between human understanding and computer processing. It means that most often human operator processes required data manually and translates it to the format suitable for machine processing.

One of the main problems with qualitative data analysis, however, is that standard databases like Excel or SQL require a certain structure. Unfortunately, unstructured data lacks this structure and traditional ORM (object-relational mapping) software can’t process it properly to fill the database.

But it doesn’t mean that we should forget about this kind of information and lose valuable insights.

When you sift the unstructured data, you get details that allow seeing the full picture of what's going on.

The information you receive after the analysis can become the cornerstone of a successful business strategy since it usually contains essential nuances about customer behavior or current trends. Let's take a look at an example.

Sources of data

Personal data

Personal data is anything that is specific to you. It covers your demographics, your location, your email address and other identifying factors.

Lots of different companies collect your personal data (especially social media sites), anytime you have to put in your email address or credit card details you are giving away your personal data. Often they'll use that data to provide you with personalized suggestions to keep you engaged. Facebook for example uses your personal information to suggest content you might like to see based on what other people similar to you like.

In addition, personal data is aggregated (to depersonalize it somewhat) and then sold to other companies, mostly for advertising and competitive research purposes. That's one of the ways you get targeted ads and content from companies you've never even heard of.

Transactional data

Transactional data is anything that requires an action to collect. You might click on an ad, make a purchase, visit a certain web page, etc.

Pretty much every website you visit collects transactional data of some kind, either through Google Analytics, another 3rd party system or their own internal data capture system.

Transactional data is incredibly important for businesses because it helps them to expose variability and optimize their operations for the highest quality results. By examining large amounts of data, it is possible to uncover hidden patterns and correlations. These patterns can create competitive advantages, and result in business benefits like more effective marketing and increased revenue.

Web data

Web data is a collective term which refers to any type of data you might pull from the internet, whether to study for research purposes or otherwise. That might be data on what your competitors are selling, published government data, football scores, etc. It's a catchall for anything you can find on the web that is public facing (ie not stored in some internal database). Studying this data can be very informative, especially when communicated well to management.

Web data is important because it's one of the major ways businesses can access information that isn't generated by themselves. When creating quality business models and making important BI decisions, businesses need information on what is happening internally and externally within their organization and what is happening in the wider market.

Web data can be used to monitor competitors, track potential customers, keep track of channel partners, generate ads, build apps, and much more. It's uses are still being discovered as the technology for turning unstructured data into structured data improves.

Web data can be collected by writing web scrapers to collect it, using a scraping tool, or by paying a third party to do the scraping for you. A web scraper is a computer program that takes a URL as an input and pulls the data out in a structured format – usually a JSON feed or CSV.

Sensor data

Sensor data is produced by objects and is often referred to as the Internet of Things. It covers everything from your smartwatch measuring your heart rate to a building with external sensors that measure the weather.

]By measuring what is happening around them, machines can make smart changes to increase productivity and alert people when they are in need of maintenance.

Uses of data

1) Data in business

Data can help businesses better understand their customers, improve their advertising campaigns, personalize their content and improve their bottom lines. The advantages of data are many, but you can't access these benefits without the proper data analytics tools and processes. While raw data has a lot of potential, you need data analytics to unlock the power to grow your business.

2) Data in healthcare

Data is extremely useful in the field of medical and healthcareDoctors are able to keep track about patient's history, the link to which is only accessed by the patient and his particular physician.

The data of the patient is stored in the database safe and secure and as and when required by the doctor it can be viewed. Most of the medical devices are big data oriented. The use of data has gone to such an extent that doctors can check the person through the heart and temperature monitoring watch fitted on the patient's hand and prescribe him with the following medicines

3) Data in media and entertainment

New business models are launched for different companies in the media and entertainment industry. The business model runs on collecting or creating the content, further to analyze it, then marketing and distribution of the content. As the rate of consumer's search increases, there is a need for obtaining content at any moment, in whatever place in all formats on a variety of devices. We can runs through customer's data along with observable data and gather even minute information to create a customer's detailed profile. The benefits of big data in media and entertainment industry include forecasting what the target audience wants, planning optimization, expanding acquisition and retention; suggest content on demand and new

4) Data in Transportation

In latest times, enormous amounts of data from location-oriented social networks and high speed data from telecoms have influenced travel journeys.

5) Data in Banking

Banking is a very crucial sector. The security, privacy, management and maintenance of any banking system is a challenge. Data here is very beneficial and helps in the fraud detection in banking system. Using Big Data , we can searches all the illegal activities that has taken place and can identifies the misuse of credit and debit cards, business precision, customer statistics modification, public analytics for business.

INTRODUCTION TO BIG DATA

"90% of the world's data was generated in the last fewyears."

Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in **2011**, and in every ten minutes in **2013**. This rate is still growing enormously. Though all this information produced is meaningful and can be useful when processed, it is being neglected.

Big data is a collection of large datasets that cannot be processed using traditional computing techniques. It is not a single technique or a tool, rather it has become a complete subject, which

What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

- **Black Box Data** − It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data** − Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** − The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data** − The power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data** − Transport data includes model, capacity, distance and availability of a vehicle.
- **Search Engine Data** − Search engines retrieve lots of data from different databases.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

- **Structured data** − Relational data.
- **Semi Structured data** − XML data.
- **Unstructured data** − Word, PDF, Text, Media Logs.

Benefits of Big Data

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

Characteristics of Big Data

Velocity

Velocity refers to the speed at which the data is generated, collected and analyzed. Data continuously flows through multiple channels such as computer systems, networks, social media, mobile phones etc. In today's data-driven business environment, the pace at which data grows can be best described as 'torrential' and 'unprecedented'. Now, this data should also be captured as close to real-time as possible, making the right data available at the right time. The speed at which data can be accessed has a direct impact on making timely and accurate business decisions. Even a limited amount of data that is available in real-time yields better business results than a large volume of data that needs a long time to capture and analyze.

Several Big data technologies today allow us to capture and analyze data as it is being generated in real-time.

Volume

Big data volume defines the 'amount' of data that is produced. The value of data is also dependent on the size of the data.

Today data is generated from various sources in different formats – structured and unstructured. Some of these data formats include word and excel documents, PDFs and reports along with media content such as images and videos. Due to the data explosion caused to digital and social media, data is rapidly being produced in such large chunks, it has become challenging for enterprises to store and process it using conventional methods of business intelligence and analytics. Enterprises must implement modern business intelligence tools to effectively capture, store and process such unprecendented amount of data in real-time.

Value

Although data is being produced in large volumes today, just collecting it is of no use. Instead, data fromwhich business insights are garnered add 'value' to the company. In the context of big data, value amounts to how worthy the data is of positively impacting a company's business. This is where big data analytics come into the picture. While many companies have invested in establishing data aggregation and storage infrastructure in their organizations, they fail to understand that the aggregation of data doesn't equal value addition. What you do with the collected data is what matters. With the help of advanced data analytics, useful insights can be derived from the collected data. These insights, in turn, are what add value to the decision-making process.

One way to ensure that the value of big data is considerable and worth investing time and effort into, is by conducting a cost Vs benefit analysis. By calculating the total cost of processing big data and comparing it with the ROI that the business insights are expected to generate, companies can effectively decide whether or not big data analytics will actually add any value to their business.

Variety

While the volume and velocity of data are important factors that add value to a business, big data also entails processing diverse data types collected from varied data sources. Data sources may involve external sources as well as internal business units. Generally, big data is classified as structured, semi-structured and unstructured data. While structured data is one whose format, length and volume are clearly defined, semi- structured data is one that may partially conform to a specific data format. On the other hand, unstructured data is unorganized data and doesn't conform with the traditional data formats. Data generated via digital and social media (images, videos, tweets, etc.) can be classified as unstructured data,

The sheer volume of data that organizations usually collect and generate may look chaotic and unstructured. In fact, almost 80 percent of data produced globally including photos, videos, mobile data, and social media content, is unstructured in nature.

Veracity/Validity

The Veracity of big data or Validity, as it is more commonly known, is the assurance of quality or credibility of the collected data. Can you trust the data that you have collected? Is this data credible enough to glean insights from? Should we be basing our business decisions on the insights garnered from this data? All these questions and more, are answered when the veracity of the data is known.

Since big data is vast and involves so many data sources, there is the possibility that not all collected datawill be of good quality or accurate in nature. Hence, when processing big data sets, it is important that the validity of the data is checked before proceeding for processing.

Types of Big Data Technologies:

Big Data Technology is mainly classified into two types:

1. **Operational Big Data Technologies**
2. **Analytical Big Data Technologies**

Firstly, The Operational Big Data is all about the normal day to day data that we generate. This could be the **Online Transactions, Social Media,** or the data from a **Particular Organisation** etc. You can even consider this to be a kind of Raw Data which is used to feed the **Analytical Big Data Technologies.**

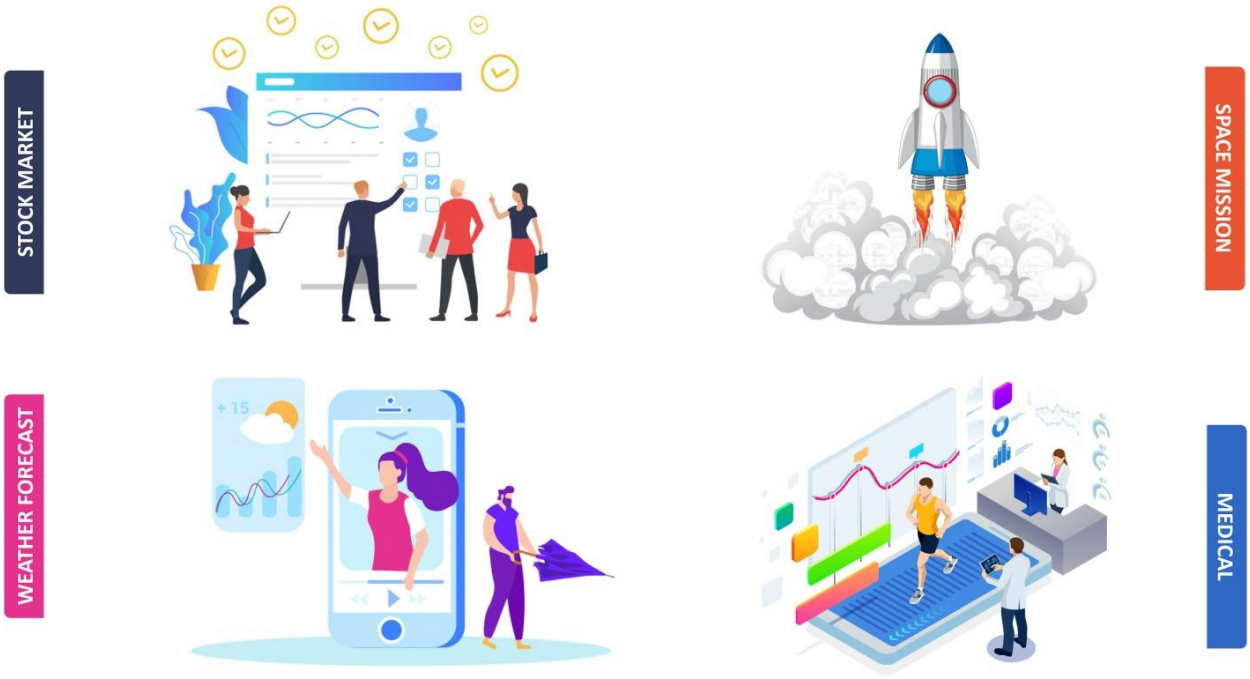
A few examples of **Operational Big Data Technologies** are as follows:

- Online ticket bookings, which includes your Rail tickets, Flight tickets, movie tickets etc.
- Online shopping which is your Amazon, Flipkart, Walmart, Snap deal and many more.
- Data from social media sites like Facebook, Instagram, what's app and a lot more.
- The employee details of any Multinational Company.

So, with this let us move into the **Analytical Big Data Technologies.**

Analytical Big Data is like the advanced version of Big Data Technologies. It is a little complex than the Operational Big Data. In short, Analytical big data is where the actual performance part comes into the picture and the crucial real-time business decisions are made by analyzing the Operational Big Data.

Few examples of **Analytical Big Data Technologies** are as follows:



- Stock marketing
- Carrying out the Space missions where every single bit of information is crucial.
- Weather forecast information.
- Medical fields where a particular patients health status can be monitored.

EVOLUTION OF DATA ANALYTICS

Introduction

The amount of data organizations process continues to increase. More data cross the internet every second than were stored in the entire internet just 20 years ago. This gives companies an opportunity to work with many petabyes of data in a single data set—and not just from the internet. For instance, it is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions. A petabyte is one quadrillion bytes, or the equivalent of about 20 million filing cabinets‘ worth of text. An exabyte is 1,000 times that amount, or one billion gigabytes. The old methods for handling such huge data won‘t work anymore.

Hence we need technologies that might handle and tame the big data. Few of such kind are as show in figure 1, but there are many more:

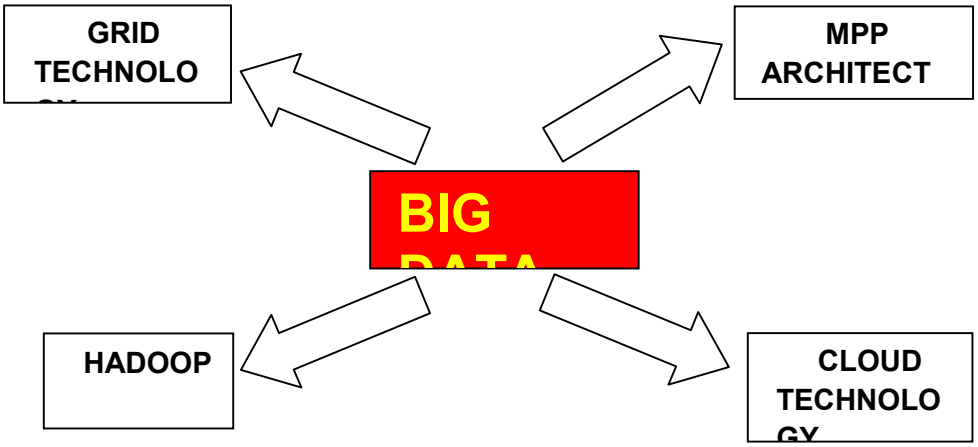


Figure 1. Big data technologies

We shall discuss about the above technologies in detail as we proceed further in this module. But before that the readers must be aware of how analytics and data environment converged together for making big data analytics more effective.

Learning Objectives

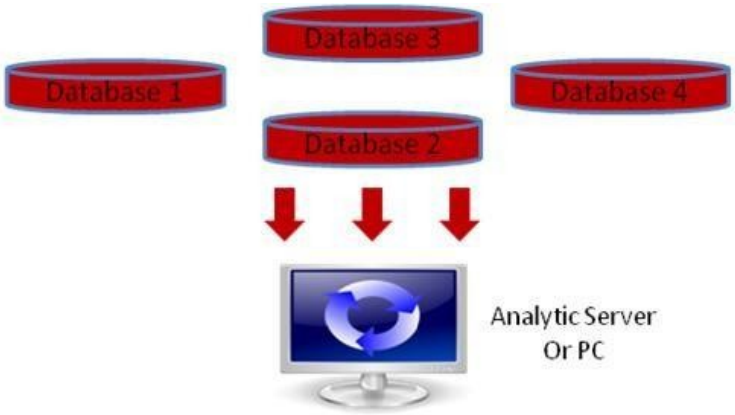
- To understand the evolution of analytical scalability
- To know about technologies that can handle big data

The Convergence of the Analytic and Data Environment

Traditional Analytic Architecture

Traditional analytics collects data from heterogeneous data sources and we had to pull all data together into a separate analytics environment to do analysis which can be an analytical server or a personal computer with more computing capability. The heavy processing occurs in the analytic environment as shown in figure 2.

Figure 2. Traditional Analytic Architecture



In such environments, shipping of data becomes a must, which might result in issues related with security of data and its confidentiality.

Modern In-Database Architecture

Data from heterogeneous sources are collected, transformed and loaded into data warehouse for final analysis by decision makers. The processing stays in the database where the data has been consolidated. The data is presented in aggregated form for querying. Queries from users are submitted to OLAP (online analytical processing) engines for execution. Such in-database architectures are tested for their query throughput rather than transaction throughput as in traditional database environments. More of metadata is required for directing the queries which helps in reducing the time taken for answering queries and hence increase the query throughput. Moreover the

data in consolidated form are free from anomalies, since they are preprocessed before loading into warehouses which may be used directly for analysis. The modern in- database architecture is shown in figure 3.

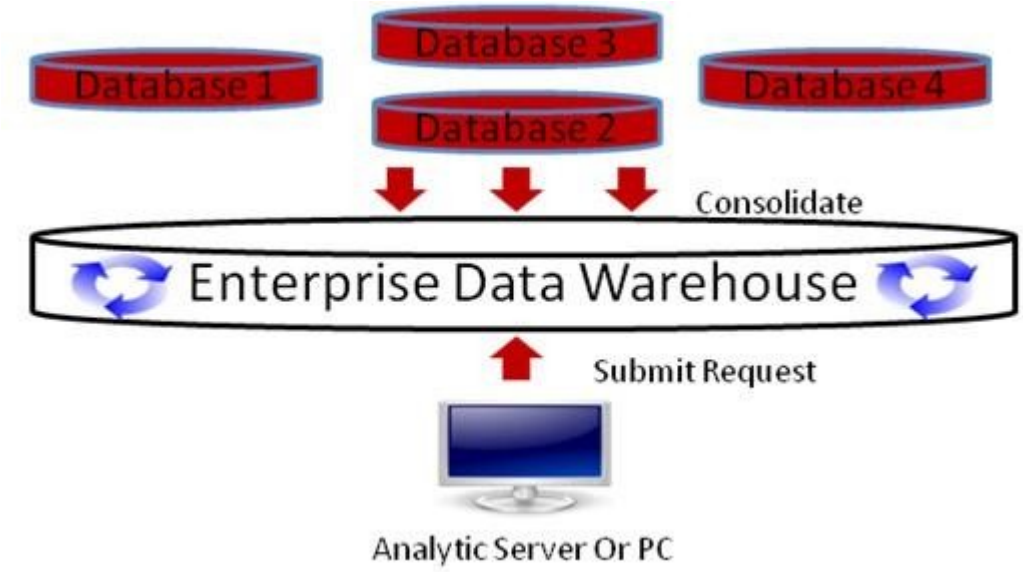


Figure 3. In-Database Architecture

Massively Parallel Processing (MPP)

Massive Parallel Processing (MPP) is the ―shared nothing‖ approach of parallel computing. It is a type of computing wherein the process is being done by many CPUs working in parallel to execute a single program.

One of the most significant differences between a Symmetric Multi-Processing or SMP and Massive Parallel Processing is that with MPP, each of the many CPUs has its own memory to assist it in preventing a possible hold up that the user may experience with using SMP when all of the CPUs attempt to access the memory simultaneously. The MPP architecture is given in figure 4.

The salient feature of MPP systems are:

- Loosely coupled nodes
- Nodes linked together by a high speed connection
- Each node has its own memory

- Disks are not shared, each being attached to only one node – shared nothing architectures

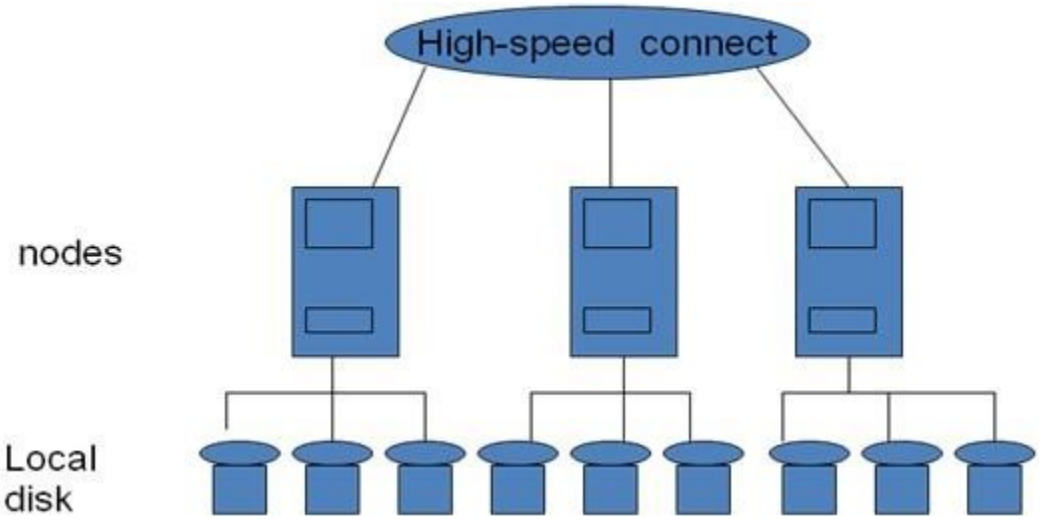


Figure 4. MPP Architecture

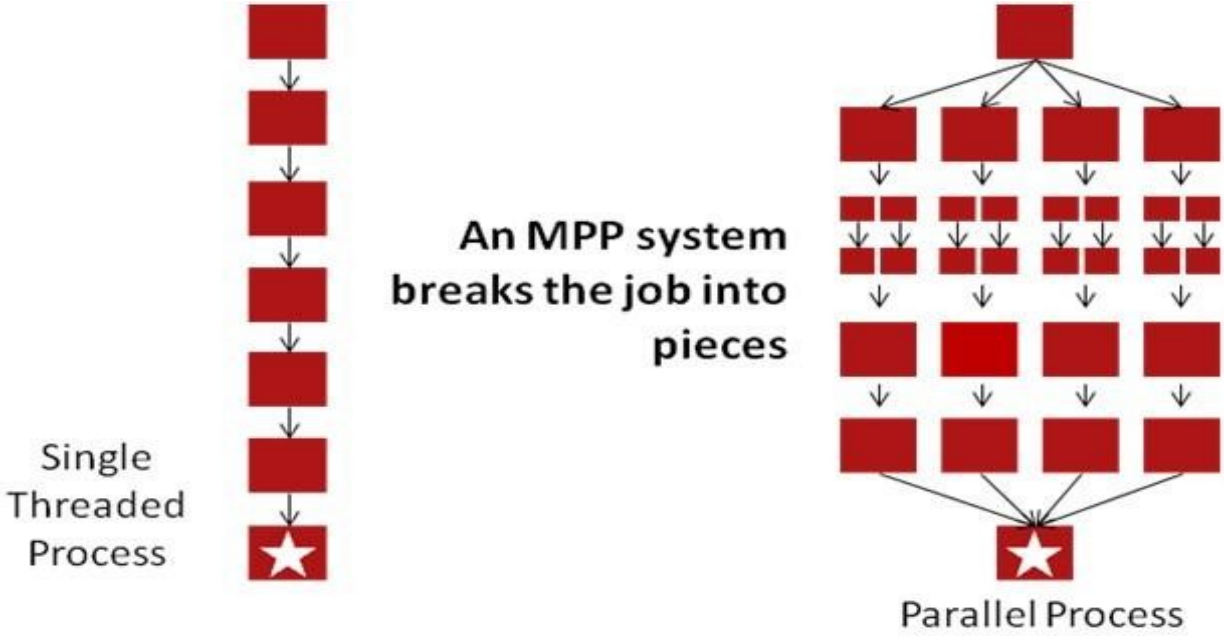


Figure 5. Execution of tasks in MPP

The idea behind MPP is really just that of the general parallel computing (figure 5) wherein the simultaneous execution of some combination of multiple instances of programmed instructions and data on multiple processors in such a way that the result can be obtained more effectively.

MPP uses shared distributed lock manager to maintain the integrity of the distributed resources across the system. CPU power that can be made available in a MPP is dependent upon number of nodes that can be connected. MPP systems build in redundancy to make recovery easy. MPP systems have resource management tools to manage the CPU and disk space and Query optimization.

The Cloud Computing

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.

McKinsey and Company in the paper _Clearing the Air on Cloud Computing‘ (March 2009) has indicated the following as characteristic features of cloud:

1) Mask the underlying infrastructure from the user
2) Be elastic to scale on demand
3) On a pay-per-use basis
4) National Institute of Standards and Technology (NIST)
5) On-demand self-service
6) Broad network access
7) Resource pooling
8) Rapid elasticity
9) Measured service

There are two types of cloud environment:

1. Public Cloud
   - The services and infrastructure are provided off-site over the internet
   - Greatest level of efficiency in shared resources
   - Less secured and more vulnerable than private clouds
2. Private Cloud
   - Infrastructure operated solely for a single organization
   - The same features of a public cloud

- Offer the greatest level of security and control
- Necessary to purchase and own the entire cloud infrastructure

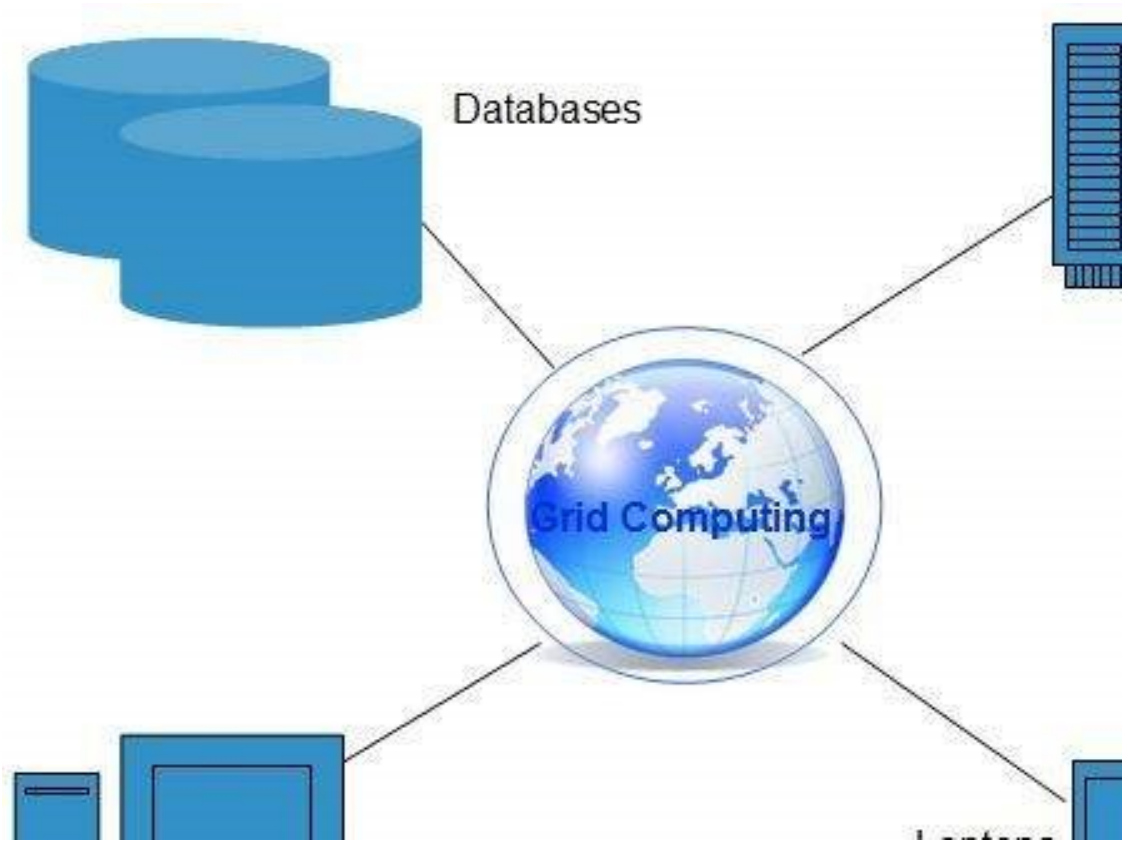
Grid Computing

Grid computing (also called "distributed computing") is a collection of computers working together to perform various tasks. It distributes the workload across multiple systems, allowing computers to contribute their individual resources to a common goal.

A computing grid is similar to a cluster, but each system (or node) on a grid has its own resource manager. In a cluster, the resources are centrally managed, typically by a single system. Additionally, clusters are usually located in a single physical space (such as a LAN), whereas grid computing often incorporates systems in several different locations (such as a WAN).

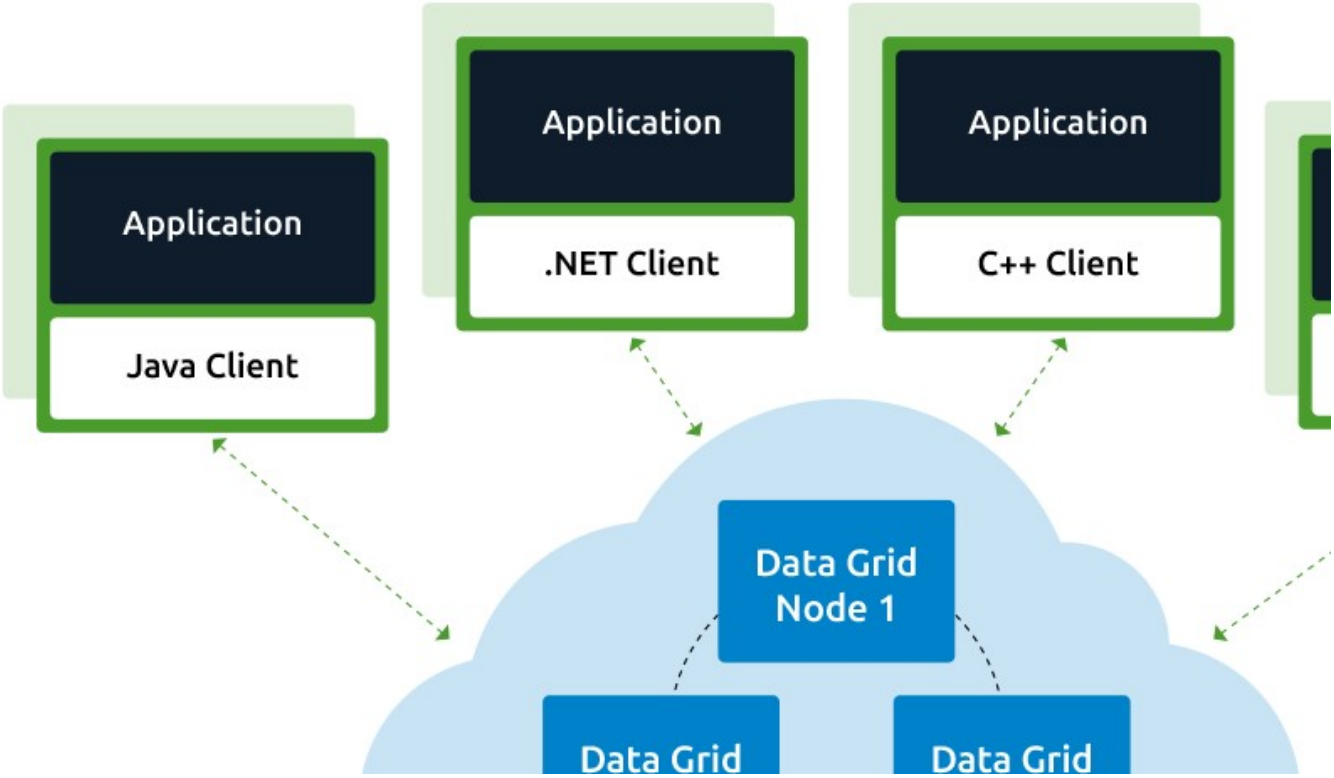
In order for systems in a computing grid to work together, they must be physically connected (over a network or the Internet) and run software that allows them to communicate. The software used in grid computing is called middleware since it translates the information passed from one system to another into a recognizable format. This allows the data computed by one node within the grid to be stored or processed by another system on the grid.

Grid computing has many different scientific applications. For example, it is used to model the changes in molecular structures, analyze brain behavior, and compute complex physics models. It is also used to perform weather and economic simulations. Some companies also use grid computing to process internal data and provide services over the Internet. Cloud computing, for instance, is considered to be a subset of grid computing.



How Does Grid Computing Work?

Grid computing works by running specialized software on every computer that participates in the data grid. The software acts as the manager of the entire system and coordinates various tasks across the grid. Specifically, the software assigns subtasks to each computer so they can work simultaneously on their respective subtasks. After the completion of subtasks, the outputs are gathered and aggregated to complete a larger-scale task. The software lets each computer communicate over the network with the other computers so they can share information on what portion of the subtasks each computer is running, and how to consolidate and deliver outputs.



Difference between Grid Computing and Cloud Computing

1. Technology involved in Grid Computing and Cloud Computing

– Grid computing is form of computing which follows a distributed architecture which means a single task is broken down into several smaller tasks through a

distributed system involving multiple computer networks. Cloud computing, on the other hand, is a whole new class of computing based on network technology where every user of the cloud has its own private resource that is provided by the specific service provider.

2. Terminology of Grid Computing and Cloud Computing

– Both are network based computing technologies that share similar characteristics such as resource pooling, however, they are very different from each other in terms of architecture, business model, interoperability, etc. Grid computing is a collection of computer resources from multiple locations to process a single task. The grid acts as a distributed system for collaborative sharing of resources. Cloud computing, on the other hand, is a form of computing based on virtualized resources which are located over multiple locations in clusters.

3. Computing Resources in Grid Computing and Cloud Computing

– Grid computing is based on a distributed system which means computing resources are distributed among different computing units which are located across different sites, countries, and continents. In cloud computing, computing resources are managed centrally which are located over multiple servers in clusters in cloud providers' private data centers.

4. Research Community

– In grid computing, computing resources are provided as a utility with grids as a computing platform that are distributed geographically and are grouped in virtual organization with multiple user communities to solve large-scale problems over the internet. Grid involves more resources than just computers and networks. Cloud computing, on the other hand, involves a common group of system administrators that manage the entire domain.

5. Function of of Grid Computing and Cloud Computing

– The main function of grid computing is job scheduling using all kinds of computing resources where a task is divided into several independent sub-tasks and each machine on a grid is assigned with a task. After all the sub-tasks are completed they are sent back to the main machine which handles and processes all the tasks. Cloud computing involves resource pooling through grouping resources on an as-needed basis from clusters of servers.

6. Application of Grid Computing and Cloud Computing

– The term “cloud” refers to the internet in cloud computing and as a whole it means internet-based computing. The cloud manages data, security requirements, job queues, etc. by eliminating the needs and complexity of buying hardware and software needed to build applications which are to be delivered as a service over the cloud. Grid computing is mostly used by academic research and is able to handle large sets of limited duration jobs that involve huge volumes of data

GRID COMPUTING
VERSUS
CLOUD COMPUTING

GRID COMPUTING	CLOUD COMPUTING
It is for Application Oriented.	It is for Service C
The resources are distributed among different computing units for processing a single task.	The computing res managed centrally ar over multiple servers
Grids are generally owned and managed by an organization within its premises.	The cloud servers ar Infrastructure provi placed in physically locations
It operates within a corporate network.	It can also be accessed internet.
It provides a shared pool of computing resources on an as-needed basis.	It involves dealing wi problem using varyin computing reso

Hadoop

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. Two main building blocks inside this runtime environment are MapReduce and Hadoop Distributed File System (HDFS). Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

Typically the compute nodes and the storage nodes are the same, that is, the MapReduce framework and the HDFS are running on the same set of nodes. This configuration allows the framework to effectively schedule tasks on the nodes where data is already present, resulting in very high aggregate bandwidth across the cluster.

The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master.

Minimally, applications specify the input/output locations and supply map and reduce functions via implementations of appropriate interfaces and/or abstract-classes. These, and other job parameters, comprise the job configuration. The Hadoop job client then submits the job (jar/executable etc.) and configuration to the JobTracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client.

Working of MapReduce

Let‘s assume there are 20 terabytes of data and 20 MapReduce server nodes for a project. The steps are summarized as follows as diagrammatically represented as in figure 5:

a) Distribute a terabyte to each of the 20 nodes using a simple file copy process

b) Submit two programs(Map, Reduce) to the scheduler

c) The map program finds the data on disk and executes the logic it contains

d) The results of the map step are then passed to the reduce process to summarize
and aggregate the final answers

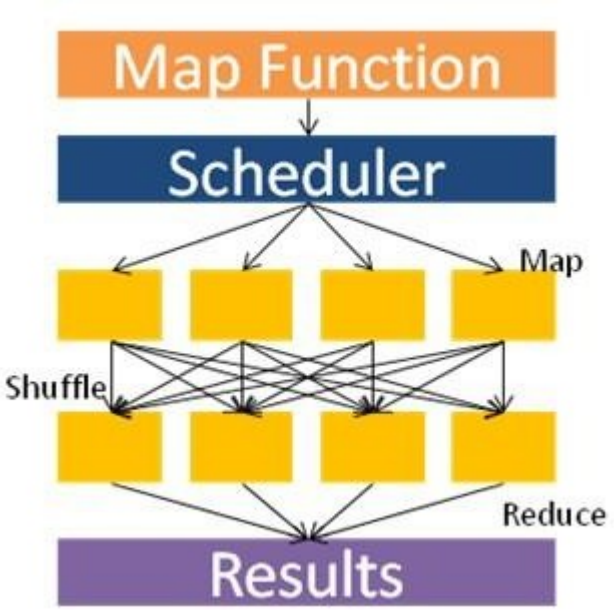


Figure 5. Working of MapReduce In

general MapReduce is Good for the following:

a) Lots of input, intermediate, and output data

b) Batch oriented datasets (ETL: Extract, Load, Transform)

c) Cheap to get up and running because of running on commodity

hardware However, MapReduce is not suitable for the following:

a) Fast response time

b) Large amounts of shared data

c) CPU intensive operations (as opposed to data intensive)

d) NOT a database

   i. No built-in security

   ii. No indexing, No query or process optimizer

   iii. No knowledge of other data that exists

Finally in real world we find few technologies that can integrate and work together for aiding big data analytics:

1) Databases running in the cloud, e.g., SimpleDB, Microsoft SQL Azure

2) Databases including MapReduce functionality, e.g., Teradata Aster, Oracle

3) MapReduce can be run against data sourced from a database, e.g., CouchDB, mongoDB

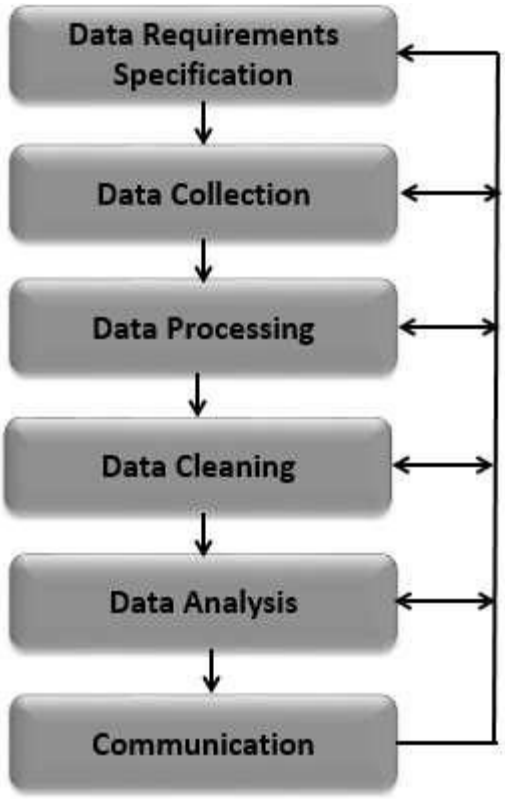
4) MapReduce can also run against data in the cloud, e.g., cloudmapreduce

Data Analytics Process

Data Analysis is a process of collecting, transforming, cleaning, and modeling data with the goal of discovering the required information. The results so obtained are communicated, suggesting conclusions, and supporting decision-making. Data visualization is at times used to portray the data for the ease of discovering the useful patterns in the data. The terms Data Modeling and Data Analysis mean the same.

Data Analysis Process consists of the following phases that are iterative in nature −

- Data Requirements Specification
- Data Collection
- Data Processing
- Data Cleaning
- Data Analysis
- Communication



Data Requirements Specification

The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical.

Data Collection

Data Collection is the process of gathering information on targeted variables identified as data requirements. The emphasis is on ensuring accurate and honest collection of data. Data Collection ensures that data gathered is accurate such that the related decisions are valid. Data Collection provides both a baseline to measure and a target to improve.

Data is collected from various sources ranging from organizational databases to the information in web pages. The data thus obtained, may not be structured and may contain irrelevant information. Hence, the collected data is required to be subjected to Data Processing and Data Cleaning.

Data Processing

The data that is collected must be processed or organized for analysis. This includes structuring the data as required for the relevant Analysis Tools. For example, the data might have to be placed into rows and columns in a table within a Spreadsheet or Statistical Application. A Data Model might have to be created.

Data Cleaning

The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors. There are several types of Data Cleaning that depend on the type of data. For example, while cleaning the financial data, certain totals might be compared against reliable published numbers or defined thresholds. Likewise, quantitative data methods can be used for outlier detection that would be subsequently excluded in analysis.

Data Analysis

Data that is processed, organized and cleaned would be ready for the analysis. Various data analysis techniques are available to understand, interpret, and derive conclusions based on the requirements. Data Visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

Statistical Data Models such as Correlation, Regression Analysis can be used to identify the relations among the data variables. These models that are descriptive of the data are helpful in simplifying analysis and communicate results.

The process might require additional Data Cleaning or additional Data Collection, and hence these activities are iterative in nature.

Communication

The results of the data analysis are to be reported in a format as required by the users to support their decisions and further action. The feedback from the users might result in additional analysis.

The data analysts can choose data visualization techniques, such as tables and charts, which help in communicating the message clearly and efficiently to the users. The analysis tools provide facility to highlight the required information with color codes and formatting in tables and charts.

Modern Analytical tools

The growing demand and importance of data analytics in the market have generated many openings worldwide. It becomes slightly tough to shortlist the top data analytics tools as the open source tools are more popular, user-friendly and performance oriented than the paid

version. There are many open source tools which doesn't require much/any coding and manages to deliver better results than paid versions e.g. – R programming in data mining and Tableau public, Python in data visualization. Below is the list of top 10 of data analytics tools, both open source and paid version, based on their popularity, learning and performance.

1. R Programming

R is the leading analytics tool in the industry and widely used for statistics and data modeling. It can easily manipulate your data and present in different ways. It has exceeded SAS in many ways like capacity of data, performance and outcome. R compiles and runs on a wide variety of platforms viz -UNIX, Windows and MacOS. It has 11,556 packages and allows you to browse the packages by categories. R also provides tools to automatically install all packages as per user requirement, which can also be well assembled with Big data.

2. Tableau Public:

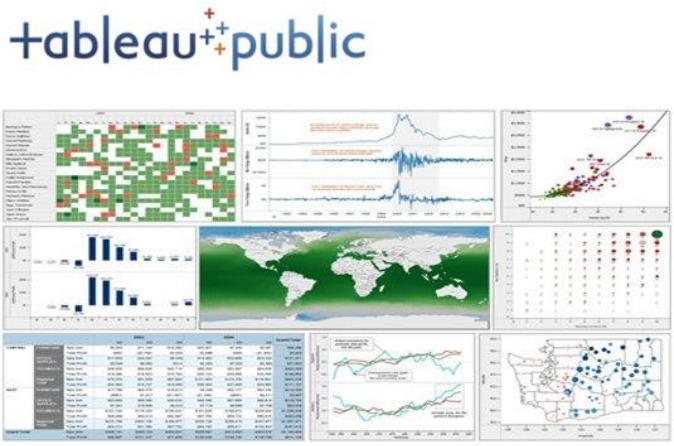
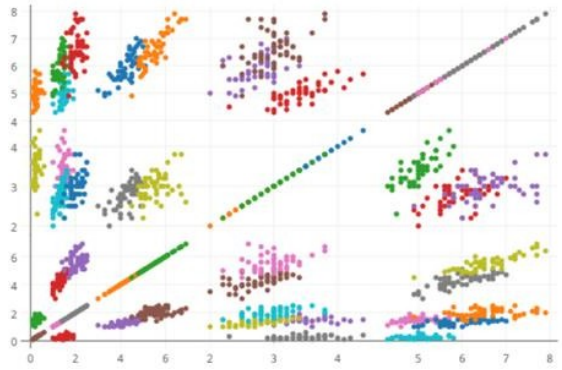


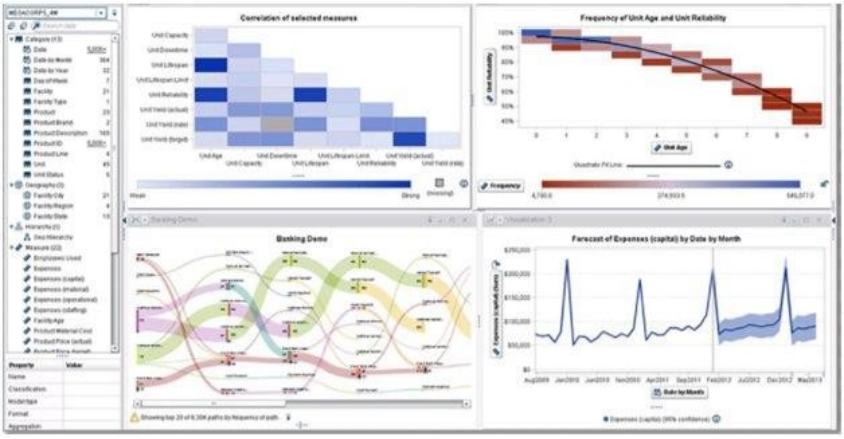
Tableau Public is a free software that connects any data source be it corporate Data Warehouse, Microsoft Excel or web-based data, and creates data visualizations, maps, dashboards etc. with real-time updates presenting on web. They can also be shared through social media or with the client. It allows the access to download the file in different formats. If you want to see the power of tableau, then we must have very good data source. Tableau's Big Data capabilities makes them important and one can analyze and visualize data better than any other data visualization software in the market.

3.Python



Python is an object-oriented scripting language which is easy to read, write, maintain and is a free open source tool. It was developed by Guido van Rossum in late 1980's which supports both functional and structured programming methods.
Phython is easy to learn as it is very similar to JavaScript, Ruby, and PHP. Also, Python has very good machine learning libraries viz. Scikitlearn, Theano, Tensorflow and Keras. Another important feature of Python is that it can be assembled on any platform like SQL server, a MongoDB database or JSON. Python can also handle text data very well.

4. SAS:

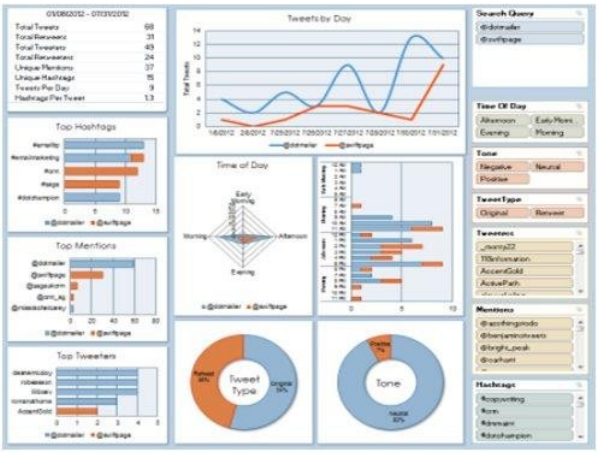


Sas is a programming environment and language for data manipulation and a leader in analytics, developed by the SAS Institute in 1966 and further developed in 1980's and 1990's. SAS is easily accessible, managable and can analyze data from any sources. SAS introduced a large set of products in 2011 for customer intelligence and numerous SAS modules for web, social media and marketing analytics that is widely used for profiling customers and prospects. It can also predict their behaviors, manage, and optimize communications.

5. Apache Spark

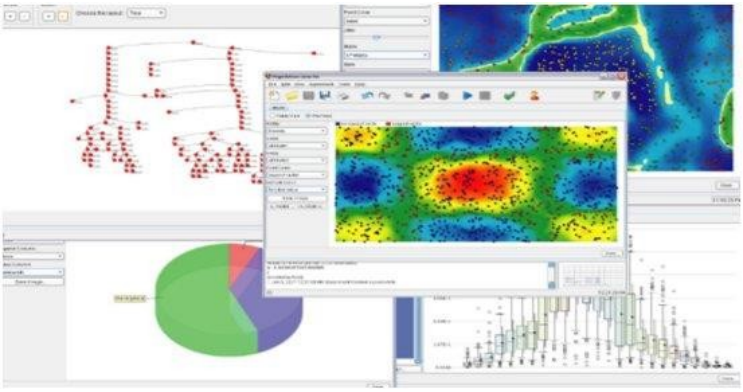
The University of California, Berkeley's AMP Lab, developed Apache in 2009. Apache Spark is a fast large-scale data processing engine and executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk. Spark is built on data science and its concept makes data science effortless. Spark is also popular for data pipelines and machine learning models development.
Spark also includes a library – MLlib, that provides a progressive set of machine algorithms for repetitive data science techniques like Classification, Regression, Collaborative Filtering, Clustering, etc.

6. Excel



Excel is a basic, popular and widely used analytical tool almost in all industries. Whether you are an expert in Sas, R or Tableau, you will still need to use Excel. Excel becomes important when there is a requirement of analytics on the client's internal data. It analyzes the complex task that summarizes the data with a preview of pivot tables that helps in filtering the data as per client requirement. Excel has the advance business analytics option which helps in modelling capabilities which have prebuilt options like automatic relationship detection, a creation of DAX measures and time grouping.

7. RapidMiner:



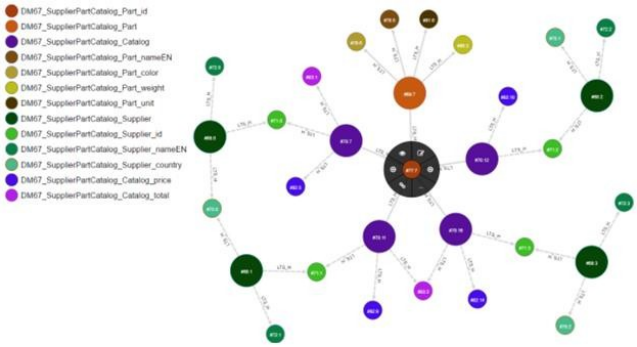
RapidMiner is a powerful integrated data science platform developed by the same company that performs predictive analysis and other advanced analytics like data mining, text analytics,

machine learning and visual analytics without any programming. RapidMiner can incorporate with any data source types, including Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase, IBM DB2, Ingres, MySQL, IBM SPSS, Dbase etc. The tool is very powerful that can generate analytics based on real-life data transformation settings, i.e. you can control the formats and data sets for predictive analysis.

8. KNIME

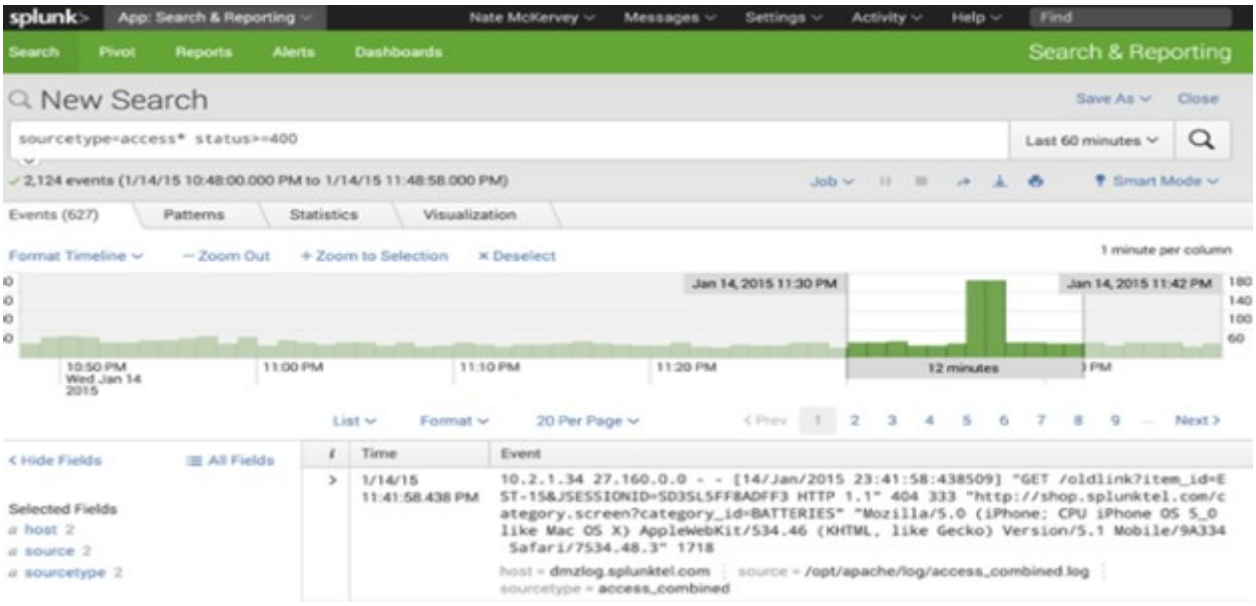
KNIME Developed in January 2004 by a team of software engineers at University of Konstanz. KNIME is leading open source, reporting, and integrated analytics tools that allow you to analyze and model the data through visual programming, it integrates various components for data mining and machine learning via its modular data-pipelining concept.

9. QlikView



QlikView has many unique features like patented technology and has in-memory data processing, which executes the result very fast to the end users and stores the data in the report itself. Data association in QlikView is automatically maintained and can be compressed to almost 10% from its original size. Data relationship is visualized using colors – a specific color is given to related data and another color for non-related data.

10. Splunk:



Splunk is a tool that analyzes and search the machine-generated data. Splunk pulls all text-based log data and provides a simple way to search through it, a user can pull in all kind of data, and perform all sort of interesting statistical analysis on it, and present it in different formats.

Reporting vs Analysis

Living in the era of digital technology and big data has made organizations dependent on the wealth of information data can bring. You might have seen how reporting and analysis are used interchangeably, especially the manner which outsourcing companies market their services. While both areas are part of web analytics (note that analytics isn't similar to analysis), there's a vast difference between them, and it's more than just spelling.

It's important that we differentiate the two because some organizations might be selling themselves short in one area and not reap the benefits, which web analytics can bring to the table. The first core component of web analytics, reporting, is merely organizing data into summaries. On the other hand, analysis is the process of inspecting, cleaning, transforming, and modeling these summaries (reports) with the goal of highlighting useful information.

Simply put, reporting translates data into information while analysis turns information into insights. Also, reporting should enable users to ask "What?" questions about the information, whereas analysis should answer to "Why"" and "What can we do about it?"

Here are five differences between reporting and analysis:

1. Purpose

Reporting helps companies monitor their data even before digital technology boomed. Various organizations have been dependent on the information it brings to their business, as reporting extracts that and makes it easier to understand.

Analysis interprets data at a deeper level. While reporting can link between cross-channels of data, provide comparison, and make understand information easier (think of a dashboard, charts, and graphs, which are reporting tools and not analysis reports), analysis interprets this information and provides recommendations on actions.

2. Tasks

As reporting and analysis have a very fine line dividing them, sometimes it's easy to confuse tasks that have analysis labeled on top of them when all it does is reporting. Hence, ensure that your analytics team has a healthy balance doing both.

Here's a great differentiator to keep in mind if what you're doing is reporting or analysis:

Reporting includes building, configuring, consolidating, organizing, formatting, and summarizing. It's very similar to the above mentioned like turning data into charts, graphs, and linking data across multiple channels.

Analysis consists of questioning, examining, interpreting, comparing, and confirming. With big data, predicting is possible as well.

3. Outputs

Reporting and analysis have the push and pull effect from its users through their outputs. Reporting has a push approach, as it pushes information to users and outputs come in the forms of canned reports, dashboards, and alerts.

Analysis has a pull approach, where a data analyst draws information to further probe and to answer business questions. Outputs from such can be in the form of ad hoc responses and analysis presentations. Analysis presentations are comprised of insights, recommended actions, and a forecast of its impact on the company—all in a language that's easy to understand at the level of the user who'll be reading and deciding on it.

This is important for organizations to realize truly the value of data, such that a standard report is not similar to a meaningful analytics.

4. Delivery

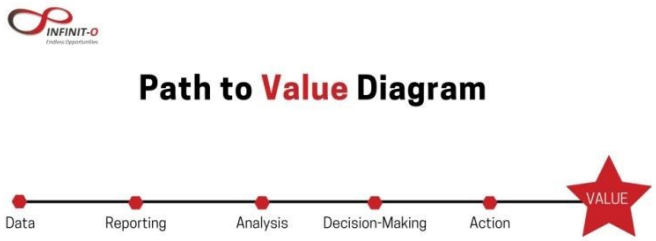
Considering that reporting involves repetitive tasks—often with truckloads of data, automation has been a lifesaver, especially now with big data. It's not surprising that the first thing outsourced are data entry services since outsourcing companies are perceived as data reporting experts.

Analysis requires a more custom approach, with human minds doing superior reasoning and analytical thinking to extract insights, and technical skills to provide efficient steps towards accomplishing a specific goal. This is why data analysts and scientists are demanded these days, as organizations depend on them to come up with recommendations for leaders or business executives make decisions about their businesses.

5. Value

This isn't about identifying which one brings more value, rather understanding that both are indispensable when looking at the big picture. It should help businesses grow, expand, move forward, and make more profit or increase their value.

This Path to Value diagram illustrates how data converts into value by reporting and analysis such that it's not achievable without the other.



Data alone is useless, and action without data is baseless. Both reporting and analysis are vital to bringing value to your data and operations.

Reporting and Analysis are Valuable

Not to undermine the role of reporting in web analytics, but organizations need to understand that reporting itself is just numbers. Without drawing insights and getting reports aligned with your organization's big picture, you can't make decisions based on reports alone.

Data analysis is the most powerful tool to bring into your business. Employing the powers of analysis can be comparable to finding gold in your reports, which allows your business to increase profits and further develop.

Areas where Data Analytics Applications have been employed:

Below are the various areas where data analytics applications have been employed:

1.) Policing/Security

Several cities all over the world have employed predictive analysis in predicting areas that would likely witness a surge in crime with the use of geographical data and historical data. This has seemed to work in major cities such as Chicago, London, Los Angeles, etc. Although, it is not possible to make arrests for every crime committed but the availability of data has made it possible to have police officers within such areas at a certain time of the day which has led to a drop in crime rate.

This shows that this kind of data analytics application will make us have safer cities without police putting their lives at risk.

2.) Transportation

A few years back at the London Olympics, there was a need for handling over 18 million journeys made by fans in the city of London and fortunately, it were sorted out.

How was this feat achieved? The TFL and train operators made use of data analytics to ensure the large numbers of journeys went smoothly. They were able to input data from events that took place and forecasted a number of persons that were going to travel; transport was being run efficiently and effectively so that athletes and spectators can be transported to and from the

respective stadiums.

3.) Fraud and Risk Detection

This has been known as one of the initial applications of data science which was extracted from the discipline of Finance. So many organizations had very bad experiences with debt and were so fed up with it. Since they already had data that was collected during the time their customers applied for loans, they applied data science which eventually rescued them from the losses they had incurred. This led to banks learning to divide and conquer data from their customers' profiles, recent expenditure and other significant information that were made available to them.

This made it easy for them to analyze and infer if there was any probability of customers defaulting.

4.) Manage Risk

In the insurance industry, risk management is the major focus. What most people aren't aware of is that when insuring a person, the risk involved is not obtained based on mere information but data that has been analyzed statistically before a decision is made. Data analytics gives insurance companies information on claims data, actuarial data and risk data covering all important decision that the company needs to take. Evaluation is done by an underwriter before an individual insured then the appropriate insurance is set.

These days, analytical software is used for detecting the various forms of fraudulent claims. Risky claims are detected by red flag indicators which can be examined. It is very essential to bring such claims to the attention of administrators, due to the manner at which automation is improving claims processing efficiency.

5.) Delivery Logistics

Well, data science and analytics have no limited applications. There are several logistic companies working all over the world such as UPS, DHL, FedEx, etc. that make use of data for improving their efficiency in operations. From data analytics applications, these companies have found the most suitable routes for shipping, the best delivery time, most suitable means of transport to select so as to gain cost efficiency and many others. Also, data generated by these companies through the use of GPS gives them enough opportunities to take advantage of data analytics and data science.

6.) Web Provision

There is this general belief that "Smart Cities" have fast internet speed provided either by their government or companies present there, therefore declaring them smart. Well, just because people can access Facebook or YouTube at the speed of lightning does not necessarily make a city smart.

Although there may be the presence of fast internet but this is just one thing; it needs to be present in the appropriate place and accessed by the right people as well. The key component of this is being able to shift bandwidth at the right time and location. This can only be achieved by the use of data.

The main assumption is that commercial and financial areas should have the highest bandwidth during weekdays while residential areas should get such on weekends. The real truth is that

this situation is more complex than it looks and this can only be solved by data analytics application. For example, if a particular community wants to get the attention of web development companies and high-tech industries and make them establish there, a higher
bandwidth would be required; only data analytics could get this done effectively.

7.) Proper Spending

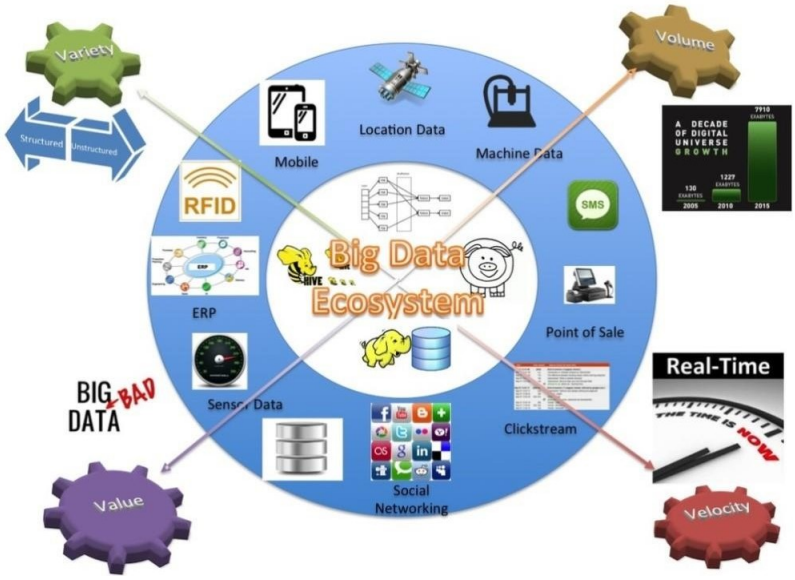
Another issue with Smart Cities is the large amount of money spent on little work. Small changes or landmark remodeling which one could dismiss as unnecessary projects consume so much money. Data analytics applications would target where taxpayers' money would have a major impact on and the kind of work that would be adequate for it. The targeting of where this money should be spent would lead to the entire city's infrastructure getting a facelift with a
reduction of excess money spent.

8.) Customer Interactions

This is another one of the applications of data analytics in insurance. Insurers can determine a lot about their services by conducting regular customer surveys mainly after interacting with claim handlers. They could use this to know which of their services are good and the ones that would need improvement. Various demographics may desire diverse methods of communication like in person interactions, websites, phone or just email. Taking the analysis of customer demographics with feedback can help insurers improve on customer experience depending on customer behavior and proven insights.

A study recently carried out showed that a lack of investment in technology was the cause customer dissatisfaction of the present generation of insurance customers because they prefer using mobile and online channels, social media and other recent mediums to interact with their agents. However, the older generation still prefers the use of the telephone. To improve the overall experience of customers, it is best for insurance companies to provide a wide range of

communication methods for their customers.



9.) City Planning

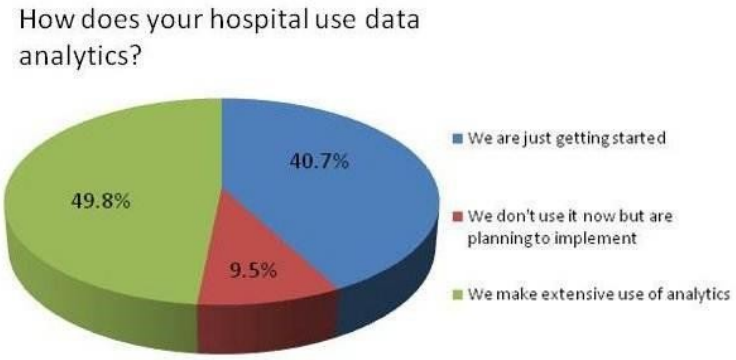
One big mistake being made in many places is that analytics is not considered when pursuing city planning. As a matter of fact, web traffic and marketing are still being used instead of the creation of spaces and buildings. This really causes a lot of issues to power over data due to its influence on things like building zoning and amenity creation. Models that are built will maximize the accessibility of specific areas or services while the risk of overloading significant elements of the infrastructure in the city is minimized. This implies that it creates efficiency.

We usually see buildings that are built on spots that look suitable but actually have a negative effect on other places. This is because such issues were not considered during the period of planning. Data analytics applications, as well as modeling, would make it easy to mark the
outcome of erecting a structure on any spot.

10.) Healthcare

One challenge most hospitals face is coping with cost pressures in treating as many patients as possible, considering the quality of healthcare's improvement. Machine and instrument data use has risen drastically so as to optimize and track treatment, patient flow as well as the use of equipment in hospitals. There is an estimation that a 1% efficiency gain will be achieved and

would result to over $63 billion in worldwide health care services.



11.) Travel

Data analytics applications help in the optimization of traveler's buying experience via social media and mobile/weblog data analysis. This is because customers' preferences and desires can be obtained from this, therefore, making companies sell products from the correlation of the current sales to recent browse-to-buy conversion through customized offers and packages. Data analytics applications can also deliver personalized travel recommendations depending
on the outcome from social media data.

12.) Energy Management

We are in an era where firms make apply data analytics to energy management and cover areas like energy optimization, smart-grid management, distribution of energy and building automation for utility companies. Data analytics application here focuses mainly on monitoring and controlling of dispatch crew, network devices and make sure service outages are properly managed. Utilities get the ability to integrate as much as millions of data points within the performance of the network which allows the engineers make use of the analytics in monitoring
the network.

13.) Internet/Web Search

When one mentions the word 'search', the first thing that comes to the mind is 'Google'. In fact, Google to some point can be used in place of 'search on the internet' by saying 'Google it'. Well, apart from Google, there are several other search engines such as Bing, Yahoo, Duckduckgo, AOL, Ask, etc. Each of these search engines is as a result of data science applications because they use algorithms to deliver the best results for any search query directed at them in just a split second. In respect to this, Google is known to process over 20

petabytes of data daily. Of course, without analytics and data science, this feat wouldn't have been possible.

14.) Digital Advertisement

Apart from web search, there is another area where data analytics and data science serves a very important purpose – digital advertisements. From the banners displayed on several websites to the digital billboards seen in the big cities; all are controlled by data algorithms.

This shows why digital adverts get more CTR than the conventional way of advertisements. Targets depend solely on the past behavior of users.

The importance of data analytics applications cannot be overemphasized because it is used in almost all areas of life today. We can see that having data is very important before making certain decisions so as to avoid unnecessary issues.

Also, handling valuable data inefficiently could lead to several problems like different departments in an organization not understanding how to make use of it which would lead to data not used to its full potential or serving any purpose.

BACKGROUND AND OVERVIEW OF DATA ANALYTICS LIFECYCLE

The Data Analytics Lifecycle defines analytics process best practices spanning discovery to project completion. The lifecycle draws from established methods in the realm of data analytics and decision science. This synthesis was developed after gathering input from data scientists and consulting established approaches that provided input on pieces of the process. Several of the processes that were consulted include these:

The scientific method, in use for centuries, still provides a solid framework for thinking about and deconstructing problems into their principal parts. One of the most valuable ideas of the scientific method relates to forming hypotheses and

finding ways to test ideas.

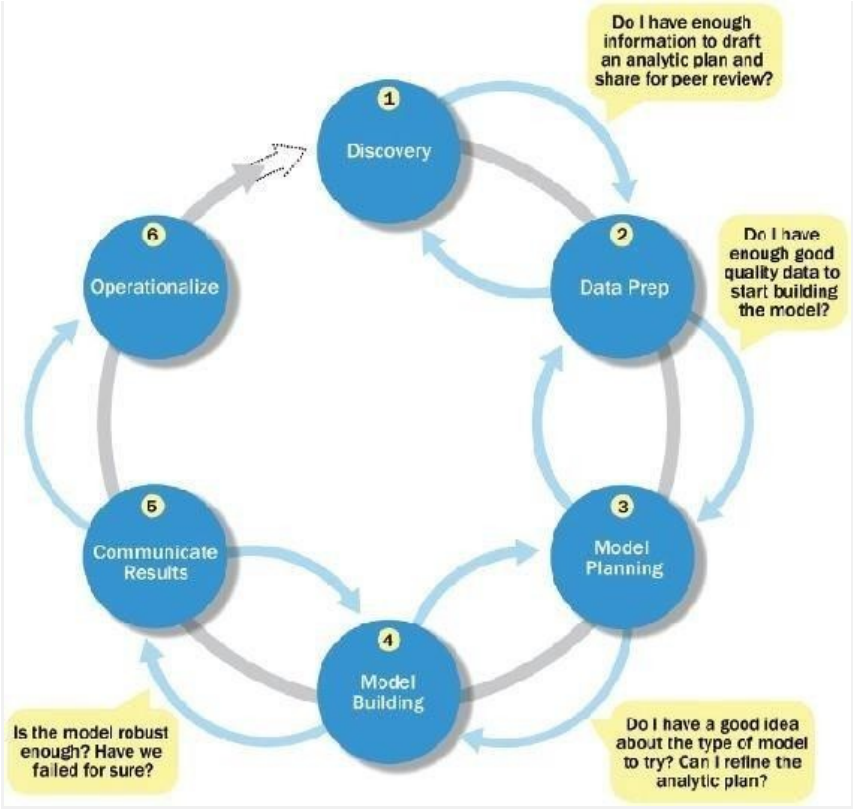
CRISP-DM provides useful input on ways to frame analytics problems and is a popular approach for data mining.

Tom Davenport's **DELTA** framework: The DELTA framework offers an approach for data analytics projects, including the context of the organization's skills, datasets, and leadership engagement.

Doug Hubbard's **Applied Information Economics (AIE)** approach: AIE provides a framework for measuring intangibles and guides on developing decision models, calibrating expert estimates, and deriving the expected value of information.

"MAD Skills" by Cohen et al. offers input for several of the techniques mentioned in Phases 2–4 that focuses on model planning, execution, and key findings.

Here is a brief overview of the main phases of the Data Analytics Lifecycle:



Phase 1 — Discovery: In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important

activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

Phase 2 — Data preparation: Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform

(ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed into the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data.

Phase 3 — Model planning: Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

Phase 4 — Model building: In Phase 4, the team develops data-sets for testing, training, and production purposes. Also, in this phase, the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows.

Phase 5 — Communicate results: In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in

Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Phase 6 — Operationalize: In Phase 6, the team delivers final reports, briefings, code, and technical documents. Also, the team may run a pilot project to implement the models in a production environment.

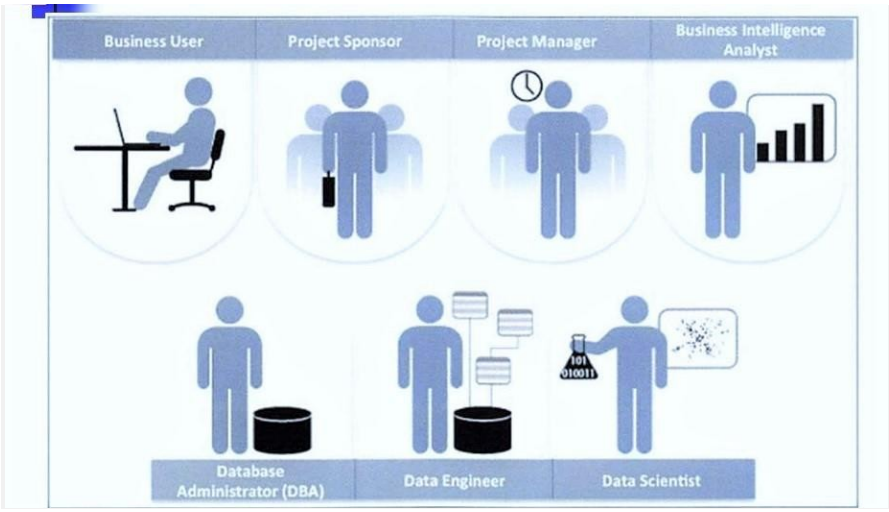
Key Roles for a Successful Analytics Project

There are seven key roles that are needed to be fulfilled for a high-functioning data science team to execute analytic projects successfully.

Each plays a critical part in a successful analytics project. Although seven roles are listed, fewer or more people can accomplish the work depending on the scope of the project, the organizational structure, and the skills of theparticipants.

For example, on a small, versatile team, these seven roles may be fulfilled by only 3 people, but a very large project may require 20 or morepeople.

The seven roles follow.



- **Business User:** Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team in the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.

• **Project Sponsor:** Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired

outputs.

• **Project Manager:** Ensures that key milestones and objectives are met on time and at the expected quality.

• **Business Intelligence Analyst:** Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics, and business intelligence from a reporting perspective. Business Intelligence Analysts generally create dashboards and reports and know the data feeds and sources.

• **Database Administrator (DBA):** Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

• **Data Engineer:** Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox, which was discussed in Chapter 1, "Introduction to Big Data Analytics." Whereas the DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

. Data Scientist: Provides subject matter expertise for analytical techniques, data modelling, and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the project.