ASSISNMENT QUESTIONS UNIT-I

Introduction to Big Data Platform:

1. What is Big Data, and where does it come from? How does it work? Discuss.

   [7M-R20-SET-4-July 2023][Remembering]
2. List and explain the characteristics of Big data. Illustrate by considering an example?

   [7M-R20-SET-1-July 2023] [Understanding]
3. Define data, Web data. Also explain structured, semi structured and unstructured data?

   [7M-R20-SET-2-July 2023] [Remembering]

Challenges of Conventional Systems:

4. What are the challenges of conventional systems in handling big data? Explain.

   [7M-R20-SET-1,3-July-2023] [Remembering]
5. Comparison of Big Data with Conventional Data and explain Three challenges of Conventional systems?**[Create]**

6. What are the Advantages and Disadvantages of Big Data? **][Remembering]**

Intelligent data analysis:

7. Discuss in detail about Intelligent Data Analysis?**[7M-R20-SET-2-April 2023] [Create]**

Nature of Data:

8. Explain in detail about Nature of Data and its applications?**[7M-R20-SET-3-April-2023] [Understanding]**
9. Explain about the data selection and data conversion in nature of data?**[Understanding]**

Analytic Processes and Tools:

10. Explain about Analytic Process and Tools in detailed?**[Understanding]**

Analysis vs Reporting:

11. Describe the Analysis Vs Reporting?**[7M-R20-SET-1-July 2023][Create]**
12. Explain about Big data and Analytic archicture. **[Create]**

UNIT-I

Introduction to big data: Introduction to Big Data Platform, Challenges of Conventional Systems, Intelligent data analysis, Nature of Data, Analytic Processes and Tools, Analysis vs Reporting.

What is big data analytics?

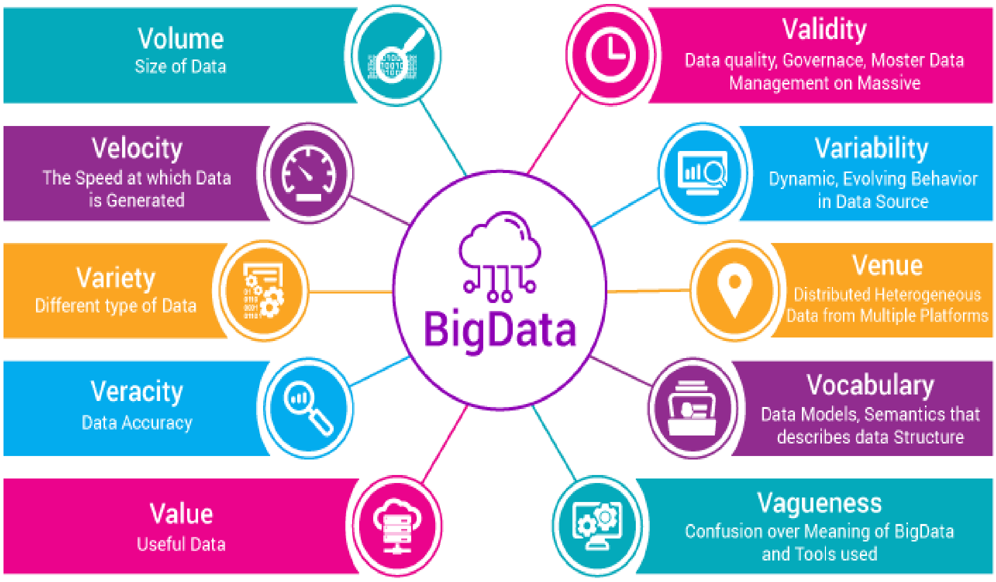
Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Big Data is a massive amount of data sets that cannot be stored, processed, or analyzed using traditional tools.

Data also exists in different formats, like structured data, semi-structured data, and unstructured data. For example, in a regular Excel sheet, data is classified as structured data—with a definite format. In contrast, emails fall under semi-structured, and your pictures and videos fall under unstructured data. All this data combined makes up Big Data.

Big data analytics is the process of collecting, examining, and analysing large amounts of data to discover market trends, insights, and patterns that can help companies make better business decisions.

This information is available quickly and efficiently so that companies can be agile in crafting plans to maintain their competitive advantage.


Volume
Size of Data
Validity
Data quality, Governace, Moster Data Management on Massive
Velocity
The Speed at which Data is Generated
Variability
Dynamic, Evolving Behavior in Data Source
Variety
Different type of Data
BigData
Venue
Distributed Heterogeneous Data from Multiple Platforms
Veracity
Data Accuracy
Vocabulary
Data Models, Semantics that describes data Structure
Value
Useful Data
Vagueness
Confusion over Meaning of BigData and Tools used

Big data analytics uses advanced analytics on large structured and unstructured data collections to produce valuable business insights. It is used widely across industries as varied as health care, education, insurance, artificial intelligence, retail, and manufacturing to understand what's working and what's not to improve processes, systems, and profitability.

The New York Stock Exchange is an example of Big Data that generates about one terabyte of new trade data per day.

The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

Where does it come from big data?

Big data comes from myriad sources -- some examples are transaction processing systems, customer databases, documents, emails, medical records, internet clickstream logs, mobile apps and social networks.

How Does Big Data Work?

Big data involves collecting, processing, and analyzing vast amounts of data from multiple sources to uncover patterns, relationships, and insights that can inform decision-making. The process involves several steps:

Data Collection

Big data is collected from various sources such as social media, sensors, transactional systems, customer reviews, and other sources.

Data Storage

The collected data then needs to be stored in a way that it can be easily accessed and analyzed later. This often requires specialized storage technologies capable of handling large volumes of data.

Data Processing

Once the data is stored, it needs to be processed before it can be analyzed. This involves cleaning and organizing the data to remove any errors or inconsistencies, and transform it into a format suitable for analysis.

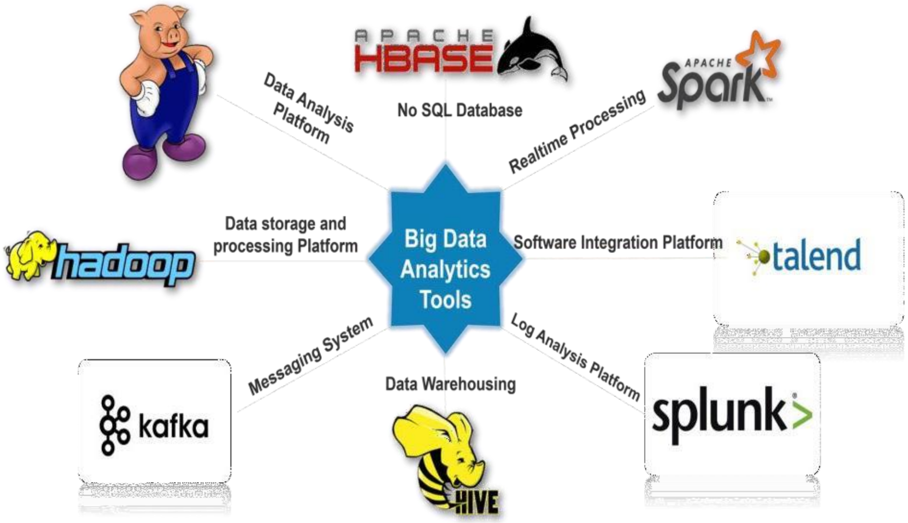
Data Analysis

After the data has been processed, it is time to analyze it using tools like statistical models and machine learning algorithms to identify patterns, relationships, and trends.

Data Visualization

The insights derived from data analysis are then presented in visual formats such as graphs, charts, and dashboards, making it easier for decision-makers to understand and act upon them.

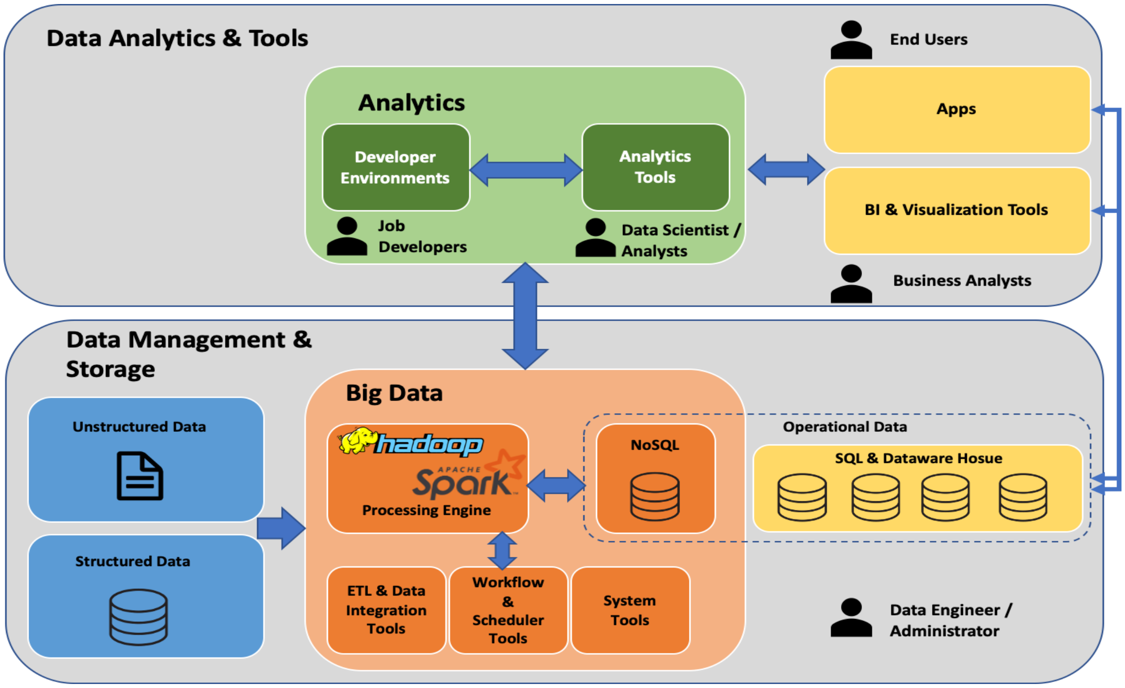
Big Data Analytics Tools



NoSQL databases (not just SQL) or non-relational are mostly used for collecting and analyzing big data. The data in a NoSQL database is handled for the dynamic organization of unstructured data versus relational databases' structured and tabular design.

Big data analysis requires a software framework for distributed storage and processing of big data. The following tools are considered software solutions for big data analytics:

Big Data and Analytics Architecture



In general terms in any BD&A architecture we need to consider the following:

Data Sources. Data sources are all the entities that are generating data such as third-party systems, machinery, sensors, social networks, among others.

ETL modules. ETL will allow us to collect the data from the different sources and in some way transform it into the appropriate formats to be used by the system to produce insightful information.

Processing engines. Processing engines such as Spark or Hadoop are required during different stages of data processing. First during the ETLing for cleaning, blending, and pre-processing the data. But also, during the actual data processing and analysis processing engines are needed.

Data Storage. Data needs to be stored in someplace, and depending on the nature of such data and the objective of the analysis, the storage needs to have certain characteristics, not only at the hardware level but also at Data Base Management System (DBMS) level. In most cases, a hybrid solution including SQL and NoSQL is needed.

Data Analytics platform. To produce insightful data some frameworks or tools to analyze the data will be needed in our solution. This includes data browsers, statistical analysis tools, and machine learning libraries.

Data Visualization framework. One of the most important stages is to present the analytics to the end-user. If data is not correctly presented, then it doesn’t matter if you are using the best algorithm or the best data processing technology, because the end-user is not going to understand it properly. Therefore, having a strong visualization framework is always very important.

Advantages and Disadvantages of Big Data

Advantages of Big Data

- Improved decision-making: Big data can provide insights and patterns that help organizations make more informed decisions.

- Increased efficiency: Big data analytics can help organizations identify inefficiencies in their operations and improve processes to reduce costs.

- Better customer targeting: By analyzing customer data, businesses can develop targeted marketing campaigns that are relevant to individual customers, resulting in better customer engagement and loyalty.

- New revenue streams: Big data can uncover new business opportunities, enabling organizations to create new products and services that meet market demand.

- Competitive advantage: Organizations that can effectively leverage big data have a competitive advantage over those that cannot, as they can make faster, more informed decisions based on data-driven insights.

Disadvantages of Big Data

Privacy concerns: Collecting and storing large amounts of data can raise privacy concerns, particularly if the data includes sensitive personal information.

- Risk of data breaches: Big data increases the risk of data breaches, leading to loss of confidential data and negative publicity for the organization.
- Technical challenges: Managing and processing large volumes of data requires specialized technologies and skilled personnel, which can be expensive and time-consuming.
- Difficulty in integrating data sources: Integrating data from multiple sources can be challenging, particularly if the data is unstructured or stored in different formats.
- Complexity of analysis: Analyzing large datasets can be complex and time-consuming, requiring specialized skills and expertise.

Big Data Characteristics:

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many multinational companies to process the data and business of many organizations. The data flow would exceed 150 exabytes per day before replication.

There are five v's of Big Data that explains the characteristics.

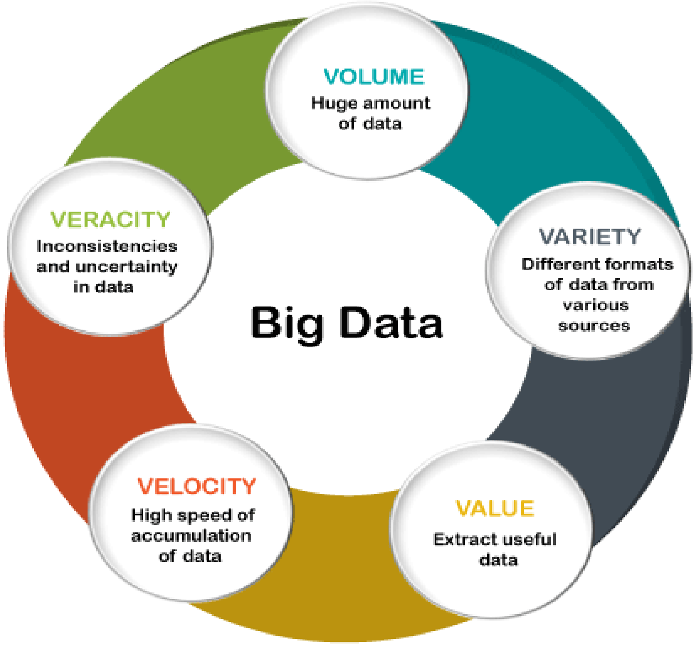
5 V's of Big Data

- **Volume**
- **Veracity**
- **Variety**
- **Value**
- **Velocity**

Volume

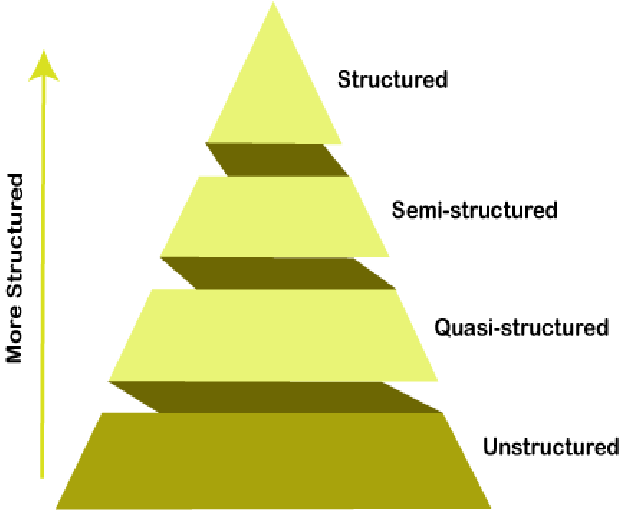
The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more.

Facebook can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.


VOLUME
Huge amount of data
VARIETY
Different formats of data from various sources
VALUE
Extract useful data
VELOCITY
High speed of accumulation of data
VERACITY
Inconsistencies and uncertainty in data
Big Data

Variety

Big Data can be structured, unstructured, and semi-structured that are being collected from different sources. Data will only be collected from databases and sheets in the past, But these days the data will comes in array forms, that are PDFs, Emails, audios, SM posts, photos, videos, etc.



The data is categorized as below:

Structured data: In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

Semi-structured: In Semi-structured, the schema is not appropriately defined, e.g., JSON, XML, CSV, TSV, and email. OLTP (Online Transaction Processing) systems are built to work with semi-structured data. It is stored in relations, i.e., tables.

Unstructured Data: All the unstructured files, log files, audio files, and image files are included in the unstructured data. Some organizations have much data available, but they did not know how to derive the value of data since the data is raw.

Quasi-structured Data: The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

Example: Web server logs, i.e., the log file is created and maintained by some server that contains a list of activities.

Veracity

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development.

For example, **Facebook posts** with hashtags.

Value

Value is an essential characteristic of big data. It is not the data that we process or store. It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.

Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in real-time. It contains the linking of incoming data sets speeds, rate of change, and activity bursts. The primary aspect of Big Data is to provide demanding data rapidly.

Types Of Big Data

Following are the types of Big Data:

Structured
Unstructured
Semi-structured

What is Structured Data?

- Any data that can be stored, accessed and processed in the form of fixed format istermed as a 'structured' data.
- Developed techniques for working with such kind of data (where the format is wellknown in advance) and also deriving value out of it.
- Foreseeing issues of today :
    - when a size of such data grows to a huge extent, typical sizes are being in the rageof multiple zetta bytes.
- Do you know?
- ***10^{21} bytes*** equal to ***1 zettabyte*** or ***one billion terabytes*** forms ***a zettabyte***.
    - That is why the name Big Data is given and imagine the challenges involved inits storage and processing?
- Do you know?
- Data stored in a relational database management system is one example ofa **'structured'** data

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

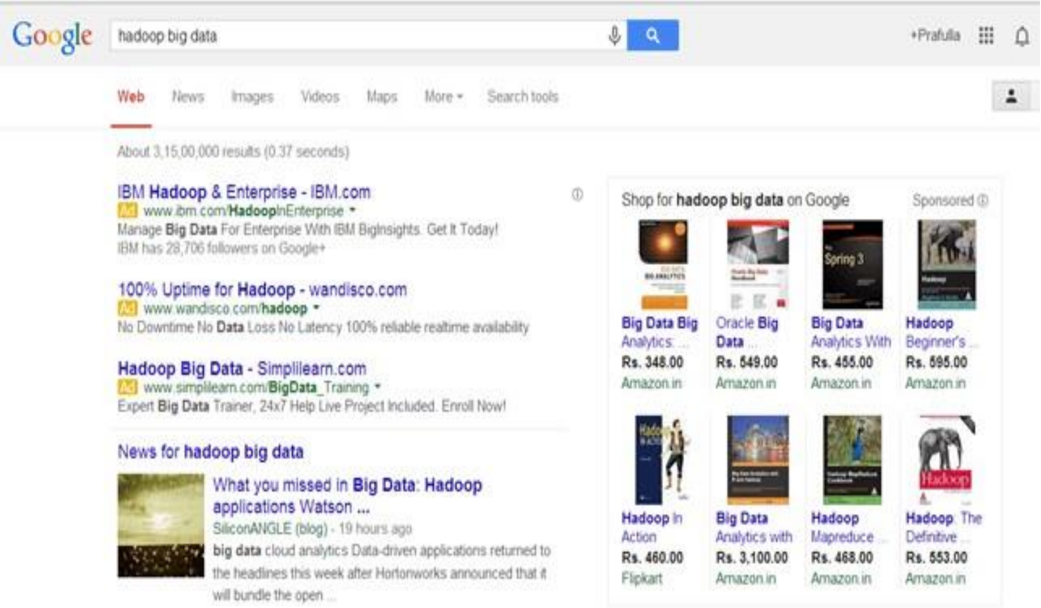
Employee_ID	**Employee_Name**	**Gender**	**Department**	**Salary_In_lacs**
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured Data

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

Example of Unstructured data

- The output returned by 'Google Search'



hadoop big data

About 3,15,00,000 results (0.37 seconds)

IBM Hadoop & Enterprise - IBM.com
Ad www.ibm.com/HadoopInEnterprise
Manage Big Data For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
Ad www.wandisco.com/hadoop
No Downtime No Data Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/BigData_Training
Expert Big Data Trainer, 24x7 Help Live Project Included. Enroll Now!

News for hadoop big data

What you missed in Big Data: Hadoop applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for hadoop big data on Google Sponsored

Big Data Big Analytics: ...
Rs. 348.00
Amazon.in

Oracle Big Data ...
Rs. 549.00
Amazon.in

Big Data Analytics With
Rs. 455.00
Amazon.in

Hadoop Beginner's ...
Rs. 595.00
Amazon.in

Hadoop In Action
Rs. 460.00
Flipkart

Big Data Analytics with
Rs. 3,100.00
Amazon.in

Hadoop Mapreduce ...
Rs. 468.00
Amazon.in

Hadoop: The Definitive ...
Rs. 553.00
Amazon.in

Example Of Un-structured Data

Semi-structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g.

a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

Examples Of Semi-structured Data

- Personal data stored in an XML file.

```
<rec>
<name>Prashant Rao</name>
<sex>Male</sex>
<age>35</age>
</rec>
<rec>
<name>Seema R.</name>
<sex>Female</sex>
<age>41</age>


</rec>
<rec>
<name>Satish Mane</name>
<sex>Male</sex>
<age>29</age>
```

```
</rec>
<rec>
<name>Subrato Roy</name>
<sex>Male</sex>
<age>26</age>
</rec>
<rec>
<name>Jeremiah J.</name>
<sex>Male</sex>
<age>35</age></rec>
```

Basics of Bigdata Platform

•Big Data platform is IT solution which combines several Big Data tools and utilities into one packaged solution for managing and analyzing Big Data.

•Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.

•It is an enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure /environment.

What is Big Data Platform?

- Big Data Platform is integrated IT solution for Big Data management which combines several software system, software tools and hardware to provide easy to use tools systemto enterprises.
- It is a single one-stop solution for all Big Data needs of an enterprise irrespective of size and data volume. Big Data Platform is enterprise class IT solution for developing, deploying and managing Big Data.
- There are several Open source and commercial Big Data Platform in the market with varied features which can be used in Big Data environment.
- Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.
- It is an enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure /environment.
- Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities
- It also supports customdevelopment, querying and integration with other systems.

- ➢ The primary benefit behind a big data platform is to reduce the complexity of multiplevendors/ solutions into a one cohesive solution.
- ➢ Big data platform are also delivered through cloud where the provider provides an allinclusive big data solutions and services.

Features of Big Data Platform

Here are most important features of any good Big Data Analytics Platform:

- ➢ Big Data platform should be able to accommodate new platforms and tool based on the business requirement. Because business needs can change due to new technologies or due to change in business process.
- ➢ It should support linear scale-out
- ➢ It should have capability for rapid deployment
- ➢ It should support variety of data format
- ➢ Platform should provide data analysis and reporting tools
- ➢ It should provide real-time data analysis software
- ➢ It should have tools for searching the data through large data sets

Big data is a term for data sets that are so large or complex that traditional data processingapplications are inadequate.

Challenges include

- **Analysis,**
- **Capture,**
- **Data Curation,**
- **Search,**
- **Sharing,**
- **Storage,**
- **Transfer,**
- **Visualization,**
- **Querying,**
- **Updating**

List of BigData Platforms

a) **Hadoop**
b) Cloudera
c) **Amazon Web Services**
d) Hortonworks
e) **MapR**

f) ***IBM Open Platform***
g) **Microsoft HDInsight**
h) ***Intel Distribution for Apache Hadoop***
i) **Datastax Enterprise Analytics**
j) ***Teradata Enterprise Access for Hadoop***
k) **Pivotal HD**

a) ***Hadoop***

What is Hadoop?

- Hadoop is open-source, Java based programming framework and server software which is used to save and analyze data with the help of 100s or even 1000s of commodity servers in a clustered environment.
- Hadoop is designed to storage and process large datasets extremely fast and in fault tolerant way.
- Hadoop uses HDFS (Hadoop File System) for storing data on cluster of commodity computers. If any server goes down it know how to replicate the data and there is no lossof data even in hardware failure.
- Hadoop is Apache sponsored project and it consists of many software packages whichruns on the top of the Apache Hadoop system.
- Top Hadoop based Commercial Big Data Analytics Platform
- Hadoop provides set of tools and software for making the backbone of the Big Data analytics system.
- Hadoop ecosystem provides necessary tools and software for handling and analyzing Big Data.
- On the top of the Hadoop system many applications can be developed and plugged-in toprovide ideal solution for Big Data needs.

b) ***Cloudera***

Cloudra is one of the first commercial Hadoop based Big Data Analytics Platform offering Big Data solution.

Its product range includes Cloudera Analytic DB, Cloudera Operational DB, ClouderaData Science & Engineering and Cloudera Essentials.

All these products are based on the Apache Hadoop and provides real-time processingand analytics of massive data sets.

Independent scaling of storage from expensive compute resources

c) ***Amazon Web Services***

Amazon is offering Hadoop environment in cloud as part of its Amazon Web Servicespackage.

AWS Hadoop solution is hosted solution which runs on Amazon's Elastic Cloud Compute and Simple Storage Service (S3).

Enterprises can use the Amazon AWS to run their Big Data processing analytics in thecloud environment.

Amazon EMR allows companies to setup and easily scale Apache Hadoop, Spark, HBase, Presto, Hive, and other Big Data Frameworks using its cloud hosting environment.

d) Hortonworks

Hortonworks is using 100% open-source software without any propriety software. Hortonworks were the one who first integrated support for Apache HCatalog.

The Hortonworks is a Big Data company based in California.

This company is developing and supports application for Apache Hadoop.

Hortonworks Hadoop distribution is 100% open source and its enterprise ready

with followingfeatures:

- Centralized management and configuration of clusters
- Security and data governance are built in feature of the system

Centralized security administration across the systemWebsite: https://hortonworks.com/

e) MapR

MapR is another Big Data platform which us using the Unix file system for handlingdata.

It is not using HDFS and this system is easy to learn anyone familiar with the Unix system.

This solution integrates Hadoop, Spark, and Apache Drill with a real-time dataprocessing feature.

Website: https://mapr.com

f) IBM Open Platform

IBM also offers Big Data Platform which is based on the Hadoop eco-system software.

IBM well knows company in software and data computing.

It uses the latest Hadoop software and provides following features (IBM Open PlatformFeatures):

Based on 100% Open source software

Native support for rolling Hadoop upgrades

Support for long running applications within YEARN.

Support for heterogeneous storage which includes HDFS for in-memory and SSD in addition to HDD

Native support for Spark, developers can use Java, Python and Scala to written program

Platform includes Ambari, which is a best tool for provisioning, managing & monitoringApache Hadoop clusters

IBM Open Platform includes all the software of Hadoop ecosystem e.g. HDFS, YARN, MapReduce, Ambari, Hbase, Hive, Oozie, Parquet, Parquet Format, Pig, Snappy, Solr, Spark, Sqoop, Zookeeper, Open JDK, Knox, Slider

Developer can download the trial Docker Image or Native installer for testing and learning the system

Application is well supported by IBM technology team

Website: https://www.ibm.com/analytics/us/en/technology/hadoop/

g) Microsoft HDInsight

The Microsoft HDInsight is also based on the Hadoop distribution and it's a commercial Big Data platform from Microsoft.

Microsoft is software giant which is into development of windows operating system for Desktop users and Server users.

This is the big Hadoop distribution offering which runs on the Windows and Azure environment.

It offer customized, optimized open source Hadoop based analytics clusters which uses Spark, Hive, MapReduce, HBase, Strom, Kafka and R Server which runs on the Hadoop system on windows/Azure environment

h) Distribution for Apache Hadoop

- Intel also offers its package distribution of Hadoop software which includes company'sGraph builder and Analytics toolkit.
- This distribution can be purchased with various channel partners and come with supportand yearly subscription.

i) Datastax Enterprise Analytics

- Datastax Enterprise Analytics is another play in the Big Data Analytics platform which offers its own distribution which is based on Apache Cassandra database management system which runs on the top of Apache Hadoop installation.

- It also included propriety system with a dashboard which is used for security management, searching data, dashboard for viewing various details and visualization engine.

j) Teradata Enterprise Access for Hadoop

- Teradata Enterprise Access for Hadoop is another player into Big Data Platform and it offers package Hadoop distribution which again based on Hortonworks distribution.

- Teradata Enterprise Access for Hadoop offers Hardware and software in its Big Data solution which can be used by enterprise to process its data sets.

Open Source Big Data Platform

There are various open-source Big Data Platform which can be used for Big Data handling and data analytics in real-time environment.

Both small and Big Enterprise can use these tools for managing their enterprise data for getting best value from their enterprise data.

Apache Hadoop

Apache Hadoop is Big Data platform and software package which is Apache sponsoredproject.

Under Apache Hadoop project various other software is being developed which runs onthe top of Hadoop system to provide enterprise grade data management and analytics solutions to enterprise.

Apache Hadoop is open-source, distributed file system which provides data processing and analysis engine for analyzing large set of data.

Hadoop can run on Windows, Linux and OS X operating systems, but it is mostly used on Ubunut and other Linux variants.

MapReduce

The MapReduce engine was originally written by Google and this is the system which enables the developers to write program which can run in parallel on 100 or even 1000s of computer nodes to process vast data sets.

After processing all the job on the different nodes it comes the results and return it to the program which executed the MapReduce job.

This software is platform independent and runs on the top of Hadoop ecosystem. It can process tremendous data at very high speed in Big Data environment.

Grid Gain

GridGain is another software system for parallel processing of data just like MapRedue. GridGain is an alternative of Apache MapReduce.

CHALLENGES OF CONVENTIONAL SYSTEMS

Introduction to Conventional Systems

What is Conventional System?

The **system** consists of one or more zones each having either manually operated call points or automatic detection devices, or a combination of both.

Big data is **huge** amount of **data** which is beyond the processing capacity of **conventional data** base **systems** to manage and analyze the **data** in a specific timeinterval.

Difference between conventional computing and intelligent computing

The conventional computing functions logically with a set of rules and calculations while the neural computing can function via images, pictures, and concepts.

Conventional computing is often unable to manage the variability of data obtained in the real world.

On the other hand, neural computing, like our own brains, is well suited to situations that have no clear algorithmic solutions and are able to manage noisy imprecise data. This allows them to excel in those areas that conventional computing often finds difficult.

Comparison of Big Data with Conventional Data

Big Data	Conventional Data
Huge data sets	Data set size in control.
Unstructured data such as text, video, and audio.	Normally structured data such as numbersand categories, but it can take other formsas well.
Hard-to-perform queries and analysis	Relatively easy-to-perform queries and analysis.

Needs a new methodology for analysis.	Data analysis can be achieved by using Conventional methods.
Need tools such as Hadoop, Hive, Hbase, Pig, Sqoop, and so on.	Tools such as SQL, SAS, R, and Excel alone may be sufficient.
The aggregated or sampled or filtered data.	Raw transactional data.
Used for reporting, basic analysis, and text mining. Advanced analytics is only ina starting stage in big data.	Used for reporting, advanced analysis, and predictive modeling .
Big data analysis needs both programming skills (such as Java) and analytical skills to perform analysis.	Analytical skills are sufficient for conventional data; advanced analysis tools don't require expert programing skills.
Petabytes/exabytes of data.	Millions/billions of accounts.
Billions/trillions of transactions.	Megabytes/gigabytes of data.
Thousands/millions of accounts.	Millions of transactions
Generated by big financial institutions, Facebook, Google, Amazon, eBay, Walmart, and so on.	Generated by small enterprises and small banks.

List of challenges of Conventional Systems

The following list of challenges has been dominating in the case Conventional systems in realtime scenarios:

1) ***Uncertainty of Data Management Landscape***
2) **The Big Data Talent Gap**
3) ***The talent gap that exists in the industry Getting data into the big data platform***
4) **Need for synchronization across data sources**
5) ***Getting important insights through the use of Big data analytics***

1) Uncertainty of Data Management Landscape:

- Because big data is continuously expanding, there are new companies and technologiesthat are being developed everyday.
- A big challenge for companies is to find out which technology works bests for them without the introduction of new risks and problems.

2) *The Big Data Talent Gap:*

- While Big Data is a growing field, there are very few experts available in this field.

- This is because Big data is a complex field and people who understand the complexity and intricate nature of this field are far few and between.

3) ***The talent gap that exists in the industry Getting data into the big data platform:***

- Data is increasing every single day. This means that companies have to tackle limitless amount of data on a regular basis.
- The scale and variety of data that is available today can overwhelm any data practitioner and that is why it is important to make data accessibility simple and convenient for brand mangers and owners.

4) ***Need for synchronization across data sources:***

- As data sets become more diverse, there is a need to incorporate them into an analyticalplatform.
- If this is ignored, it can create gaps and lead to wrong insights and messages.

5) ***Getting important insights through the use of Big data analytics:***

- It is important that companies gain proper insights from big data analytics and it is important that the correct department has access to this information.
- A major challenge in the big data analytics is bridging this gap in an effective fashion.

Other Three challenges of Conventional systems

Three Challenges That big data face.

1. Data
2. Process
3. Management

1. Data Challenges

Volume

1. The volume of data, especially machine-generated data, is exploding,

2. how fast that data is growing every year, withnew sources of data that are emerging.

3. For example, in the year 2000, 800,000petabytes (PB) of data were stored in the world, and itis expected to reach 35 zetta bytes (ZB) by2020 (according to IBM).

Social media plays a key role: Twitter generates 7+ terabytes (TB) of data every day. Facebook,10 TB.

• Mobile devices play a key role as well, as there were estimated 6 billion mobile phones in 2011.

• The challenge is how to deal with the size of Big Data.

Variety,CombiningMultipleDataSets

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

• Today, companies are looking to leverage a lot more•data from a wider variety of sources both inside and outside the organization.

• Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.

Variety•A lot of this data is unstructured, or has a complex structure that’s hard to represent in rows and columns.

2. Processing

- More than 80% of today’s information isunstructured and it is typically too big to manage effectively.
- Today, companies are looking to leverage a lot more data from a wider variety of sources both inside and outside the organization.
- Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.

3. Management

- A lot of this data is unstructured, or has acomplex structure that’s hard to represent in rows and columns.

Big Data Challenges

– The challenges include capture, duration, storage, search, sharing, transfer, analysis, and visualization.

Big Data is trend to larger data sets due to the additional information derivable from analysis of a single large set of related data,as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to

"spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway trafficconditions.”

Challenges of Big Data

The following are the five most important challenges of the Big Data

a) **Meeting the need for speed**

In today's hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly.

b) Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.

c) **The challenge only *grows as the degree of granularity increases.*** One possible solution is hardware. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly

d) **Understanding the data**

- It takes a lot of understanding to get data in the **RIGHT SHAPE** so that you can use
- visualization as part of data analysis.

e) **Addressing data quality**

- Even if you can find and analyze data quickly and put it in the proper context for the
- audience that will be consuming the information, the value of data for **DECISION-MAKING PURPOSES** will be jeopardized if the data is not accurate or timely.

e) **Displaying meaningful results**

Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information.

For example, imagine you have 10 billion rows of retail SKU data that you're trying to compare. The user trying to view 10 billion plots on the screen will have a hard time

Seeing so many data points. Grouping the data together, or "binning," you can more effectively visualize thedata.

f) **Dealing with outliers**

The graphical representations of data made possible by visualization can communicate trends and outliers much faster than tables containing numbers and text.

Users can easily spot issues that need attention simply by glancing at a chart. Outliers typically represent about 1 to 5 percent of data, but when you're working with massive amounts of data, viewing 1 to 5 percent of the data is rather difficult

We can also bin the results to both view the distribution of data and see the outliers. While outliers may not be representative of the data, they may also reveal previously unseen and potentially valuable i

INTELLIGENT DATA ANALYSIS

Intelligent Data Analysis (IDA) is one of the hot issues in the field of artificial intelligence and information

IDA is an interdisciplinary study concerned with the effective analysis of data;

- used for extracting useful information from large quantities of online data; extracting desirable knowledge or interesting patterns from existing databases;
- The distillation of information that has been collected, classified, organized, integrated, abstracted and value-added;
- At a level of abstraction higher than the data, and information on which it is based and can be used to deduce new information and new knowledge;
- Usually in the context of human expertise used in solving problems.
- the distillation of information that has been collected, classified, organized, integrated, abstracted and value-added;
- at a level of abstraction higher than the data, and information on which it is based and can be used to deduce new information and new knowledge;

1,3,2 Uses / Benefits of IDA

Intelligent Data Analysis provides a forum for the examination of issues related to the research and applications of Artificial Intelligence techniques in data analysis across a variety of disciplines and the techniques include (but are not limited to):

The benefit areas are:

- **Data Visualization**
- **Data pre-processing (fusion, editing, transformation, filtering, sampling)**
- **Data Engineering**
- **Database mining techniques, tools and applications**
- **Use of domain knowledge in data analysis**
- **Big Data applications**
- **Evolutionary algorithms**
- **Machine Learning(ML)**
- **Neural nets**
- **Fuzzy logic**
- **Statistical pattern recognition**
- **Knowledge Filte**

Intelligent Data Analysis

Knowledge Acquisition

The process of eliciting, analyzing, transforming, classifying, organizing and integrating knowledge and representing that knowledge in a form that can be used in a computer system.

Intelligent Data Examples:

- Epidemiological study (1970-1990)
- Sample of examinees died from cardiovascular diseases during the period

Question: Did they know they were ill?

- they were healthy
- they were ill (drug treatment, positive clinical and laboratory findings)

Illustration of IDA by using See5

- **application.names** - lists the classes to which cases may belong and the attributes used to describe each case.
- **Attributes are of two types:** discrete attributes have a value drawn from a set of possibilities, and continuous attributes have numeric values.
- **application.*data*** - provides information on the *training* cases from which See5 will extract patterns.
  The entry for each case consists of one or more lines that give the values for all attributes.
- **application.*test*** - provides information on the *test* cases (used for evaluation of results).
  The entry for each case consists of one or more lines that give the values for all attributes.

Goal 1.1 : ***application.names – example***

gender:M,

F activity:1,2,3

age: continuous

smoking: No, Yes

…

Goal:1,2 **:**

application.*data* – example

M,1,59,Yes,0,0,0,0,119,73,103,86,247,87,15979,?,?,?,1,73,2.5

M,1,66,Yes,0,0,0,0,132,81,183,239,?,783,14403,27221,19153,23187,1,73,2.6

M,1,61,No,0,0,0,0,130,79,148,86,209,115,21719,12324,10593,11458,1,74,2.5

… …

Result:

Results – example

Rule 1: (cover 26)

gender = MSBP > 111

oil_fat > 2.9

-> class 1 [0.929]

Rule 1: (cover 26)

gender = MSBP > 111

oil_fat > 2.9

-> class 1 [0.929]

Rule 4: (cover 14)

smoking = YesSBP > 131

glucose > 93

glucose <= 118

oil_fat <= 2.9

-> class 2 [0.938]

Rule 15: (cover 2)

SBP <= 111

oil_fat > 2.9

-> class 2 [0.750]

Evaluation on training

data (199 cases):

- *(b) <-classified as*

```
 ---- ----

  107    3    (a): class 1
   17   72    (b): class 2
```

Results on (training set):

Sensitivity=0.97

Specificity=0.81

Sensitivity=0.97
Specificity=0.81

Sensitivity=0.98
Specificity=0.90

Evaluation of IDA results

- ***Absolute & relative accuracy***
- **Sensitivity & specificity**
- ***False positive & false negative***
- **Error rate**
- ***Reliability of rules***
- **Etc.**

1. 4 NATURE OF DATA

Data

- **Data** is a set of values of qualitative or quantitative variables; restated, pieces of **data** are individual pieces of information.
- **Data** is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images.

Properties of Data

For examining the properties of data, reference to the various definitions of data. Reference to these definitions reveals that following are the properties of data:

a) **Amenability of use**
b) **Clarity**
c) **Accuracy**
d) **Essence**
e) **Aggregation**
f) **Compression**
g) **Refinement**

.a) **Amenability of use:** From the dictionary meaning of data it is learnt that data are facts used in deciding something. In short, data are meant to be used as a base for arriving atdefinitive conclusions.

a) **Clarity:** Data are a crystallized presentation. Without clarity, the meaning desired to be communicated will remain hidden.

b) **Accuracy:** Data should be real, complete and accurate. Accuracy is thus, an essential property of data.

c) **Essence:** A large quantities of data are collected and they have to be Compressed and refined. Data so refined can present the essence or derived qualitative value, of the matter.

d) **Aggregation:** Aggregation is cumulating or adding up.

e) **Compression:** Large amounts of data are always compressed to make them more meaningful. Compress data to a manageable size.Graphs and charts are some examplesof compressed data.

f) **Refinement:** Data require processing or refinement. When refined, they are capable of leading to conclusions or even generalizations. Conclusions can be drawn only when data are processed or refined.

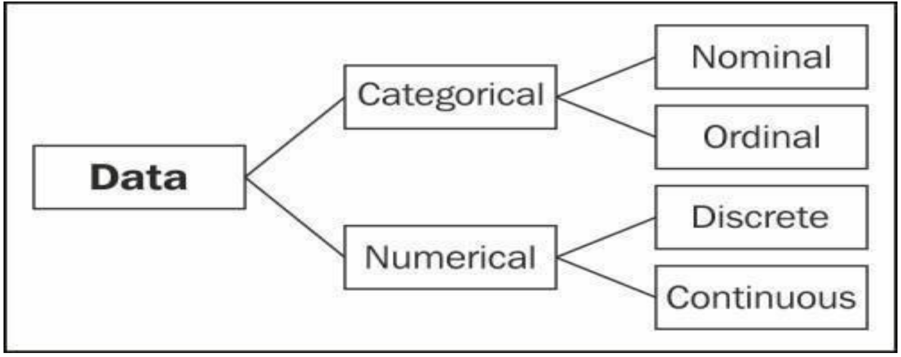
TYPES OF DATA

- In order to understand the nature of data it is necessary to categorize them into various types. Different categorizations of data are possible.
- The first such categorization may be on the basis of disciplines, e.g., Sciences, Social Sciences, etc. in which they are generated.
- Within each of these fields, there may be several ways in which data can be categorized intotypes.

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 beperformed.



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 uninterruptable.

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.

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 uninterruptable because it only appears in the scale from the transformation.

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

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

Ratio Scales

Ratio scales appear as nominal scales with 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 exactlyequals our expenses!)

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

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

The table below summarizes the characteristics of all four types of scales.

	Nominal	Ordinal	Interval	Ratio
Order Matters	No	Yes	Yes	Yes
Distance Is Interpretable	No	No	Yes	Yes
Zero Exists	No	No	No	Yes

DATA CONVERSION

- We can convert or transform our data from **ratio** to **interval** to **ordinal** to **nominal**. However, we *cannot* convert or transform our data from **nominal** to **ordinal** to **interva**l to **ratio.**

Scaled data can be measured in **exact amounts.**

For example, 60 degrees , 12.5 feet, 80 Miles per hour

Scaled data can be measured **with equal intervals.**

For example, Between **0** and **1** is **1 inch,** Between **13** and **14** is also **1 inch**

Ordinal or ranked data *provides*

comparative AmountsExample:

1st Place *2nd Place* *3rd Place*

- ***Not equal intervals***

1st Place	**2nd Place**	**3rd Place**
19.6 feet	**18.2 feet**	**12.4 feet**

DATA SELECTION

Another Example that handle the question as :

What is the average driving **speed** of teenagers on the freeway?

a) Scaled
b) Ordinal

Scaled – Speed:- ***Speed*** *can be measured in* ***exact amounts with equal intervals.***

Example :

60 degrees **12.5 feet** **80 Miles per hour**

- ***Ordinal or ranked data*** *provides* ***comparative amounts.***

For example, ***1st Place*** ***2nd Place*** ***3rd Place***

- ***Percentiles*** *provide* ***comparative amounts.***

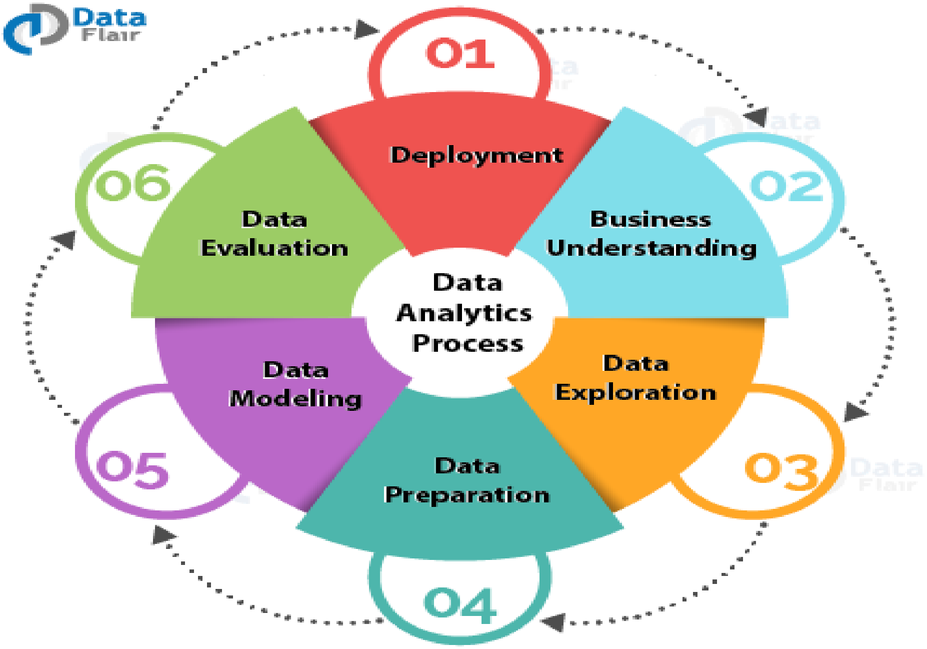
In this case, 93% of all hospital have lower patient satisfaction scores than Eastridge hospital.31% have lower satisfaction scores than Westridge Hospital

ANALYTIC PROCESS AND TOOLS

Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions.

There are 6 analytic processes:

1. **Deployment**
2. **Business Understanding**
3. **Data Exploration**
4. **Data Preparation**
5. **Data Modeling**
6. **Data Evaluation**



Step 1: Deployment

Big data deployment involves distributed computing, multiple clusters, networks, and firewalls.

Plan the deployment and monitoring and maintenance, we need to produce a final report and review the project.

– In this phase,

- ➢ We deploy the results of the analysis.
- ➢ This is also known as reviewing the project**.**

Step 2: Business Understanding

The very first step consists of business understanding.

Business processes generate large volumes of data.

Whenever any requirement occurs, firstly we need to determine the business objective, assess the situation, determine data mining goals and then produce the project plan as per the requirement.

Business objectives are defined in this phase.

Step 3: Data Exploration

The second step consists of Data understanding.

For the further process, we need to gather initial data, describe and explore the data and verify data quality to ensure it contains the data we *require*.

Data collected from the various sources is described in terms of its application and the need for the project in this phase. This is also known as data exploration.

This is necessary to verify the quality of data collected.

Step 4: Data Preparation

From the data collected in the last step,

we need to select data as per the need, clean it, construct it to get useful information and then integrate it all.

Finally, we need to format the data to get the appropriate data.Data is selected, cleaned, and integrated into the format finalized for the analysis in thisphase.

Step 5: Data Modeling

We need to select a modeling technique, generate test design, build a model and assess the model built.

The data model is build to analyze relationships between various selected objects in the data,

Test cases are built for assessing the model and model is tested and implemented on the data in this phase.

- Where processing is hosted?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is stored?
  - Distributed Storage (e.g. Amazon S3)
- What is the programming model?
  - Distributed Processing (e.g. MapReduce)
- How data is stored & indexed?
  - High-performance schema-free databases (e.g. MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing

Step 6:.Data Evaluation

The data evaluation process determines whether data is usable for calculating risk estimates. Data that is unusable for calculating the risk estimates still may provide useful information for determining the distribution.A distribution describes the probability or likelihood of any potential value.

Some examples include administrative records, surveillance systems, or surveys. There are many types and sources of data. One isn't necessarily better than the other, and you can pick and choose indicators, data, and trends that are most relevant for your program and evaluation questions.

ANALYSIS AND REPORTING

What is Analysis?

The process of exploring data and reports in order **to extract meaningful insights.**

which can be **used to better understand and improve business performance.**

What is Reporting?

Reporting is "the process of organizing data into informational summaries in order to monitor how different areas of a business are performing."

COMPARING ANALYSIS WITH REPORTING

Reporting is "the process of organizing data into informational summaries in order to monitor how different areas of a business are performing."

Measuring core metrics and presenting them — whether in an email, a slidedeck, or online dashboard — falls under this category.

Analytics is "the process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance."

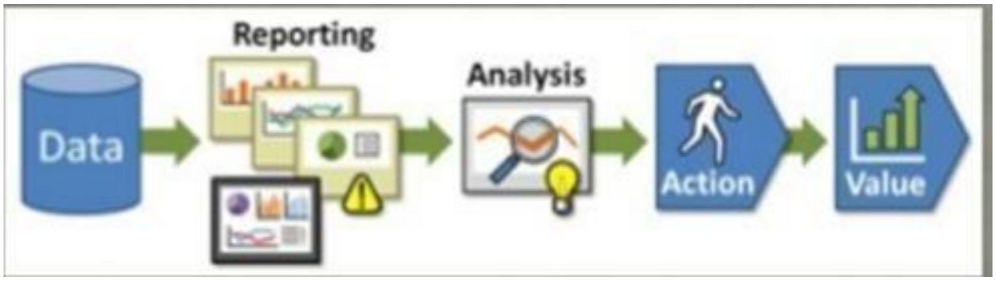
Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges.

Good reporting

Should **raise questions** about the business from its end users.The goal of analysis is to **answer questions** by interpreting the data at a deeper level and providingactionable recommendations.

A firm may be focused on the general area of analytics (strategy, implementation, reporting, etc.).but not necessarily on the specific aspect of analysis.

It's almost like some organizations run out of gas after the initial set-up-related activities and don't make it to the analysis stage



CONTRAST BETWEEN ANALYSIS AND REPORTING

The basis differences between Analysis and Reporting are as follows:

Analysis	Reporting
Provides what is needed	Provides what is asked for
Is typically customized	Is Typically standardized
Involves a person	Does not involve a person
Dig deeper: Analytics goes beyond summaries, uncovering hidden insights	Summarize data: Reporting organizes data neatly, making it easy to understand.
Transforms the information into insights & recommendations.	Transforms your data into information
Is extremely flexible	Is fairly Inflexible

- Reporting translates raw data into **information**.
- Analysis transforms data and information into **insights**.
- reporting shows you ***what is happening***
- while analysis focuses on explaining ***why it is happening*** and ***what you can do about it.***

- Reports are like Robots n monitor and alter you and where as analysis is like parents - c an figure out what is going on (hungry, dirty diaper, no pacifier, , teething, tired, ear infection, etc).
- Reporting and analysis can go hand-in-hand:
- Reporting provides no limited context about what is happening in the data. Context is critical to good analysis.
- Reporting translate a raw data into information
- Reporting usually raises a question – **What is happening ?**

\Analysis transforms the data into insights - **Why is it happening ? What you can doabout it?**