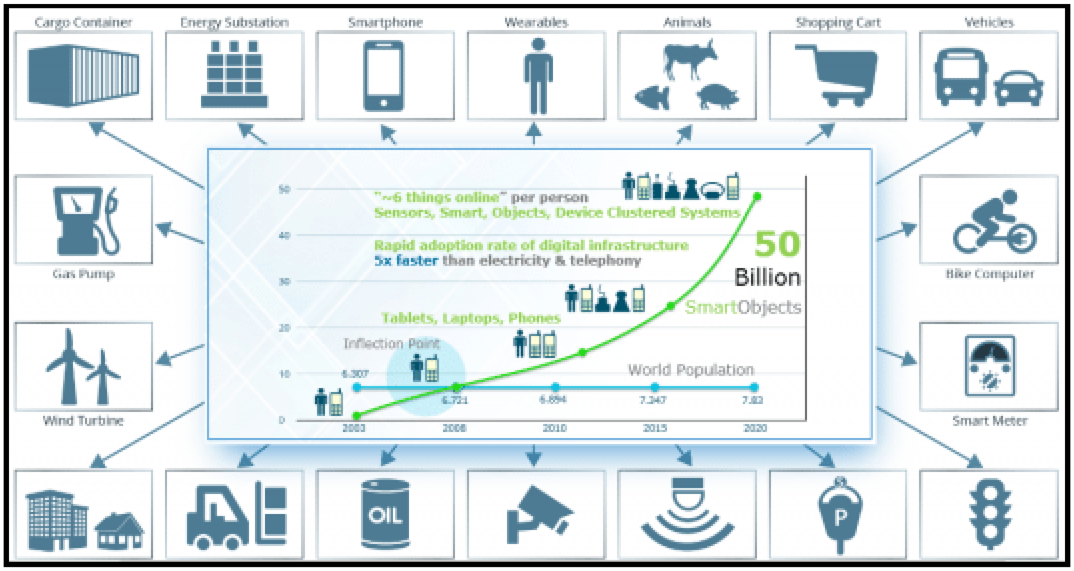
Big Data

Unit-1

Introduction to Big Data

Big Data Driving Factors



The quantity of data on planet earth is growing exponentially for many reasons. Various sources and our day to day activities generate lots of data. With invent of the web, the whole world has gone online, every single thing we do leaves a digital trace. With the smart objects going online, the data growth rate has increased rapidly. The major sources of Big Data are social media sites, sensor networks, digital images/videos, cell phones, purchase transaction records, web logs, medical records, archives, military surveillance, e-Commerce, complex scientific research and so on. By 2020, the data volumes will be around 40 Zettabytes which is equivalent to adding every single grain of sand on the planet multiplied by seventy-five.

Some of Big Data Drivers are:

Business: So what drivers make businesses tick? 1. Data driven initiatives: They are primarily categorized into 3 broad areas: a. Data Driven Innovation: I particularly like the Innovation aspect with being data driven. Imagine being able to learn from your customer first what they need and having the ability to drive innovation through those uber targeted data indicators. b. Data Driven Decision Making: Data driven decision-making is the inherent ability of analytics to sieve through globs of data and identify the best path forward. Whether in terms of finding the best route to validating the current route and estimating the success/failure in current strategy. It takes decision making away from gut and focus on data backed reasoning for higher chances of success. c. Data Driven Discovery: Your data know a whole lot about you than you image. Having a discovery mechanism will help you understand hidden insights that were not visible through traditional means.

2. Data Science as a competitive advantage: I had the fortune of interacting with couple of mid size company's executives from commodity businesses. There had been a consistent outcry on having to build big data as a capability to add to their competitive advantage. With a proper data driven framework, businesses could build sustainable capabilities and further leverage these capabilities as a competitive edge. If businesses were able to master big data driven capabilities, businesses could use these capabilities to establish secondary source of revenues by selling it to other businesses.

3. Sustained processes: Data driven approach creates sustainable processes, which gives a huge endorsement to big data analytics strategy as a go for enterprise adoption. Randomness kills businesses and adds scary risks, while data driven strategy reduces the risk by bringing statistical models, which are measurable.

4. Cost advantages of commodity hardware & open source software: Cost advantage is music to CXO's ears. How about the savings your IT will enjoy from moving things to commodity hardware and leverage more open source platforms for cost effective ways to achieve enterprise level computations and beyond. No more overpaying of premium hardware when similar or better analytical processing could be done using commodity and open source systems.

5. Quick turnaround and less bench times: Have you dealt with IT folks in your company? Mo and mo people, complex processes and communication charter gives you hard time connecting with someone who could get the task done. Things take forever long and cost fortunes with substandard quality.;

What is Big Data?

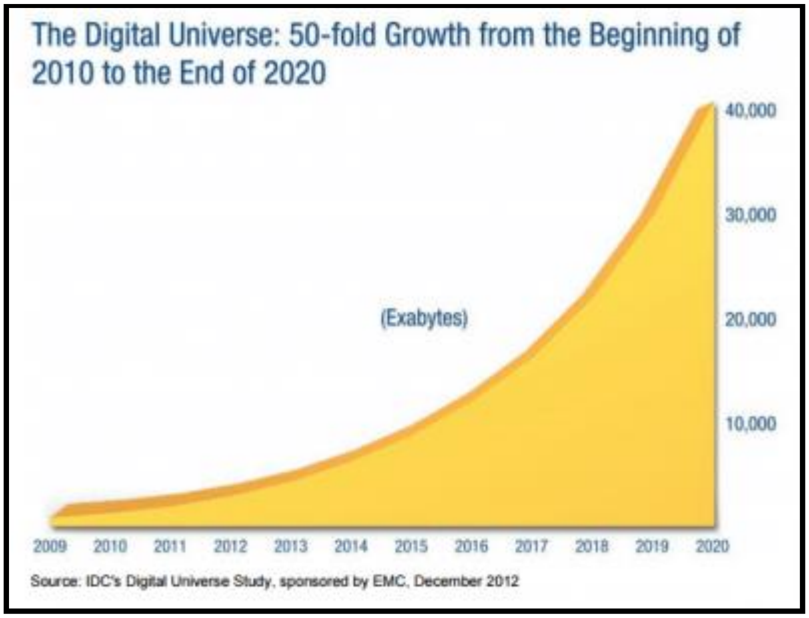
Big Data is a term used for a collection of data sets that are large and complex, which is difficult to store and process using available database management tools or traditional data processing applications. The challenge includes capturing, curating, storing, searching, sharing, transferring, analyzing and visualization of this data.

Big Data Characteristics

The five characteristics that define Big Data are: Volume, Velocity, Variety, Veracity and Value.

1. ***VOLUME***

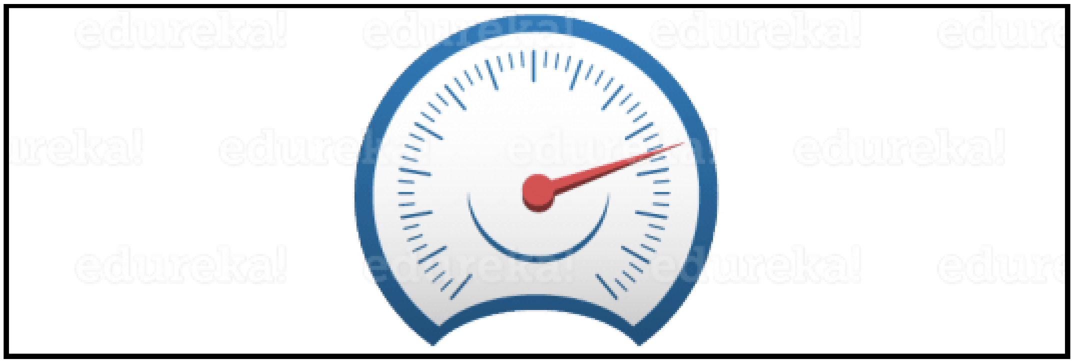
   Volume refers to the 'amount of data', which is growing day by day at a very fast pace. The size of data generated by humans, machines and their interactions on social media itself is massive. Researchers have predicted that 40 Zettabytes (40,000 Exabytes) will be generated by 2020, which is an increase of 300 times from 2005.



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

2. ***VELOCITY***

Velocity is defined as the pace at which different sources generate the data every day. This flow of data is massive and continuous. There are 1.03 billion Daily Active Users (Facebook DAU) on Mobile as of now, which is an increase of 22% year-over-year. This shows how fast the numbers of users are growing on social media and how fast the data is getting generated daily. If you are able to handle the velocity, you will be able to generate insights and take decisions based on real-time data.

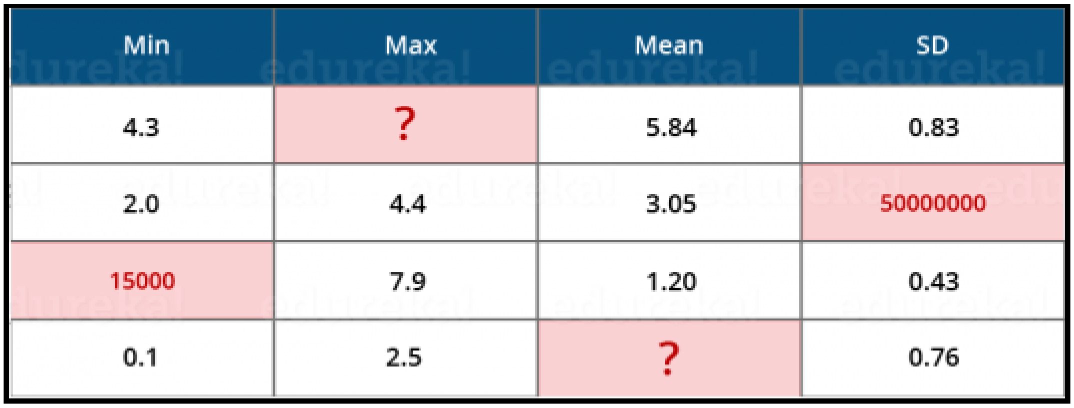


3. ***VARIETY***

As there are many sources which are contributing to Big Data, the type of data they are generating is different. It can be structured, semi-structured or unstructured. Hence, there is a variety of data which is getting generated every day. Earlier, we used to get the data from excel and databases, now the data are coming in the form of images, audios, videos, sensor data etc. as shown in below image. Hence, this variety of unstructured data creates problems in capturing, storage, mining and analyzing the data.

4. ***VERACITY***

Veracity refers to the data in doubt or uncertainty of data available due to data inconsistency and incompleteness. In the image below, you can see that few values are missing in the table. Also, a few values are hard to accept, for example – 15000 minimum values in the 3rd row, it is not possible. This inconsistency and incompleteness is Veracity.



Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Data available can sometimes get messy and maybe difficult to trust. With many forms of big data, quality and accuracy are difficult to control like Twitter posts with hashtags, abbreviations, typos and colloquial speech. The volume is often the reason behind for the lack of quality and accuracy in the data.

- Due to uncertainty of data, 1 in 3 business leaders don't trust the information they use to make decisions.
- It was found in a survey that 27% of respondents were unsure of how much of their data was inaccurate.
- Poor data quality costs the US economy around $3.1 trillion a year.

5. ***VALUE***

After discussing Volume, Velocity, Variety and Veracity, there is another V that should be taken into account when looking at Big Data i.e. Value. It is all well and good to have access to big data but unless we can turn it into value it is useless. By turning it into value I mean, is it adding to the benefits of the organizations who are analyzing big data? Is the organization working on Big Data achieving high ROI (Return on Investment)? Unless, it adds to their profits by working on Big Data, it is useless.

Types of Big Data

Big Data could be of three types:

- Structured
- Semi-Structured
- Unstructured

1. ***Structured***

   The data that can be stored and processed in a fixed format is called as Structured Data. Data stored in a relational database management system (RDBMS) is one example of 'structured' data. It is easy to process structured data as it has a fixed schema. Structured Query Language (SQL) is often used to manage such kind of Data.

2. ***Semi-Structured***

   Semi-Structured Data is a type of data which does not have a formal structure of a data model, i.e. a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that make it easier to analyze. XML files, HTML files or JSON documents are examples of semi-structured data.

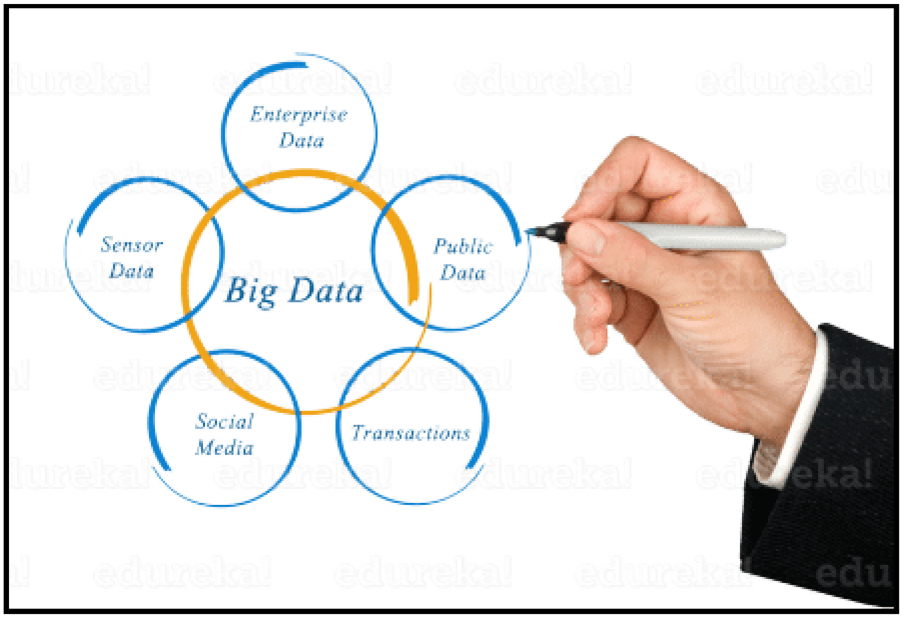
3. ***Unstructured***

   The data which have unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data. Text Files and multimedia contents like images, audios, videos are example of unstructured data. The unstructured data is growing quicker than others, experts say that 80 percent of the data in an organization are unstructured.

Till now, I have just covered the introduction of Big Data. Furthermore, this Big Data tutorial talks about examples, applications and challenges in Big Data.

Examples of Big Data

Daily we upload millions of bytes of data. 90 % of the world's data has been created in last two years.



- Walmart handles more than **1 million** customer transactions every hour.
- Facebook stores, accesses, and analyzes **30+ Petabytes** of user generated data.
- **230+ millions** of tweets are created every day.
- More than **5 billion** people are calling, texting, tweeting and browsing on mobile phones worldwide.
- YouTube users upload **48 hours** of new video every minute of the day.
- Amazon handles **15 million** customer click stream user data per day to recommend products.
- **294 billion** emails are sent every day. Services analyses this data to find the spams.
- Modern cars have close to **100 sensors** which monitors fuel level, tire pressure etc. , each vehicle generates a lot of sensor data.

Applications of Big Data

We cannot talk about data without talking about the people, people who are getting benefited by Big Data applications. Almost all the industries today are leveraging Big Data applications in one or the other way.

- **Smarter Healthcare**: Making use of the Petabytes of patient's data, the organization can extract meaningful information and then build applications that can predict the patient's deteriorating condition in advance.

- **Telecom**: Telecom sectors collects information analyzes it and provides solutions to different problems. By using Big Data applications, telecom companies have been able to significantly reduce data packet loss, which occurs when networks are overloaded, and thus, providing a seamless connection to their customers.

- **Retail**: Retail has some of the tightest margins, and is one of the greatest beneficiaries of big data. The beauty of using big data in retail is to understand consumer behavior. Amazon's recommendation engine provides suggestion based on the browsing history of the consumer.

- **Traffic control**: Traffic congestion is a major challenge for many cities globally. Effective use of data and sensors will be a key to managing traffic better as cities become increasingly densely populated.

- **Manufacturing**: Analyzing big data in the manufacturing industry can reduce component defects, improve product quality, increase efficiency, and save time and money.

- **Search Quality**: Every time we are extracting information from google, we are simultaneously generating data for it. Google stores this data and uses it to improve its search quality.

Challenges with Big Data

Let me tell you few challenges which come along with Big Data:

1. **Data Quality –** The problem here is the 4th V i.e. Veracity. The data here is very messy, inconsistent and incomplete. Dirty data cost $600 billion to the companies every year in the United States.

2. **Discovery –** Finding insights on Big Data is like finding a needle in a haystack. Analyzing Petabytes of data using extremely powerful algorithms to find patterns and insights are very difficult.

3. **Storage –** The more data and organization has the more complex the problems of managing it can become. The question that arises here is "Where to store it?"We need a storage system which can easily scale up or down on-demand.

4. **Analytics –** In the case of Big Data, most of the time we are unaware of the kind of data we are dealing with, so analyzing that data is even more difficult.

5. **Security –** Since the data is huge in size, keeping it secure is another challenge. It includes user authentication, restricting access based on a user, recording data access histories, proper use of data encryption etc.

6. **Lacks of Talent – There are** a lot of Big Data projects in major organizations, but a sophisticated team of developers, data scientists and analysts who also have sufficient amount of domain knowledge is still a challenge.

What is Big Data Technology?

Big Data Technology can be defined as a Software-Utility that is designed to **Analyze**, **Process** and **Extract** the information from an extremely complex and large data sets which the **Traditional Data Processing Software** could never deal with.

We need Big Data Processing Technologies to analyze this huge amount of Real-time data and come up with Conclusions and Predictions to reduce the risks in the future.

Categories in which the **Big Data Technologies** are classified:

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 **Particular Organization** 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.

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 Analytical 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.

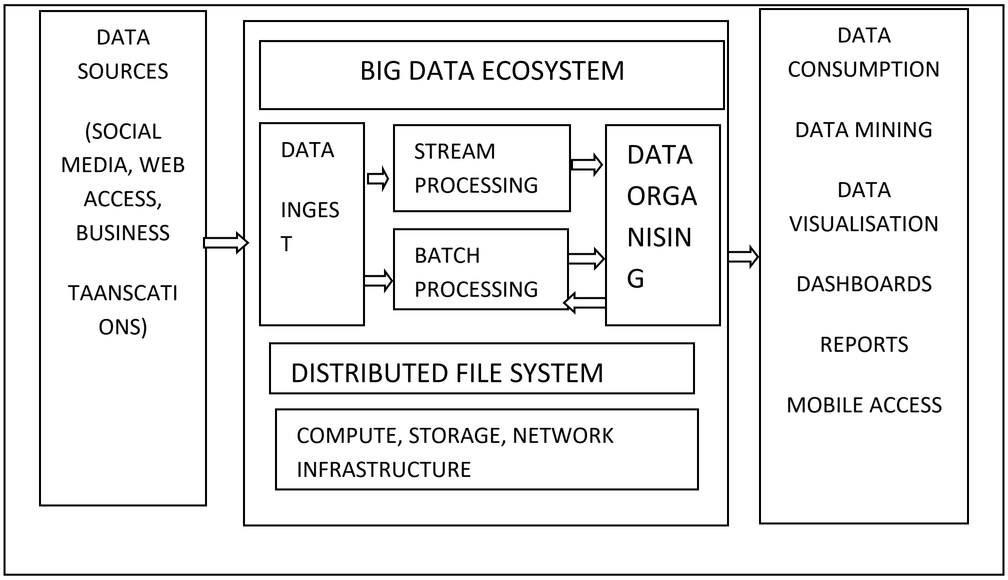
- Medical fields where a particular patients health status can be monitored.

Top Big Data Technologies

Top big data technologies are divided into **4** fields which are classified as follows:

- **Data Storage**
- **Data Mining**
- **Data Analytics**
- **Data Visualization**

Big Data architecture



- SOURCE LAYER:

  The choice of data for an application depends upon what data is required to perform the kind of analysis you need. Big data vary in origin, size, speed, form and function. Data sources can be internal or external to the organization. The Scope of access to data available could be limited. The level of structure could be high or low. The Speed of data and its quantity will also by high or low depending upon the data source.

- DATA INGEST LAYER:

The layer is responsible for acquiring data from the data sources . The data is received through a scalable set of input points that can acquire data at various speeds. The data is sent to a batch processing system, a stream processing system or directly to a storage file system such as Hadoop, compliance regulations and government policies impact what can be stored and for how long.

- BATCH PROCESSING LAYER:

  The analysis layer receives data from the ingest point or from the file system or from the No SQL Databases. Data is processed using parallel programming techniques (such as Map Reduce) to process it and produce the desired results. The batch processing layer thus needs to understands the data sources and data types, the algorithms that would work on that data and the format of the desired outcomes. The output of this layer could be sent for instant reporting or stored in a No SQL Databases for an on demand report for the client.

- STREAM PROCESSING LAYER:

  The technology layer receives data directly from the ingest point. Data is processed using parallel processing techniques to process it in real time and produce the desired results. This layer, thus needs to understand the data sources and data types extremely well and the super light algorithms that would work on that data to produce the desired results. The outcome of this layer too could be stored in the NOSQL Databases.

- DATA ORGANISING LAYERS:

  This layer receives data from both the batch and stream processing layers. Its objective is to organize the data for easy access. It is represented by the NO SQL Databases. There are a variety of NO SQL Databases to suit different needs. SQL like languages like PIG and HIVE can be used easily access data and generate reports from these databases.

- INFRASTRUCTURE LAYER:

  At the bottom there is a layer that manages the raw resources of storage, compute and communication. This is increasingly provided through a cloud computing paradigm.

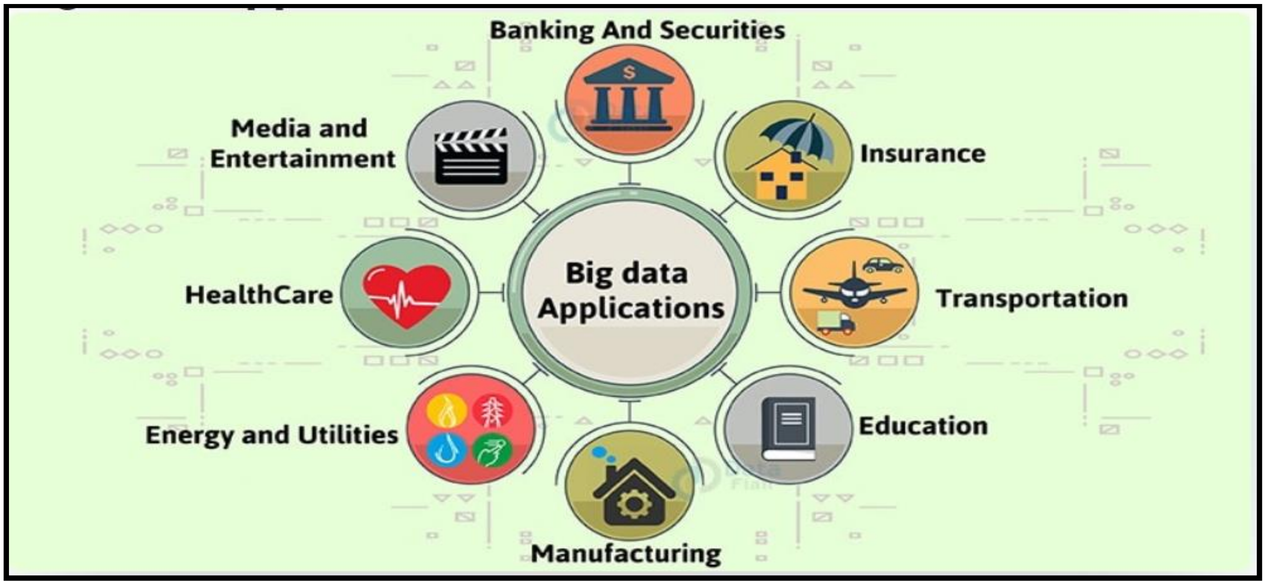
- DISTRIBUTED FILE SYSTEM LAYER:

  This is a heart of a Big Data System. It would store huge quantities of data and make it quickly and securely, available and accessible to the other layers. HDFS is the primary technology in this layer. It would include supporting applications such as YARN that enable the efficient access to data storage and its transfer.

- DATA CONSUMPTION LAYER:

This is the final layer and it consumes the output provided by the analysis layers directly or through the organizing layer. The outcome could be standard reports, data analytics, dash boards and other visualization applications, recommendation engine on mobile or other devices.

Big Data Applications



- **Government**
  Big data analytics has proven to be very useful in the government sector. Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign. Also most recently, Big data analysis was majorly responsible for the BJP and its allies to win a highly successful Indian General Election 2014. The Indian Government utilizes numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation.
- **Social Media Analytics**
  The advent of social media has led to an outburst of big data. Various solutions have been built in order to, analyze social media activity like IBM's Cognos Consumer Insights a point solution running on IBM's BigInsights Big Data platform, can make sense of the chatter. Social media can provide valuable real-time insights into how the market is responding to products and campaigns. With the help of these insights, the companies can adjust their pricing, promotion, and campaign placements accordingly. Before utilizing the big data there needs to be some preprocessing to be done on the big data in order to derive some intelligent and valuable results. Thus to know the consumer mindset the application of intelligent decisions derived from big data is necessary.
- **Technology**

The technological applications of big data comprise of the following companies which deal with huge amounts of data every day and put them to use for business decisions as well. For example, eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay‟s 90PB data warehouse. Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005, they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB. Facebook handles 50 billion photos from its user base. Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

- **Fraud detection**

  For businesses whose operations involve any type of claims or transaction processing, fraud detection is one of the most compelling Big Data application examples. Historically, fraud detection on the fly has proven an elusive goal. In most cases, fraud is discovered long after the fact, at which point the damage has been done and all that's left is to minimize the harm and adjust policies to prevent it from happening again. Big Data platforms that can analyze claims and transactions in real time, identifying large-scale patterns across many transactions or detecting anomalous behavior from an individual user, can change the fraud detection game.
- **Call Center Analytics**

  Now we turn to the customer-facing Big Data application examples, of which call center analytics are particularly powerful. What's going on in a customer's call center is often a great barometer and influencer of market sentiment, but without a Big Data solution, much of the insight that a call center can provide will be overlooked or discovered too late. Big Data solutions can help identify recurring problems or customer and staff behavior patterns on the fly not only by making sense of time/quality resolution metrics but also by capturing and processing call content itself.
- **Banking**

  The use of customer data invariably raises privacy issues. By uncovering hidden connections between seemingly unrelated pieces of data, big data analytics could potentially reveal sensitive personal information. Research indicates that 62% of bankers are cautious in their use of big data due to privacy issues. Further, outsourcing of data analysis activities or distribution of customer data across departments for the generation of richer insights also amplifies security risks. Such as customers' earnings, savings, mortgages, and insurance policies ended up in the wrong hands. Such incidents reinforce concerns about data privacy and discourage customers from sharing personal information in exchange for customized offers.
- **Agriculture**

A biotechnology firm uses sensor data to optimize crop efficiency. It plants test crops and runs simulations to measure how plants react to various changes in condition. Its data environment constantly adjusts to changes in the attributes of various data it collects, including temperature, water levels, soil composition, growth, output, and gene sequencing of each plant in the test bed. These simulations allow it to discover the optimal environmental conditions for specific gene types.

- **Marketing**

  Marketers have begun to use facial recognition software to learn how well their advertising succeeds or fails at stimulating interest in their products. A recent study published in the Harvard Business Review looked at what kinds of advertisements compelled viewers to continue watching and what turned viewers off. Among their tools was "a system that analyses facial expressions to reveal what viewers are feeling." The research was designed to discover what kinds of promotions induced watchers to share the ads with their social network, helping marketers create ads most likely to "go viral" and improve sales.
- **Smart Phones**

  Perhaps more impressive, people now carry facial recognition technology in their pockets. Users of I Phone and Android smartphones have applications at their fingertips that use facial recognition technology for various tasks. For example, Android users with the remember app can snap a photo of someone, then bring up stored information about that person based on their image when their own memory lets them down a potential boon for salespeople.
- **Telecom**

  Now a day's big data is used in different fields. In telecom also it plays a very good role. Operators face an uphill challenge when they need to deliver new, compelling, revenue-generating services without overloading their networks and keeping their running costs under control. The market demands new set of data management and analysis capabilities that can help service providers make accurate decisions by taking into account customer, network context and other critical aspects of their businesses. Most of these decisions must be made in real time, placing additional pressure on the operators. Real-time predictive analytics can help leverage the data that resides in their multitude systems, make it immediately accessible and help correlate that data to generate insight that can help them drive their business forward.
- **Healthcare**

  Traditionally, the healthcare industry has lagged behind other industries in the use of big data, part of the problem stems from resistance to change providers are accustomed to making treatment decisions independently, using their own clinical judgment, rather than relying on protocols based on big data. Other obstacles are more structural in nature. This is one of the best place to set an example for Big Data Application. Even within a single hospital, payor, or pharmaceutical company, important information often remains siloed within one group or department because organizations lack procedures for integrating data and communicating findings.

Health care stakeholders now have access to promising new threads of knowledge. This information is a form of “big data,” so called not only for its sheer volume but for its complexity, diversity, and timelines. Pharmaceutical industry experts, payers, and providers are now beginning to analyze big data to obtain insights. Recent technologic advances in the industry have improved their ability to work with such data, even though the files are enormous and often have different database structures and technical characteristics.

Big Data features

1. Data Processing

Data processing features involve the collection and organization of raw data to produce meaning. Data modeling takes complex data sets and displays them in a visual diagram or chart. This makes it digestible and easy to interpret for users trying to utilize that data to make decisions.

Data mining allows users to extract and analyze data from different perspectives and summarize it into actionable insights. It is especially useful on large unstructured data sets collected over a period of time.

Big Data analytics tools should enable data import from sources such as Microsoft Access, Microsoft Excel, text files and other flat files. Being able to merge data from multiple sources and in multiple formats will reduce labor by preventing the need for data conversion and speed up the overall process by importing directly to the system.

The same goes for export capabilities — being able to take the visualized data sets and export them as PDFs, Excel files, Word files or .dat files is crucial to the usefulness and transferability of the data collected in earlier processes.

Modeling
Data Mining
Data File Sources
File Exporting

2. Predictive Applications

Identity management (or identity and access management) is the organizational process for controlling who has access to your data. Identity management functionality manages identifying data for everything that has access to a system including individual users, computer hardware and software applications.

Identity management also deals with issues including how users gain an identity with access, protection of those identities and support for other system protections such as network protocols and passwords. It determines whether a user has access to a system and the level of access that user has permission to utilize.

Identity management applications aim to ensure only authenticated users can access your system and, by extension, your data. It is a crucial element of any organization's security plan and will include real-time security and fraud analytics capabilities.

Fraud analytics involve a variety of fraud detection functionalities. Too many businesses are reactive when it comes to fraudulent activities — they deal with the impact rather than proactively preventing it. Data analytics tools can play a role in fraud detection by offering repeatable tests that can run on your data at any time, ensuring you'll know if anything is amiss. You also have wider coverage of your data as a whole rather than relying on spot checking at financial transactions. Analytics can be an early warning tool to quickly and efficiently identify potentially fraudulent activity before it has a chance to impact your business at large.

Identity Management
Fraud Analytics

3. Analytics

Big Data analytics tools offer a variety of analytics packages and modules to give users options. RIsk analytics, for example, is the study of the uncertainty surrounding any given action. It can be used in combination with forecasting to minimize the negative impacts of future events. Risk analytics allow users to mitigate these risks by clearly defining and understanding their organization's tolerance for and exposure to risk.

Decision management involves the decision making processes of running a business. Decision management modules treat decisions as usable assets. It incorporates technology at key points to automate parts of that decision making process.

Text analytics is the process of examining text that was written about or by customers. Analytics software helps you find patterns in that text and offers potential actions to be taken based on what you learn. This kind of analytics is particularly useful for drawing insight about your customers' wants and needs directly from their interactions with your organization.

Content analysis is very similar to text analysis but includes the analysis of all formats of documentation including audio, video, pictures, etc. Social media analytics is one form of content analysis that focuses on how your user base is interacting with your brand on social media.

Statistical analytics collects and analyzes data sets composed of numbers. The goal is to draw a sample from the total data that is representative of a total population. Statistical analysis takes place in five steps: describing the nature of the data, exploring the relation of the data to the population that provided it, creating a model to summarize the connections, proving or disproving its validity, and employing predictive analytics to guide decision-making.

Predictive analytics is a natural next step to statistical analytics. This feature takes the data collected and analyzed, offers what-if scenarios, and predicts potential future problems.

Risk Analytics
Decision Management
Text Analytics
Content Analytics
Statistical Analysis
Predictive Analytics
Social Media Analytics

4. Reporting Features

Reporting functions keep users on top of their business. Real-time reporting gathers minute-by-minute data and relays it to you, typically in an intuitive dashboard format. This allows users to make snap decisions in heavily time-constrained situations and be both more prepared and more competitive in a society that moves at the speed of light.

Dashboards are data visualization tools that present metrics and KPIs. They are often customizable to report on a specific metric or targeted data set. One example of a targeted metric is location-based insights — these are data sets gathered from or filtered by location that can garner useful information about demographics.

Real-Time Reporting
Dashboards
Location-Based Insights

5. Security Features

Keeping your system safe is crucial to a successful business. Big Data analytics tools should offer security features to ensure security and safety. One such feature is single sign-on. Also called SSO, it is an authentication service that assigns users a single set of login credentials to access multiple applications. It authenticates end user permissions and eliminates the need to login multiple times during the same session. It can also log and monitor user activities and accounts to keep track of who is doing what in the system.

Another security feature offered by Big Data analytics platforms is data encryption. Data encryption involves changing electronic information into unreadable formats by using algorithms or codes. While web browsers offer automatic encryption, you want something a bit more robust for your sensitive proprietary data. Make sure the system offers comprehensive encryption capabilities when looking for a data analytics application.

Big Data privacy and ethics

Big data analytics raises a number of ethical issues, especially as companies begin monetizing their data externally for purposes different from those for which the data was initially collected. The scale and ease with which analytics can be conducted today completely changes the ethical framework. We can now do things that were impossible a few years ago, and existing ethical and legal frameworks cannot prescribe what we should do. While there is still no black or white, experts agree on a few principles:

1. **Private customer data and identity should remain private:** Privacy does not mean secrecy, as private data might need to be audited based on legal requirements, but that private data obtained from a person with their consent should not be exposed for use by other businesses or individuals with any traces to their identity.
2. **Shared private information should be treated confidentially:** Third party companies share sensitive data — medical, financial or locational — and need to have restrictions on whether and how that information can be shared further.
3. **Customers should have a transparent view** of how our data is being used or sold, and the ability to manage the flow of their private information across massive, third-party analytical systems.
4. **Big Data should not interfere with human will:** Big data analytics can moderate and even determine who we are before we make up our own minds. Companies need to begin to think about the kind of predictions and inferences that should be allowed and the ones that should not.
5. **Big data should not institutionalize unfair biases** like racism or sexism. Machine learning algorithms can absorb unconscious biases in a population and amplify them via training samples.

Big Data Analytics

- Big data analytics is the often complex process of examining big data to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

- On a broad scale, data analytics technologies and techniques give organizations a way to analyze data sets and gather new information. Business intelligence (BI) queries answer basic questions about business operations and performance.

- Big data analytics is a form of advanced analytics, which involve complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by analytics systems.

Why is big data analytics important?

Organizations can use big data analytics systems and software to make data-driven decisions that can improve business-related outcomes. The benefits may include more effective marketing, new revenue opportunities, customer personalization and improved operational efficiency. With an effective strategy, these benefits can provide competitive advantages over rivals.

How does big data analytics work?

Data analysts, data scientists, predictive modelers, statisticians and other analytics professionals collect, process, clean and analyze growing volumes of structured transaction data as well as other forms of data not used by conventional BI and analytics programs.

Here is an overview of the four steps of the data preparation process:

1. Data professionals **collect** data from a variety of different sources. Often, it is a mix of semi-structured and unstructured data. While each organization will use different data streams, some common sources include:

- internet clickstream data;
- web server logs;
- cloud applications;
- mobile applications;
- social media content;
- text from customer emails and survey responses;

- mobile phone records; and
- machine data captured by sensors connected to the internet of things (IoT).

2. Data is **processed**. After data is collected and stored in a data warehouse or data lake, data professionals must organize, configure and partition the data properly for analytical queries. Thorough data processing makes for higher performance from analytical queries.
3. Data is **cleansed** for quality. Data professionals scrub the data using scripting tools or enterprise software. They look for any errors or inconsistencies, such as duplications or formatting mistakes, and organize and tidy up the data.
4. The collected, processed and cleaned data is **analyzed** with analytics software. This includes tools for:

- data mining, which sifts through data sets in search of patterns and relationships
- predictive analytics, which builds models to forecast customer behavior and other future developments
- machine learning, which taps algorithms to analyze large data sets
- deep learning, which is a more advanced offshoot of machine learning
- text mining and statistical analysis software
- artificial intelligence (AI)
- mainstream business intelligence software
- data visualization tools

MODERN BIG DATA ANALYTIC TOOLS:

Big Data Analytics software is widely used in providing meaningful analysis of a large set of data. This software analytical tools help in finding current market trends, customer preferences, and other information.

Here are the Best Big Data Analytics Tools with key feature:

1) Xplenty: is a cloud-based ETL solution providing simple visualized data pipelines for automated data flows across a wide range of sources and destinations. Xplenty's powerful on-platform transformation tools allow you to clean, normalize, and transform data while also adhering to compliance best practices.

Features:

- Powerful, code-free, on-platform data transformation offering
- Rest API connector - pull in data from any source that has a Rest API
- Destination flexibility - send data to databases, data warehouses, and Salesforce
- Security focused - field-level data encryption and masking to meet compliance requirements
- Rest API - achieve anything possible on the Xplenty UI via the Xplenty API
- Customer-centric company that leads with first-class support

2) Analytics

Analytics is a tool that provides visual analysis and dash boarding. It allows you to connect multiple data sources, including business applications, databases, cloud drives, and more.

Features:

- Offers visual analysis and dash boarding.
- It helps you to analyze data in depth.
- Provides collaborative review and analysis.
- You can embed reports to websites, applications, blogs, and more.

3) Microsoft HDInsight

Azure HDInsight is a Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads.

Features:

- Reliable analytics with an industry-leading SLA
- It offers enterprise-grade security and monitoring
- Protect data assets and extend on-premises security and governance controls to the cloud
- High-productivity platform for developers and scientists
- Integration with leading productivity applications
- Deploy Hadoop in the cloud without purchasing new hardware or paying other up-front costs

4) Skytree:

Skytree is one of the best big data analytics tools that empowers data scientists to build more accurate models faster. It offers accurate predictive machine learning models that are easy to use.

Features:

- Highly Scalable Algorithms
- Artificial Intelligence for Data Scientists
- It allows data scientists to visualize and understand the logic behind ML decisions
- Skytree via the easy-to-adopt GUI or programmatically in Java
- Model Interpretability
- It is designed to solve robust predictive problems with data preparation capabilities
- Programmatic and GUI Access

5) Splice Machine:

Splice Machine is one of the best big data analytics tools. Their architecture is portable across public clouds such as AWS, Azure, and Google.

Features:

- It is a big data analytics software that can dynamically scale from a few to thousands of nodes to enable applications at every scale
- The Splice Machine optimizer automatically evaluates every query to the distributed HBase regions
- Reduce management, deploy faster, and reduce risk
- Consume fast streaming data, develop, test and deploy machine learning models

6) Spark:

Apache Spark is one of the powerful open source big data analytics tools. It offers over 80 high-level operators that make it easy to build parallel apps. It is one of the open source data analytics tools used at a wide range of organizations to process large datasets.

Features:

- It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk
- It is one of the open source data analytics tools that offers lighting Fast Processing
- Support for Sophisticated Analytics
- Ability to Integrate with Hadoop and Existing Hadoop Data
- It is one of the open source big data analytics tools that provides built-in APIs in Java, Scala, or Python

7) Apache SAMOA:

Apache SAMOA is a big data analytics tool. It is one of the big data analysis tools which enables development of new ML algorithms. It provides a collection of distributed algorithms for common data mining and machine learning tasks.

8) R-Programming:

R is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

Features:

- Effective data handling and storage facility,
- It provides a suite of operators for calculations on arrays, in particular, matrices,
- It provides coherent, integrated collection of big data tools for data analysis
- It provides graphical facilities for data analysis which display either on-screen or on hardcopy

CHALLENGES OF CONVENTIONAL SYSTEM:

1. Data 2. Process 3. Management

Three challenges that Big Data Face:

VOLUME:

1. Volume of data, especially machine generated data is exploding

2. How fast that data is growing every year, with new sources of data that are emerging

3. For Example: In the year 2000, 800,000 petabytes of data were stored in the world and it is expected to reach 35 Zetta Bytes by 2020 (according to IBM)

Processing:

1. More than 80% of today's information is unstructured and it is typically too big to manage effectively.

2. Today's Companies are looking to leverage a lot more data from a variety of sources both inside and outside the organization.

3. Things like contracts, documents, machine data, social media data, health records, emails etc. This List is endless.

Management:

A lot of data is unstructured or has a complex structure that is hard to represent in rows and columns.

INTELLIGENT DATA ANALYSIS:

Intelligent data analysis reveals implicit, previously unknown and potentially valuable information or knowledge from large amounts of data. Intelligent data analysis is also a kind of decision support process. Based on artificial intelligence, machine learning, pattern recognition, statistics, database and visualization technology mainly, IDA automatically extracts useful information, necessary knowledge and interesting models from a lot of online data in order to help decision makers make the right choices.

The process of IDA generally consists of the following three stages: (1) data preparation; (2) rule finding or data mining; (3) result validation and explanation. Data preparation involves selecting the required data from the relevant data source and integrating this into a data set to be used for data mining. Rule finding is working out rules contained in the data set by means of certain methods or algorithms. Result validation requires examining these rules, and result explanation is giving intuitive, reasonable and understandable descriptions using logical reasoning.

As the goal of intelligent data analysis is to extract useful knowledge, *the process demands a combination of extraction, analysis, conversion, classification, organization, reasoning, and so on. It is challenging and fun working out how to choose appropriate methods to resolve the difficulties encountered in the process. Intelligent data analysis methods and tools, as well as the authenticity of obtained results* pose us continued *challenges.*

NATURE OF DATA:

So, we have to start with the basics: the nature of data. There are four types of data:

- Nominal
- Ordinal
- Interval
- Ratio

Each offers a unique set of characteristics, which impacts the type of analysis that can be performed.

The distinction between the four types of scales center on three different characteristics:

1. The **order** of responses – whether it matters or not
2. The **distance between observations** – whether it matters or is interpretable
3. The presence or inclusion of a **true zero**

Nominal Scales

Nominal scales measure categories and have the following characteristics:

- **Order:** The order of the responses or observations does not matter.
- **Distance:** Nominal scales do not hold distance. The distance between a 1 and a 2 is not the same as a 2 and 3.
- **True Zero:** There is no true or real zero. In a nominal scale, zero is uninterpretable.

Consider traffic source (or last touch channel) as an example in which visitors reach our site through a mutually exclusive channel, or last point of contact. These channels would include:

1. Paid Search
2. Organic Search

3. Email
4. Display

(This list looks artificially short, but the logic and interpretation would remain the same for nine channels or for 99 channels.)

If we want to know that each channel is *simply somehow different*, then we could count the number of visits from each channel. Those counts can be considered *nominal in nature*.

Suppose the counts looked like this:

Channel	Count of Visits
Paid Search	2,143
Organic Search	3,124
Email	1,254
Display	2,077

With nominal data, the order of the four channels would not change or alter the interpretation. Suppose we, instead, viewed the data like this:

Channel	Count of Visits
Display	2,077
Paid Search	2,143
Email	1,254
Organic Search	3,124

The order of the categories does not matter.

And, the *distance between* the categories is not relevant. Display is not four times as much as paid search and organic search is not half of organic search. While there is an arithmetic relationship between these counts, that is only relevant if we treat the scales as *ratio scales* (see the Ratio Scales section below).

Finally, zero holds no meaning. We could not interpret a zero because it does not occur in a nominal scale.

Appropriate statistics for nominal scales: mode, count, frequencies

Displays: histograms or bar charts

Ordinal Scales

At the risk of providing a tautological definition, ordinal scales measure, well, order. So, our characteristics for ordinal scales are:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ordinal scales do not hold distance. The distance between first and second is unknown as is the distance between first and third along with all observations.
- **True Zero:** There is no true or real zero. An item, observation, or category cannot finish zero.

Let's work through our traffic source example and rank the channels based on the number of visits to our site, with "1" being the highest number of visits:

Channel	Count of Visits
Organic Search	1
Paid Search	2
Display	3
Email	4

Again, for this example, we are limiting ourselves to four channels, but the logic would remain the same for ranking nine channels or 99 channels.

By ranking the channel from most to least number of visitors in terms of last point of contact, we've established an *order*.

However, the distance between the rankings appears unknown. Organic Search could have one more visit compared to Paid Search or one hundred more visitors. The distance between the two items appears unknown.

Finally, zero holds no meaning. We could not interpret a zero because it does not occur in an ordinal scale. An item such as Organic Search could not maintain a zero ranking.

Appropriate statistics for ordinal scales: count, frequencies, mode

Displays: histograms or bar charts

Interval Scales

Interval scales provide insight into the variability of the observations or data. Classic interval scales are Likert scales (e.g., 1 - strongly agree and 9 - strongly disagree) and Semantic Differential scales (e.g., 1 - dark and 9 - light). In an interval scale, users could respond to "I enjoy opening links to the website from a company email" with a response ranging on a scale of values.

The characteristics of interval scales are:

- **Order:** The order of the responses or observations does matter.
- **Distance:** Interval scales do offer distance. That is, the distance from 1 to 2 appears the same as 4 to 5. Also, six is twice as much as three and two is half of four. Hence, we can perform arithmetic operations on the data.

- **True Zero:** There is no zero with interval scales. However, data can be rescaled in a manner that contains zero. An interval scales measure from 1 to 9 remains the same as 11 to 19 because we added 10 to all values. Similarly, a 1 to 9 interval scale is the same a -4 to 4 scale because we subtracted 5 from all values. Although the new scale contains zero, zero remains uninterpretable because it only appears in the scale from the transformation.

Unless a web analyst is working with survey data, it is doubtful he or she will encounter data from an interval scales. More likely, a web analyst will deal with ratio scales (next section).

Appropriate statistics for interval scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

An Illustrative Side Note About Temperature

An argument exists about temperature. Is it an interval scale or an ordinal scale? Many researchers argue for temperature as an interval scale. It offers order (e.g., 212◦◦ F is hotter than 32◦◦ F), distance (e.g., 40◦◦ F to 44◦◦ F is the same as 100◦◦ F to 104◦◦ F), and lacks a true zero (e.g., 0◦◦ F is not the same as 0◦◦ C). However, other researchers argue for temperature as an ordinal scale because of the issue related to distance. 200◦◦ F is not twice as 100 F. The human brain registers both temperatures as equally hot (if standing outside) or mild (if touching a stove). Finally, we would not say that 80 F is twice as warm as 40◦◦ F or that 30◦◦ F is a third colder as 90◦◦ F.

Ratio Scales

Ratio scales appear as nominal scales with a true zero. They have the following characteristics:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ratio scales do do have an interpretable distance.
- **True Zero:** There is a true zero.

Income is a classic example of a ratio scale:

- Order is established. We would all prefer $100 to $1!
- Zero dollars means we have no income (or, in accounting terms, our revenue exactly equals our expenses!)
- Distance is interpretable, in that $20 appears as twice $10 and $50 is half of a $100.

In web analytics, the number of visits and the number of goal completions serve as examples of ratio scales. A thousand visits is a third of 3,000 visits, while 400 goal completions are twice as many as 200 goal completions. Zero visitors or zero goal completions should be interpreted as just that: no visits or completed goals (uh-oh… did someone remove the page tag?!).

For the web analyst, the statistics for ratio scales are the same as for interval scales.

Appropriate statistics for ratio scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

An Important Note: *Don’t let the term “ratio” trip you up. Laypeople (aka, “non-statisticians”) are taught that ratios represent a relationship between two numbers. For instance, conversion rate is the “ratio” of orders to visits. But, as illustrated above, that is an overly narrow definition when it comes to statistics.*

Analysis:

1. It interprets data at a deeper level. It interprets the information and provide recommendations on actions.

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

3. 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.

4. 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.

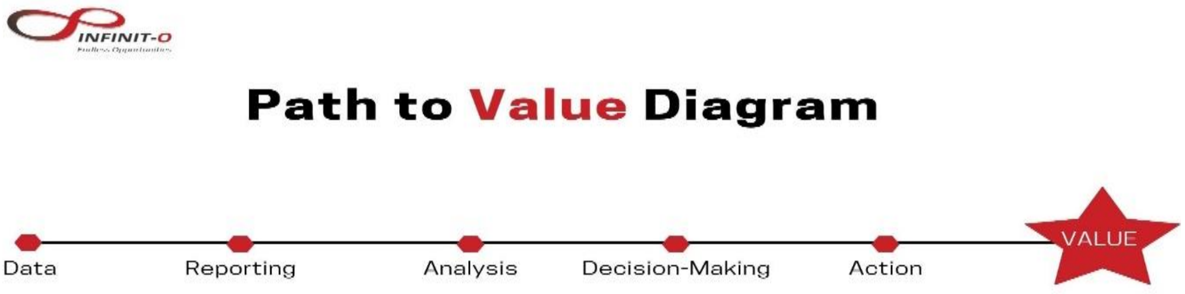
Reporting:

1. 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.

2. 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.

3. Reporting has a push approach, as it pushes information to users and outputs come in the forms of canned reports, dashboards, and alerts.

4. 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.


INFINIT-O
Endless Opportunities
Path to Value Diagram
Data
Reporting
Analysis
Decision-Making
Action
VALUE