DIGITAL NOTES ON
# BUSINESS ANALYTICS BASICS
B.TECH III YEAR – II SEM

(2023-2024)

DEPARTMENT OF AERONAUTICAL ENGINEERING

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(Autonomous Institution – UGC, Govt. of India)

Recognized under 2(f) and 12 (B) of UGC ACT 1956

(Affiliated to JNTUH, Hyderabad, Approved by AICTE - Accredited by NBA & NAAC – 'A' Grade - ISO 9001:2015 Certified)

Maisammaguda, Dhulapally (Post Via.   Hakimpet), Secunderabad – 500100, Telangana State, India

MALLA REDDY COLLEGE OF ENGINEERING AND TECHNOLOGY

B.Tech -III Year-II Sem (ANE) **L/T/P/C**

3/-/-/3

BUSINESS ANALYTICS BASICS

COURSE OBJECTIVES

To help students in under standing how the managersuse business analytics for managerial decision making.

LearningOutcome/s:

The students will be familiar with the practices of analyzing and reporting the business data useful for the insights of business growth and development.

Unit-I:Understanding Business Analytics

Introduction: Meaning of Analytics - Evolution of Analytics - Need of Analytics - BusinessAnalysis vs. Business Analytics - Categorization of Analytical Models - Data Scientist vs. DataEngineervs. BusinessAnalyst-BusinessAnalytics inPractice-Types ofData-RoleofBusinessAnalyst.

Unit-II:Dealing with Data and DataScience

Data: Data Collection - Data Management - Big Data Management - Organization/Sources ofData-ImportanceofDataQuality-DealingwithMissingorIncompleteData-DataVisualization-Data Classification.

DataScienceProjectLifeCycle:BusinessRequirement-DataAcquisition-DataPreparation - Hypothesis and Modeling - Evaluation and Interpretation - Deployment - Operations - Optimization-Applications forDataScience

Unit-III:Data Mining and Machine Learning

Data Mining: The Origins of Data Mining - Data Mining Tasks - OLAP and MultidimensionalDataAnalysis- Basic ConceptofAssociation AnalysisandCluster Analysis.

Machine Learning: History and Evolution - AI Evolution - Statistics vs. Data Mining vs. DataAnalytics vs. Data Science - Supervised Learning - Unsupervised Learning - ReinforcementLearning-FrameworksforBuildingMachine Learning Systems.

Unit-IV:Applications of Business Analytics

Overview of Business Analytics Applications: Financial Analytics - Marketing Analytics - HRAnalytics - Supply Chain Analytics - Retail Industry - Sales Analytics - Web & Social MediaAnalytics-HealthcareAnalytics-EnergyAnalytics-TransportationAnalytics-LendingAnalytics -SportsAnalytics-Futureof Business Analytics.

Unit-V:Ethical, Legal and Organizational Issues

Issues&Challenges:BusinessAnalyticsImplementationChallenges-PrivacyandAnonymizaiton-HackingandInsider Threats - MakingCustomer Comfortable.

REFERENCES:

- JamesREvans,BusinessAnalytics,GlobalEdition,PearsonEducation
- UDineshKumar, BusinessAnalytics, WileyIndiaPvt.Ltd.,NewDelhi
- GerKoole,AnIntroductiontoBusinessAnalytics,Lulu.com,2019
- J.D.Camm,J.J.Cochran,M.J.Fry,J.W.Ohlmann,D.R.Anderson,D.J.Sweeney,T.A.Williams-Essentials ofBusiness Analytics,2e;Cengage Learning.
- VipinKumar,IntroductiontoDataMining,Pang-NingTan,MichaelSteinbach,PearsonEducationIndia
-
    BhimasankaramPochiraju,SridharSeshadri,EssentialsofBusinessAnalytics:AnIntroductiontothe Methodology anditsApplication, Springer

UNIT 1
Understanding Business Analytics

Introduction – Meaning of Analytics-Evolution of Analytics-Need of Analytics- Business Analytics vs. Business Analytics – Categorization of Analytical Models – Data Scientist vs. Data Engineer vs. Business Analyst – Business Analytics in practice- Types of Data- Role of Business Analyst.

Introduction

The word analytics has come into the foreground in last decade or so. The increase of the internet and information technology has made analytics very relevant in the current age. Analytics is a field which combines data, information technology, statistical analysis, quantitative methods and computer-based models into one.

This all are combined to provide decision makers all the possible scenarios to make a well thought and researched decision. The computer-based model ensures that decision makers are able to see performance of decision under various scenarios.

Meaning

Business analytics (BA) is a set of disciplines and technologies for solving business problems using data analysis, statistical models and other quantitative methods. It involves an <u>iterative</u>, methodical exploration of an organization's data, with an emphasis on statistical analysis, to drive decision-making.

At its core, business analytics involves a combination of the following:

- identifying new patterns and relationships with data mining;

- using quantitative and statistical analysis to design business models;

- conducting A/B and multi-variable testing based on findings;

- forecasting future business needs, performance, and industry trends with predictive modelling; and

- Communicating your findings in easy-to-digest reports to colleagues, management, and customers.

Definition

➢ **Business analytics** (**BA**) refers to the skills, technologies, and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods.

➢ **Business Analytics** is the process of transforming data into insights to improve business decisions. Data management, data visualization, predictive modelling, data

mining, forecasting simulation, and optimization are some of the tools used to create insights from data.

Evolution of Business Analytics

➢ Business analytics has been existence since very long time and has evolved with availability of newer and better technologies. It has its roots in operations research, which was extensively used during World War II.

➢ Operations research was an analytical way to look at data to conduct military operations. Over a period of time, this technique started getting utilized for business. Here operation's research evolved into management science. Again, basis for management science remained same as operation research in data, decision making models, etc.

➢ Analytics have been used in business since the management exercises were put into place by Frederick Winslow Taylor in the late 19th century.

➢ Henry Ford measured the time of each component in his newly established assembly line. But analytics began to command more attention in the late 1960s when computers were used in decision support systems.

➢ Since then, analytics have changed and formed with the development of enterprise resource planning (ERP) systems, data warehouses, and a large number of other software tools and processes.

In later years the business analytics have exploded with the introduction of computers. This change has brought analytics to a whole new level and has brought about endless possibilities. As far as analytics has come in history, and what the current field of analytics is today, many people would never think that analytics started in the early 1900s with Mr. Ford himself.

As the economies started developing and companies became more and more competitive, management science evolved into business intelligence, decision support systems and into PC software.

❖ **Scope of Business Analytics**

Business analytics has a wide range of application and usages. It can be used for descriptive analysis in which data is utilized to understand past and present situation. This kind of descriptive analysis is used to asses' current market position of the company and effectiveness of previous business decision.

It is used for predictive analysis, which is typical used to asses' previous business performance.

Business analytics is also used for prescriptive analysis, which is utilized to formulate optimization techniques for stronger business performance.

For example, business analytics is used to determine pricing of various products in a departmental store based past and present set of information.

❖ **How business analytics works**

Before any data analysis takes place, BA starts with several foundational processes:
- Determine the business goal of the analysis.
- Select an analysis methodology.
- Get business data to support the analysis, often from various systems and sources.
- Cleanse and integrate data into a single repository, such as a <u>data warehouse</u> or <u>data mart</u>.

❖ **Need/Importance of Business Analytics**

- **Business analytics is a methodology or tool to make a sound commercial decision**. Hence it impacts functioning of the whole organization. Therefore, business analytics can help improve profitability of the business, increase market share and revenue and provide better return to a shareholder.
- Facilitates better understanding of available primary and secondary data, which again affect operational efficiency of several departments.
- Provides a competitive advantage to companies. In this digital age flow of information is almost equal to all the players. It is how this information is utilized makes the company competitive. Business analytics combines available data with various well thought models to improve business decisions.
- Converts available data into valuable information. This information can be presented in any required format, comfortable to the decision maker.

For starters, business analytics is the tool your company needs to make accurate decisions. These decisions are likely to impact your entire organization as they help you to improve profitability, increase market share, and provide a greater return to potential shareholders.

While some companies are unsure what to do with large amounts of data, business analytics works to combine this data with actionable insights to improve the decisions you make as a company

Essentially, the four main ways business analytics is important, no matter the industry, are:
- Improves performance by giving your business a clear picture of what is and isn't working
- Provides faster and more accurate decisions
- Minimizes risks as it helps a business make the right choices regarding consumer behaviour, trends, and performance
- Inspires change and innovation by answering questions about the consumer.

❖ **Essentials of business analytics**

Business analytics has many use cases, but when it comes to commercial organizations, BA is typically used to:
- Analyze data from a variety of sources. This could be anything from cloud applications to marketing automation tools and CRM software.

- Use advanced analytics and statistics to find patterns within datasets. These patterns can help you predict trends in the future and access new insights about the consumer and their behaviour.
- Monitor KPIs and trends as they change in real-time. This makes it easy for businesses to not only have their data in one place but to also come to conclusions quickly and accurately.
- Support decisions based on the most current information. With BA providing such a vast amount of data that you can use to back up your decisions, you can be sure that you are fully informed for not one, but several different scenarios.

❖ **Data for Analytics**

- Business analytics uses data from three sources for construction of the business model. It uses business data such as annual reports, financial ratios, marketing research, etc. It uses the database which contains various computer files and information coming from data analysis.

Benefits of implementing BA in your organization

Apart from having applications in various arenas, following are the benefits of Business Analytics and its impact on business –
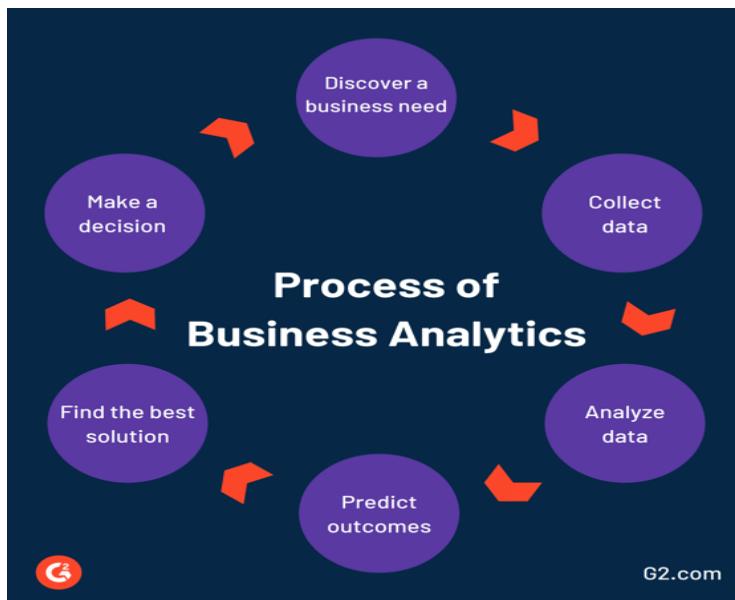
- Accurately transferring information
- Consequent improvement in efficiency
- Help portray Future Challenges
- Make Strategic decisions
- As a perfect blend of data science and analytics
- Reduction in Costs
- Improved Decisions
- Share information with a larger audience
- Ease in Sharing information with stakeholders

❖ **Challenges**

Moreover, any technology is subject to its own set of problems and challenges. Following are the challenges in implementing business analytics in an organization.

- Lack of technical skills in employees
- Fuss over acceptance of BA by staff
- Data Security and Maintenance
- Integrity of Data
- Delivering relevant information in the given time
- Inability to address complex issues
- Costs involved in implementing BA
- Investment of staff time in implementation of BA
- Lack of a proper strategy to implement BA

➢ Business analytics can be possible only on large volume of data. It is sometime difficult obtain large volume of data and not question its integrity.

➢ Business analytics depends on sufficient volumes of high-quality data.

➢ The difficulty in ensuring data quality is integrating and reconciling data across different systems, and then deciding what subsets of data to make available.

➢ Previously, analytics was considered a type of after-the-fact method of forecasting consumer behaviour by examining the number of units sold in the last quarter or the last year. This type of data warehousing required a lot more storage space than it did speed.

➢ Now business analytics is becoming a tool that can influence the outcome of customer interactions. When a specific customer type is considering a purchase, an analytics-enabled enterprise can modify the sales pitch to appeal to that consumer. This means the storage space for all that data must react extremely fast to provide the necessary data in real-time.

❖ **Application**

Business analytics has a wide range of application from customer relationship management, financial management, and marketing, supply-chain management, human-resource management, pricing and even in sports through team game strategies.

In healthcare, business analysis can be used to operate and manage clinical information systems. It can transform medical data from a bewildering array of analytical methods into useful information. Data analysis can also be used to generate contemporary reporting systems which include the patient's latest key indicators, historical trends and reference values.

- **Decision analytics**: supports human decisions with visual analytics that the user models to reflect reasoning.
- **Descriptive analytics:** gains insight from historical data with reporting, scorecards, clustering etc.
- **Predictive analytics:** employs predictive modelling using statistical and machine learning techniques

- **Prescriptive analytics:** recommends decisions using optimization, simulation, etc.

- Behavioural analytics
- Cohort analysis
- Competitor analysis
- Cyber analytics
- Enterprise optimization
- Financial services analytics
- Fraud analytics
- Health care analytics
- Key Performance Indicators (KPI's)
- Marketing analytics
- Pricing analytics
- Retail sales analytics
- Risk & Credit analytics
- Supply chain analytics
- Talent analytics
- Telecommunications
- Transportation analytics
- Customer Journey Analytics
- Market Basket Analysis

❖ **Business Analysis vs. Business Analytics**

The aim of business analytics is data and reporting—examining past business performance and forecasting future business performance. On the other hand, the business analysis focuses on functions and processes—determining business requirements and suggesting solutions.

- *Business Analysis: Definition and Activities*

Business analysis is the practice of assisting firms in resolving their technical difficulties by understanding, defining, and solving those issues.

The activities that are carried out while performing Business Analysis:

- **Company analysis**: Business analysis aims at figuring out the requirements of a firm in general and its strategic direction and determining the initiatives that will enable the business to address those strategic goals.
- **Requirements planning and management**: It focuses on planning the requirements of the development process, identifying what the top priority is for execution, and managing the changes.
- **Requirements elicitation**: It outlines techniques for collecting needs from relevant members of the project team.
- **Requirements analysis and documentation**: It explains how to establish and define the needs in detail to allow them to be effectively carried out by the team.

- **Requirements communication**: Business analysis explains methods to help stakeholders have a shared understanding of the needs and how they will be carried out.
- **Solution assessment and validation**: It also explains how a business analyst can execute a suggested solution, how to support the execution of a solution, and how to evaluate possible flaws in the implementation.

Business analysis is performed by Functional Analysts, Systems Analysts, Business Analysts, and Business Requirements Analysts.

➢ *Business Analytics: Definition and Its Applications*

Business analytics is also known as data analytics. It is a process of collecting, evaluating, and drawing valuable outcomes from the enormous amount of data available. Business analytics is widely used in the following applications:

- Finance
- Marketing
- HR
- CRM
- Manufacturing
- Banking and Credit Cards

Business analytics is performed by Data Scientists and Data Analysts.

➢ *Business Analysis vs. Business Analytics*

Most people believe that business analysis and analytics are the same, but they are not! The primary differences between business analysis and business analytics:

Business Analysis

- It mainly aims at the methods and determining the business needs.
- It is employed to figure out the organizational needs and possible problems to have productive outcomes.
- Here, the tasks are carried out by Functional Analysts, Systems Analysts, and Business Analysts.
- Business, functional, and domain skills are needed to perform business analysis.
- The architectural domains for business analysis include enterprise architecture, process architecture, technology architecture, and organization architecture.

Business Analytics

- It aims at data and reporting.
- It is widely practiced to reckon further stats and make decisions to bring improvements in the business.
- Here, the tasks are carried out by Data Scientists and Data Analysts.

- Mathematical, statistical, and programming skills are needed for executing business analytics.
- The architectural domains for business analytics include data architecture, technology architecture, and information architecture.

Business Analysis vs Business Analytics

Focuses on data and statistical analysis	Focuses on methods, techniques, and functions
Used to reckon stats and drive business decisions	Used to figure out complex business problems and produce positive outcomes
Carried out by Data Scientists and Data Analysts	Carried out by Business Analysts, Functional Analysts, and System Analysts
Skills required: Mathematical, statistical, and programming skills	Skills required: Fundamental, business, and technical skills
Architectural domains: Data, technology, and information	Architectural domains: Enterprise, process, technology, and organization

➢ *Business Analysis vs. Analytics: Similarities Explained*

Business analysis and business analytics have some commonalities. They both:

- Examine and enhance businesses
- Determine solutions to issues
- Establish things based on the requirements

Business analysis is a practice of identifying business requirements and figuring out solutions to specific business problems. This has a heavy overlap with the analysis of business needs to function normally and to enhance how they function. Sometimes, the solutions include a system's development feature. It can also incorporate business change, process enhancement or strategic planning, and policy improvement.

On the contrary, business analytics is all about the group of tools, techniques, and skills that help the investigation of previous business performance. It also aids to gain insights into future performance. In general, business analytics aims mostly at data and statistical analysis.
Categorization of Analytical Models

4 Types of Business Analytics
There are mainly four types of Business Analytics, each of these types are increasingly complex. They allow us to be closer to achieving real-time and future situation insight application. Each of these types of business analytics have been discussed below.
1. **Descriptive Analytics**
2. **Diagnostic Analytics**
3. **Predictive Analytics**
4. **Prescriptive Analytics**

1. Descriptive Analytics

It summarizes an organisation's existing data to understand what has happened in the past or is happening currently. Descriptive Analytics is the simplest form of analytics as it employs data aggregation and mining techniques. It makes data more accessible to members of an organisation such as the investors, shareholders, marketing executives, and sales managers.

It can help identify strengths and weaknesses and provides an insight into customer behaviour too. This helps in forming strategies that can be developed in the area of targeted marketing.

2. Diagnostic Analytics

This type of Analytics helps shift focus from past performance to the current events and determine which factors are influencing trends. To uncover the root cause of events, techniques such as data discovery, data mining and drill-down are employed. Diagnostic analytics makes use of probabilities, and likelihoods to understand why events may occur. Techniques such as sensitivity analysis and training algorithms are employed for classification and regression.

3. Predictive Analytics

This type of Analytics is used to forecast the possibility of a future event with the help of statistical models and ML techniques. It builds on the result of descriptive analytics to devise models to extrapolate the likelihood of items. To run predictive analysis, Machine Learning experts are employed. They can achieve a higher level of accuracy than by business intelligence alone.

One of the most common applications is sentiment analysis. Here, existing data collected from social media and is used to provide a comprehensive picture of an users opinion. This data is analysed to predict their sentiment (positive, neutral or negative).

4. Prescriptive Analytics

Going a step beyond predictive analytics, it provides recommendations for the next best action to be taken. It suggests all favourable outcomes according to a specific course of action and also recommends the specific actions needed to deliver the most desired result. It mainly relies on two things, a strong feedback system and a constant iterative analysis. It learns the relation between actions and their outcomes. One common use of this type of analytics is to create recommendation systems.

Business Analytics	Questions	Tools	Outcomes	Focus
Prescriptive (Automation)	• What should I do? • Why should I do it?	• Decision modeling • Optimization • Simulation • Expert systems	• Optimization-Best possible business decisions	• Focus on decision making and efficiency
Predictive (Foresight)	• What is likely to happen? • What will happen? • Why will it happen	• Data mining • Text/media mining • Predictive modeling • Artificial Neural Networks (ANN)	• Accurate projections of the future conditions and states	• Identify past patterns to predict the future
Diagnostic (Insight)	• Why did it happen?	• Enterprise data warehouse • Data discovery • Data mining and correlations • Drill-down/roll-up	• Accurate projections of the future conditions and states	• Identify past patterns to predict the future
Descriptive (Hindsight)	• What happened? • What is happening?	• Data modeling • Business reporting • Visualization • Dashboard • Regression	• Well defined business problems or opportunities	• Uncovering patterns that offer insight

❖ Business Analytics Tools

Business Analytics tools help analysts to perform the tasks at hand and generate reports which may be easy for a layman to understand. These tools can be obtained from open source platforms, and enable business analysts to manage their insights in a comprehensive manner. They tend to be flexible and user-friendly. Various business analytics tools and techniques like.

- **Python** is very flexible and can also be used in web scripting. It is mainly applied when there is a need for integrating the data analyzed with a web application or the statistics is to be used in a database production. The I Python Notebook facilitates and makes it easy to work with Python and data. One can share notebooks with other people without necessarily telling them to install anything which reduces code organizing overhead

- **SAS** The tool has a user-friendly GUI and can churn through terabytes of data with ease. It comes with an extensive documentation and tutorial base which can help early learners get started seamlessly.

- **R** is open source software and is completely free to use making it easier for individual professionals or students starting out to learn. Graphical capabilities or data visualization is the strongest forte of R with R having access to packages like GGPlot, RGIS, Lattice, and GGVIS among others which provide superior graphical competency.

- **Tableau** is the most popular and advanced data visualization tool in the market. Story-telling and presenting data insights in a comprehensive way has become one of the trademarks of a competent business analyst Tableau is a great platform to develop customized visualizations in no time, thanks to the drop and drag features.

Python, R, SAS, Excel, and Tableau have all got their unique places when it comes to usage.

❖ **Data Scientist vs. Data Engineer vs. Data Analyst**

1. Data scientists use their advanced statistical skills to help improve the models the data engineers implement and to put proper statistical rigour on the data discovery and analysis the customer is asking for.

- Companies extract data to analyze and gain insights about various trends and practices. In order to do so, they employ specialized data scientists who possess knowledge of statistical tools and programming skills. Moreover, a data scientist possesses ***knowledge of machine learning algorithms***.

- However, Data Science is not a singular field. It is a quantitative field that shares its background with math, statistics and computer programming. With the help of data science, industries are qualified to make careful data-driven decisions.

- These algorithms are responsible for predicting future events. Therefore, data science can be thought of as an ocean that includes all the data operations like data extraction, data processing, data analysis and data prediction to gain necessary insights.

A Data Scientist is required to perform **responsibilities** –
- Performing data pre-processing that involves data transformation as well as data cleaning.
- Using various machine learning tools to forecast and classify patterns in the data.
- Increasing the performance and accuracy of machine learning algorithms through fine-tuning and further performance optimization.
- Understanding the requirements of the company and formulating questions that needs to be addressed.
- Using robust storytelling tools to communicate results with the team members.

For becoming a Data Scientist, you must have the following key **skills** –
- Should be proficient with Math and Statistics.
- Should be able to handle structured & unstructured information.
- In-depth knowledge of tools like R, Python and SAS.
- Well versed in various machine learning algorithms.
- Have knowledge of SQL(Structured Query Language) and NoSQL(Non Structured Query Language or not only SQL)
- Must be familiar with Big Data tools.

Some of the **tools** that are used by Data Scientist **are**
- **Web Scraping**
- **Data Analytics**
- **Machine Learning**
- **Reporting**

2. A Data Engineer is a person who specializes in preparing data for analytical usage. Data Engineering also involves the development of platforms and architectures for data processing.

- In other words, a data engineer develops the foundation for various data operations. A Data Engineer is responsible for designing the format for data scientists and analysts to work on.

- Data Engineers have to work with both structured and unstructured data. Therefore, they _**need expertise in SQL**_ and NoSQL databases both. Data Engineers allow data scientists to carry out their data operations.

- Data Engineers have to deal with Big Data where they engage in numerous operations like data cleaning, management, transformation, data deduplication etc.

- A Data Engineer is more experienced with core programming concepts and algorithms. The **role of a data engineer** also follows closely to that of a software engineer. This is because a data engineer is assigned to develop platforms and architecture that utilize guidelines of software development.

For example, developing a cloud infrastructure to facilitate real-time analysis of data requires various development principles. Therefore, building an interface API is one of the job responsibilities of a data engineer.

Tools used by Data Engineers

Some of the **tools** that are used by Data Engineers **are** –
- *Hadoop*
- *Apache Spark*
- *Kubernetes*
- *Java*
- *Yarn*

A Data Engineer is supposed to have the following **responsibilities –**
- Development, construction, and maintenance of data architectures.
- Conducting testing on large scale data platforms.
- Handling error logs and building robust data pipelines.
- Ability to handle raw and unstructured data.
- Provide recommendations for data improvement, quality, and efficiency of data.
- Ensure and support the data architecture utilized by data scientists and analysts.
- Development of data processes for data modelling, mining, and data production.

Following are the key **skills** required to become a data engineer –

- Knowledge of programming tools like Python and Java.
- Solid Understanding of Operating Systems.
- Ability to develop scalable ETL packages.
- Should be well versed in SQL as well as NoSQL technologies like Cassandra and MongoDB.
- He should possess knowledge of data warehouse and big data technologies like Hadoop, Hive, Pig, and Spark.
- Should possess creative and out of the box thinking.

3. A **Data Analyst** is responsible for taking actionable that affect the current scope of the company. A **data engineer** is responsible for developing a platform those data analysts and data scientists work on. And, a **data scientist** is responsible for unearthing future insights from existing data and helping companies to make data-driven decisions.

- A **data analyst** does not directly participate in the decision-making process; rather, he helps indirectly through providing static insights about company performance. A **data engineer** is not responsible for decision making. And, a **data scientist** participates in the active decision-making process that affects the course of the company.

- A **data analyst** uses static modelling techniques that summarize the data through descriptive analysis. On the other hand, a **data engineer** is responsible for the development and maintenance of data pipelines. A **data scientist** uses dynamic techniques like **Machine learning to gain insights about the future**.

- Knowledge of machine learning is not important for **data analysts**. However, this is mandatory for **data scientists**. A **data engineer** need not require the knowledge of machine learning but he is required to have the knowledge of core computing concepts like programming and algorithms to build robust data systems.

- A **data analyst** only has to deal with structured data. However, both **data scientists and data engineers** deal with unstructured data as well.

- Data analyst and data scientists are both required to be proficient in data visualization. However, this is not required in the case of a **data engineer**.

- Both **data scientists and analysts** need not have knowledge of application development and working of the APIs. However, this is the most essential requirement for a **data engineer**.

A Data Analyst has following **responsibilities -**
- Analyzing the data through descriptive statistics.
- Using database query languages to retrieve and manipulate information.
- Perform data filtering, cleaning and early stage transformation.
- Communicating results with the team using data visualization.
- Work with the management team to understand business requirements.

In order to become a Data Analyst, you must possess the following **skills** –
- Should possess the strong mathematical aptitude
- Should be well versed with Excel, Oracle, and SQL.
- Possession of problem-solving attitude.
- Proficient in the communication of results to the team.
- Should have a strong suite of analytical skills.

Some of the **tools** that are used by Data Analyst **are**

- Talend :Talend is one of the most powerful data analytics tools available in the market and is developed in the eclipse graphical development environment. ...
- Qlik Sense. ...
- Apache Spark. ...
- Power BI. ...
- ThoughtSpot. ...
- RapidMiner. ...
- Tableau

Business Analyst

Business analysts use data to form business insights and recommend changes in businesses and other organizations. Business analysts can identify issues in virtually any part of an organization, including IT processes, organizational structures, or staff development.

As businesses seek to increase efficiency and reduce costs, business analytics has become an important component of their operations. Let's take a closer look at what business analysts do and what it takes to get a job in business analysis.

Business analysts identify business areas that can be improved to increase efficiency and strengthen business processes. They often work closely with others throughout the business hierarchy to communicate their findings and help implement changes.

Tasks and duties can include:
- Identifying and prioritizing the organization's functional and technical needs and requirements
- Using SQL and Excel to analyze large data sets
- Compiling charts, tables, and other elements of data visualization
- Creating financial models to support business decisions
- Understanding business strategies, goals, and requirements
- Planning enterprise architecture (the structure of a business)
- Forecasting, budgeting, and performing both variance analysis and financial analysis

Business analyst skills

The key skills business analysts need are:
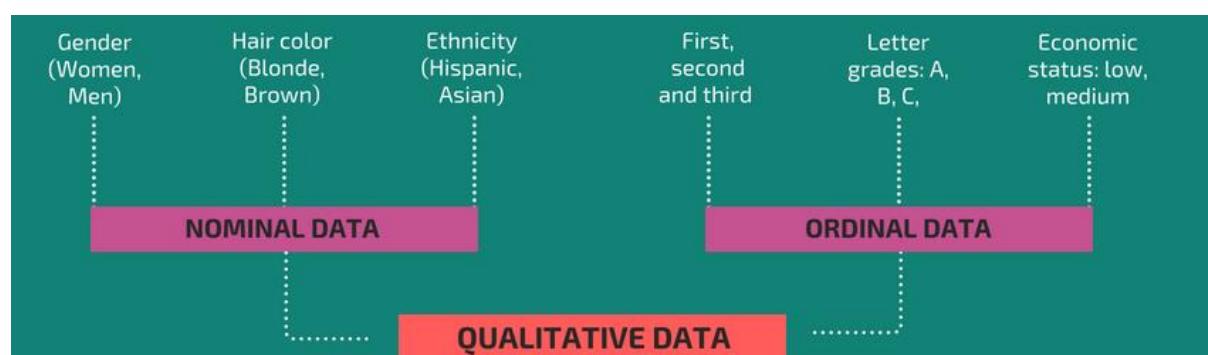
- **Technical skills:** These skills include stakeholder management, data modeling and knowledge of IT.
- **Analytical skills:** Business analysts have to analyze large amounts of data and other business processes to form ideas and fix problems.
- **Communication:** These professionals must communicate their ideas in an expressive way that is easy for the receiver to understand.
- **Problem-solving:** It is a business analyst's primary responsibility to come up with solutions to an organization's problems.
- **Research skills:** Thorough research must be conducted about new processes and software to present results that are effective.

Business analyst responsibilities
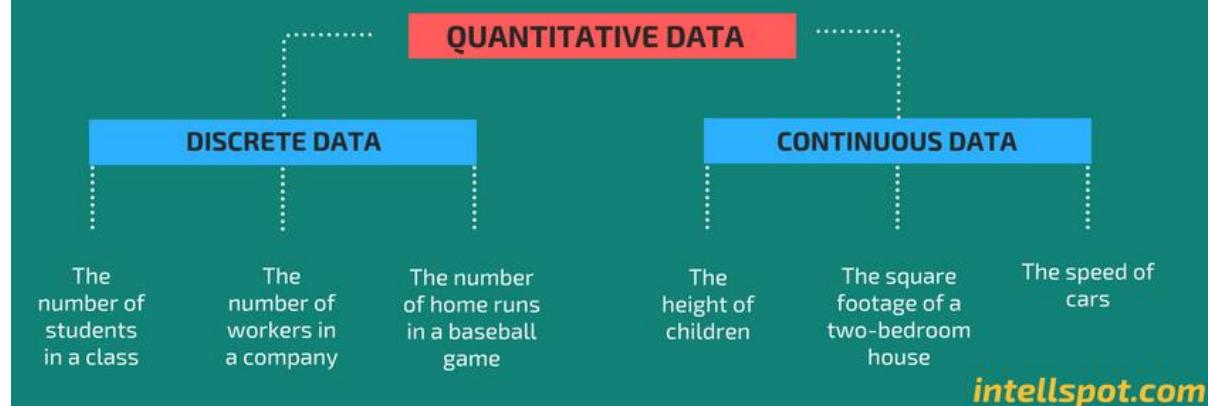
- Analyzing and evaluating the current business processes a company has and identifying areas of improvement
- Researching and reviewing up-to-date business processes and new IT advancements to make systems more modern
- Presenting ideas and findings in meetings
- Training and coaching staff members
- Creating initiatives depending on the business's requirements and needs
- Developing projects and monitoring project performance
- Collaborating with users and stakeholders
- Working closely with senior management, partners, clients and technicians

Types of Data
Qualitative vs. Quantitative Data

1. Quantitative data
- Quantitative data seems to be the easiest to explain. It answers key questions such as "how many, "how much" and "how often".
- Quantitative data can be expressed as a number or can be quantified. Simply put, it can be measured by numerical variables.
- Quantitative data are easily amenable to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as line, bar graph, scatter plot, and etc.

Examples of quantitative data:
- Scores on tests and exams e.g. 85, 67, 90 and etc.
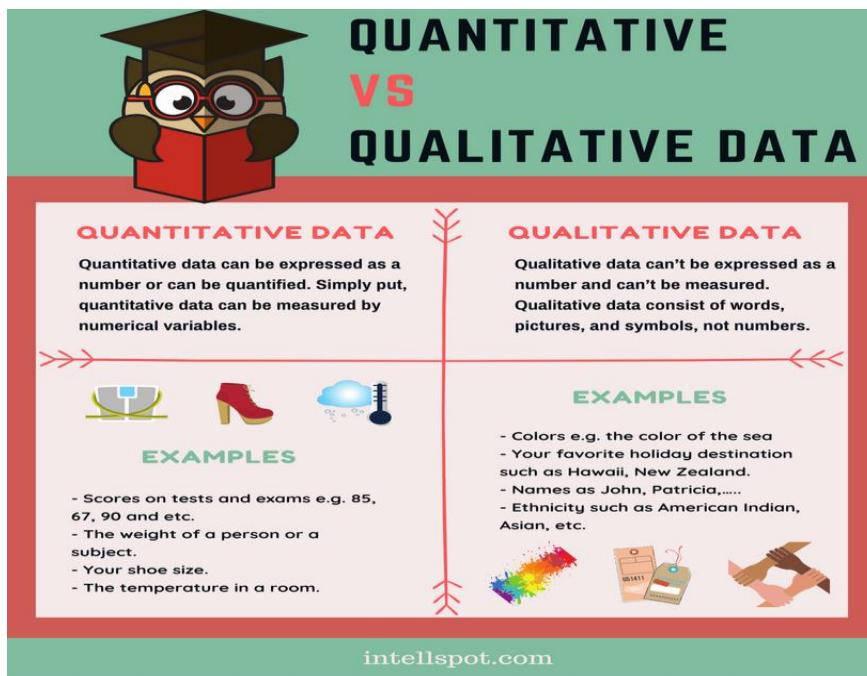
- The weight of a person or a subject.
- Your shoe size.
- The temperature in a room.

2. Qualitative data

- Qualitative data can't be expressed as a number and can't be measured. Qualitative data consist of words, pictures, and symbols, not numbers.
- Qualitative data is also called categorical data because the information can be sorted by category, not by number.
- Qualitative data can answer questions such as "how this has happened" or and "why this has happened".

Examples of qualitative data:

- Colors e.g. the color of the sea
- Your favorite holiday destination such as Hawaii, New Zealand and etc.
- Names as John, Patricia..
- Ethnicity such as American Indian, Asian, etc.

Nominal vs. Ordinal Data

3. Nominal data

Nominal data is used just for labelling variables, without any type of quantitative value. The name 'nominal' comes from the Latin word "nomen" which means 'name'.

The nominal data just name a thing without applying it to order. Actually, the nominal data could just be called "labels."

Examples of Nominal Data:

- Gender (Women, Men)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Marital status (Married, Single, Widowed)
- Ethnicity (Hispanic, Asian)

Eye color is a nominal variable having a few categories (Blue, Green, Brown) and there is no way to order these categories from highest to lowest.

4. Ordinal data

Ordinal data shows where a number is in order. This is the crucial difference from nominal types of data.

Ordinal data is data which is placed into some kind of order by their position on a scale. Ordinal data may indicate superiority.

However, **you cannot do arithmetic with ordinal numbers** because they only show sequence.

Ordinal variables are considered as "in between" qualitative and quantitative variables.
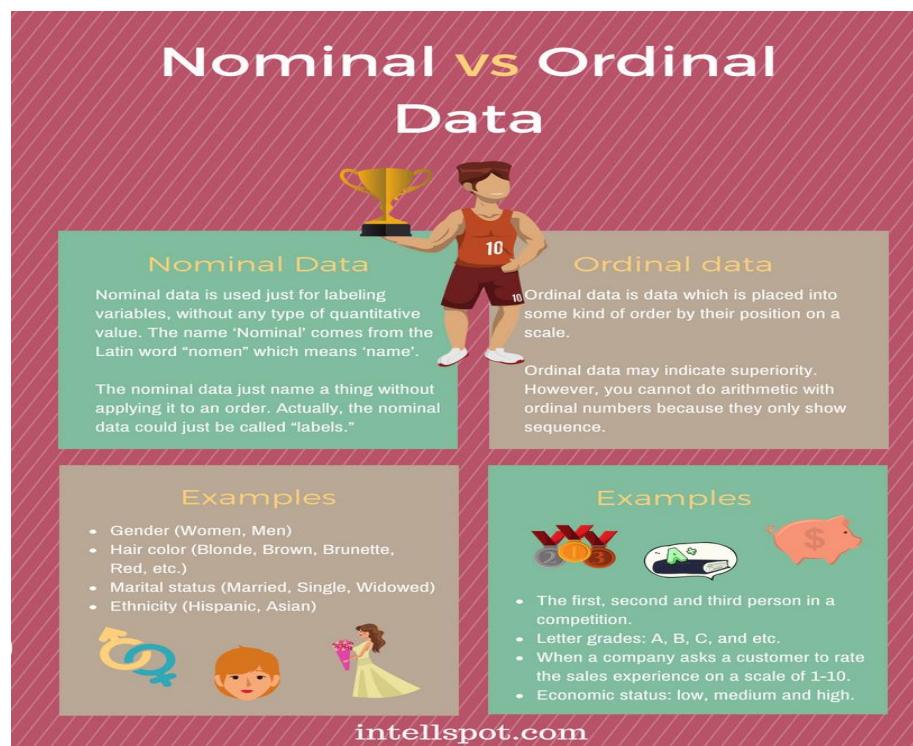
In other words, the ordinal data is qualitative data for which the values are ordered.

In comparison with nominal data, the second one is qualitative data for which the values cannot be placed in an ordered.

We can also assign numbers to ordinal data to show their relative position. But we cannot do math with those numbers. For example: "first, second, third…etc."

Examples of Ordinal Data:

▪   The first, second and third person in a competition.
▪   Letter grades: A, B, C, and etc.
▪   When a company asks a customer to rate the sales experience on a scale of 1-10.
▪   Economic status: low, medium and high.

Discrete vs. Continuous Data

In statistics, marketing research, and data science, many decisions depend on whether the basic data is discrete or continuous.

5. Discrete data

Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts.

For example, the number of children in a class is discrete data. You can count whole individuals. You can't count 1.5 kids.

To put in other words, discrete data can take only certain values. The data variables cannot be divided into smaller parts.

It has a limited number of possible values e.g. days of the month.

Examples of discrete data:

- The number of students in a class.
- The number of workers in a company.
- The number of home runs in a baseball game.
- The number of test questions you answered correctly

6. Continuous data

Continuous data is information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value.

For example, you can measure your height at very precise scales — meters, centimeters, millimeters and etc.

You can record continuous data at so many different measurements – width, temperature, time, and etc. This is where the key difference from discrete types of data lies.

The continuous variables can take any value between two numbers. For example, between 50 and 72 inches, there are literally millions of possible heights: 52.04762 inches, 69.948376 inches and etc.

A good great rule for defining if a data is continuous or discrete is that if the point of measurement can be reduced in half and still make sense, the data is continuous.

Examples of continuous data:

- The amount of time required to complete a project.
- The height of children.
- The square footage of a two-bedroom house.
- The speed of cars.

Conclusion

All of the different types of data have a critical place in statistics, research, and data science. Data types work great together to help organizations and businesses from all industries build successful data-driven decision-making process.

Working in the data management area and having a good range of data science skills involves a deep understanding of various types of data and when to apply them.

❖ ROLES OF A BUSINESS ANALYST

1. BA LEVELS

There are four levels that a business analyst in an organization comprises of:

- **Strategic management:** This is the analysis level, where a business analyst evaluates and calculates the strategic where about if a company. This is one of the most critical levels because unless the evaluation is done on the point, none of the further steps can work appropriately.
- **Analysis of business model:** This level has to do with evaluating policies that are currently being employed by the company. This not only enables us to implement what's new but also helps in checking the previous ones.
- **Designing the process:** Like an artist creates his imagination, business analysts do that with their skills. The step includes modelling the business processes, which comes out to be designing and modelling.
- **Analysis of technology:** Technical systems need a thorough analysis too. This is something that, if not taken care of, leads to severe consequences.

The key **business analyst roles and responsibilities**:

- ✓ **What does a business needs:** As a business analyst, it is his key responsibility to understand what stakeholders need and pass these requirements to the developers, and also give on the developer's expectations to the stakeholders. A business analyst's skill for this responsibility is the communication skills that can impress everyone across. While he transfers the information, he is the one who needs to put these in such words that make a difference. This responsibility is no doubt tome taking because he needs to listen and execute, which might seem easy, but only a skilled professional can handle all this.
- ✓ **Conducting meetings with developing team and stakeholders:** Business analysts are supposed to coordinate with both stakeholders and the development team whenever a new feature or update is added to a project. This may vary from project to project. This facilitates the collection of client feedback and the resolution of issues encountered by the development team when implementing new features. The **business analyst role** is to understand and explain the new feature updates to clients and take feedback for further development. Based on client feedback, Business Analyst instructs the development team to make amendments or continue as is. At times, the client requests an additional feature be added to a project, and the BA must determine whether or not it is feasible, and then assign resources if necessary to implement it.
- ✓ **System possibilities:** A business analyst might be considered one among those working in the software team, but their key responsibility Is not what the team does. He has to ensure that he figures out what a project needs. He is the one who leads the

path to the goals. He might be the one who dreams of targets, but he is also the one who knows how to make those dreams a reality. Looking for the opportunities and grabbing them before they go is what a business analyst is good at.

✓ **Present the company:** He can be called the face of a business. A business analyst is responsible for putting a business's thoughts and goals in front of the stakeholders. In short, he is the one who needs to impress the stakeholders with his presentation skills and the skill to present what the person on the other side is looking for and not what the company has in store for them.

✓ **Present the details:** A project brings with itself hundreds of minute details that might be left unseen. A business analyst is the one who is responsible for elaborating the project with the tiniest of the loopholes or hidden secrets. This is considered the most crucial role of a business analyst because unless the details are put across the stakeholders, they won't take an interest, and unless they show the part, the project is likely to take a pause.

✓ **Implementation of the project:** After going through all the steps mentioned above, the next and the most important role of a business analyst in agile is to implement whatever has been planned. Execution is not easy unless the previous steps have been taken care of in a systemized fashion.

✓ **Functional and non-functional requirements of a business:** As an organization, the main goal is to receive an end product that is productive and gives a company a long time. The role of business analyst in it company is to take care of the business's functional aspect, which includes the steps and ways to ensure the working of the project. Sideways he is also supposed to take care of the non-functional that comprise how a project or a business is supposed to work.

✓ **Testing:** The role of a business analyst is way longer than expected. Once the product is prepared, the next step is to test it among the users to know it's working capacity and quality. The Business Analyst tests the prototype/interface by involving some clients and recording their experiences with the model that has been developed, according to the role description. Based on their feedback, Business Analyst intends to make some changes to the model that will make it even better. They conduct UAT (user acceptance test) to determine whether or not the prototype meets the requirements of the project under consideration.

✓ **Decision making and problem-solving:** The **responsibilities of business analyst** range from developing the required documents to making decisions in the most stringent circumstances, job role of business analyst is **to** do it all. Moreover, a business analyst is expected to be the one who tackles things most easily and calmly because he should also be good at problem-solving, even if that's related to the stakeholders, employees, or the clients.

✓ **Maintenance:** Like they say that care is as essential as building something new. No matter how much human resources, energy, or finds you spend on a project, if the maintenance part is not taken care of properly or is neglected, it tends to spoil the entire hard work put across. What is the role of a business analyst here? Is it just limited to the maintenance of the clients or sales; it also has to ensure that the quality and the promised products are maintained throughout.

✓ **Building a team:** Everyone is born with varied skills. As a business analyst, the business analyst's responsibility is to make the team with people possessing different skills required for the project. Not only the hiring but retaining them is as essential. A well united and skilled team can do wonders. The things that are required in a great

section inside co combination, structuring, and skills. A good team tends to take the company to the heights of success.

- ✓ **Presentation and Documentation of the Final Project**: After the business project is completed, the Business Analyst must document the details of the project and share the project's findings with the client. In most cases, **BA roles and responsibilities** include preparing reports and presenting the results of a project to key stakeholders and clients. During building the project, they must also record all of the lessons learned and challenges they encountered in a concise form. This step aids the business analyst in making better decisions in the future.

CONCLUSION

A business analyst might be another position in an organization but its roles and responsibilities play a vital role in an organization's success. While he needs to be a good orator, he should possess the quality of bringing people closers to his team and across. His roles are not limited to a specific step in project management. He is required one overstep till the end. From the initial stages of evaluation to the maintenance, a company needs a business analyst's skill.

Data: Data Collection-Data Management-Big Data Management-Organization/sources of Data- Importance of Data Quality- Dealing with missing or incomplete data – Data Visualization- Data Classification.

Data Science project Life Cycle- Business Requirement – Data Acquisition- data Preparation- Hypothesis and Modelling- Evaluation and interpretation- Deployment-Operations-Optimization-Applications for Data Science.

Data
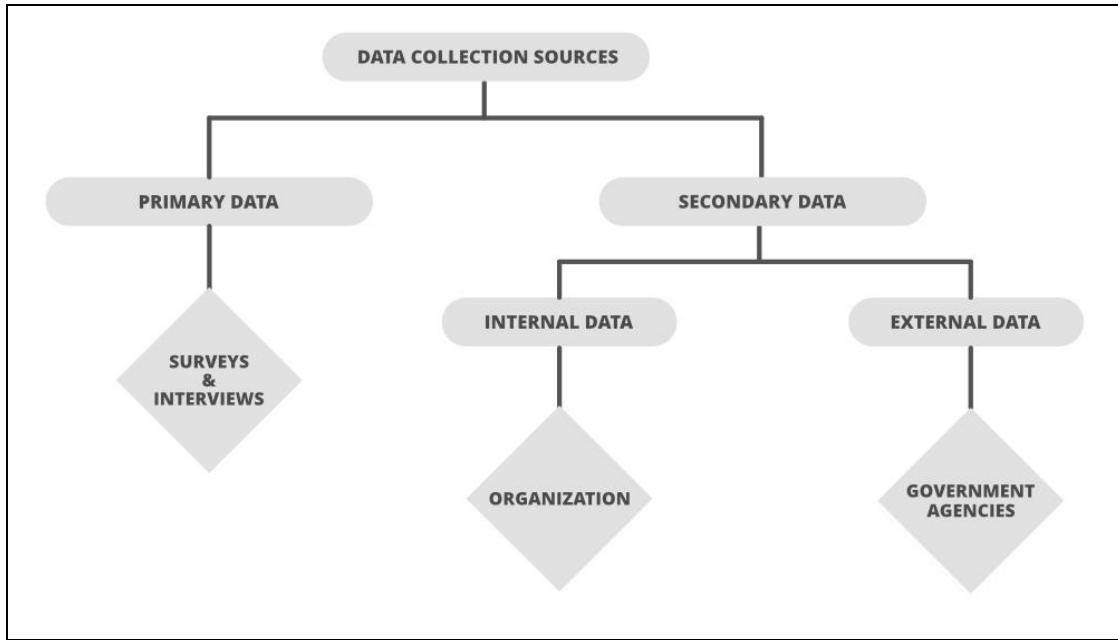
- Knowledge is power, information is knowledge, and data is information in digitized form, at least as defined in IT. Hence, data is power.
- Data are individual facts, statistics, or items of information, often numeric. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects
- Data is various kinds of information formatted in a particular way. Therefore, data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.
- Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity.
- The concept of data collection isn't a new one, as we'll see later, but the world has changed. There is far more data available today, and it exists in forms that were unheard of a century ago. The data collection process has had to change and grow with the times, keeping pace with technology.
- Data collection breaks down into two methods: 1. Primary & 2. Secondary

❖ **Data Collection**

Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis. In the process of big data analysis, "Data collection" is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

The actual data is then further divided mainly into two types known as:
1. **Primary data**
2. **Secondary data**

1. Primary data:

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

➢ **Interview method:**
The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

➢ **Survey method:**

The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

➢ **Observation method:**
The observation method is a method of data collection in which the researcher keenly observes the behaviour and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behaviour towards the products. The data obtained will be sent for processing.

➢ **Projective Technique**

Projective data gathering is an indirect interview, used when potential respondents know why they're being asked questions and hesitate to answer. For instance, someone may be reluctant

to answer questions about their phone service if a cell phone carrier representative poses the questions. With projective data gathering, the interviewees get an incomplete question, and they must fill in the rest, using their opinions, feelings, and attitudes.

➢ **Delphi Technique.**

The Oracle at Delphi, according to Greek mythology, was the high priestess of Apollo's temple, who gave advice, prophecies, and counsel. In the realm of data collection, researchers use the Delphi technique by gathering information from a panel of experts. Each expert answers questions in their field of specialty, and the replies are consolidated into a single opinion.

➢ **Focus Groups.**

Focus groups, like interviews, are a commonly used technique. The group consists of anywhere from a half-dozen to a dozen people, led by a moderator, brought together to discuss the issue.

➢ **Questionnaires.**

Questionnaires are a simple, straightforward data collection method. Respondents get a series of questions, either open or close-ended, related to the matter at hand.

➢ **Experimental method:**

The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

- **CRD- Completely Randomized design** is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.
- **RBD- Randomized Block Design** is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.
- **LSD – Latin Square Design** is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.
- **FD- Factorial design** is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

2. Secondary data:

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

i. **Internal source:**

These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

- Financial Statements
- Sales Reports
- Retailer/Distributor/Deal Feedback
- Customer Personal Information (e.g., name, address, age, contact info)
- Business Journals
- Government Records (e.g., census, tax records, Social Security info)
- Trade/Business Magazines
- The internet

ii. External source:

The data which can't be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labour bureau, syndicate services, and other non-governmental publications.

iii. Other sources:
- **Sensors data:** With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data:** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
- **Web traffic:** Due to fast and cheap internet facilities many formats of data Which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

❖ Data Collection Tools

1. Word Association.
The researcher gives the respondent a set of words and asks them what comes to mind when they hear each word.

2. Sentence Completion.
Researchers use sentence completion to understand what kind of ideas the respondent has. This tool involves giving an incomplete sentence and seeing how the interviewee finishes it.

3. Role-Playing.
Respondents are presented with an imaginary situation and asked how they would act or react if it was real.

4. In-Person Surveys.
The researcher asks questions in person.

5. Online/Web Surveys.
These surveys are easy to accomplish, but some users may be unwilling to answer truthfully, if at all.

6. Mobile Surveys.

These surveys take advantage of the increasing proliferation of mobile technology. Mobile collection surveys rely on mobile devices like tablets or smart phones to conduct surveys via SMS or mobile apps.

7. Phone Surveys.
No researcher can call thousands of people at once, so they need a third party to handle the chore. However, many people have call screening and won't answer.

8. Observation.
Sometimes, the simplest method is the best. Researchers who make direct observations collect data quickly and easily, with little intrusion or third-party bias. Naturally, it's only effective in small-scale situations.

❖ **Data Management**

Data management refers to the professional practice of constructing and maintaining a framework for ingesting, storing, mining, and archiving the data integral to a modern business. Data management is the spine that connects all segments of the information lifecycle.

Data management works symbiotically with process management, ensuring that the actions teams take are informed by the cleanest, most current data available — which in today's world means tracking changes and trends in real-time. Below is a deeper look at the practice, its benefits and challenges, and best practices for helping your organization get the most out of its business intelligence.

❖ **7 types of data management**
Data management experts generally focus on specialties within the field. These specialties can fall under one or more of the following areas:
1. **Master data management:** Master data management (MDM) is the process of ensuring the organization is always working with — and making business decisions based on — a single version of current, reliable information. Ingesting data from all of your data sources and presenting it as one constant, reliable source, as well as repropagating data into different systems, requires the right tools.

2. Data stewardship: A data steward does not develop information management policies but rather deploys and enforces them across the enterprise. As the name implies, a data steward stands watch over enterprise data collection and movement policies, ensuring practices are implemented and rules are enforced.

3. Data quality management: If a data steward is a kind of digital sheriff, a data quality manager might be thought of as his court clerk. Quality management is responsible for combing through collected data for underlying problems like duplicate records, inconsistent versions, and more. Data quality managers support the defined data management system.

4. Data security: One of the most important aspects of data management today is security. Though emergent practices like DevSecOps incorporate security considerations at every level of application development and data exchange, security specialists are still tasked with

encryption management, preventing unauthorized access, guarding against accidental movement or deletion, and other frontline concerns.

5. Data governance: Data governance sets the law for an enterprise's state of information. A data governance framework is like a constitution that clearly outlines policies for the intake, flow, and protection of institutional information. Data governors oversee their network of stewards, quality management professionals, security teams, and other people and data management processes in pursuit of a governance policy that serves a master data management approach.

6. Big data management: Big data is the catch-all term used to describe gathering, analyzing, and using massive amounts of digital information to improve operations. In broad terms, this area of data management specializes in intake, integrity, and storage of the tide of raw data that other management teams use to improve operations and security or inform business intelligence.

7. Data warehousing: Information is the building block of modern business. The sheer volume of information presents an obvious challenge: What do we do with all these blocks? Data warehouse management provides and oversees the physical and/or cloud-based infrastructure used to aggregate raw data and analyze it in-depth to produce business insights. The unique needs of any organization practicing data management may require a blend of some or all of these approaches. Familiarity with management areas provides data managers with the background they need to build solutions customized for their environments.

❖ **Benefits of data management systems**

Data management processes help organizations identify and resolve internal pain points to deliver a better customer experience.

First, data management provides businesses with a way of measuring the amount of data in play. A myriad of interactions occur in the background of any business — between network infrastructure, software applications, APIs, security protocols, and much more — and each presents a potential glitch (or time bomb) to operations if something goes wrong. Data management gives managers a big-picture look at business processes, which helps with both perspective and planning.

Once data is under management, it can be mined for informational gold: business intelligence. This helps business users across the organization in a variety of ways, including the following:

- Smart advertising that targets customers according to their interests and interactions
- Holistic security that safeguards critical information
- Alignment with relevant compliance standards, saving time and money
- Machine learning that grows more environmentally aware over time, powering automatic and continuous improvement
- Reduced operating expenses by restricting use to only the necessary storage and compute power required for optimal performance

❖ **Data management challenges**

- All these benefits don't come without climbing some hills. The ever-growing, rolling landscape of information technology is constantly changing and data managers will encounter plenty of challenges along the way.
- There are four key data management challenges to anticipate:

- **The amount of data can be (at least temporarily) overwhelming.** It's hard to overstate the volume of data that must come under management in a modern business, so, when developing systems and processes, be ready to think big. Really big. Specialized third-party services and apps for integrating big data or providing it as a platform are crucial allies.
- **Many organizations silo data.** The development team may work from one data set, the sales team from another, operations from another, and so on. A modern data management system relies on access to all this information to develop modern business intelligence. Re
- Real-time data platform services help stream and share clean information between teams from a single, trusted source.
- **The journey from unstructured data to structured data can be steep.** Data often pours into organizations in an unstructured way. Before it can be used to generate business intelligence, data preparation has to happen: Data must be organized, de-duplicated, and otherwise cleaned. Data managers often rely on third-party partnerships to assist with these processes, using tools designed for on-premises, cloud, or hybrid environments.
- **Managing the culture is essential to managing data.** All of the processes and systems in the world won't do you much good if people don't know how — and perhaps just as importantly, why — to use them. By making team members aware of the benefits of data management (and the potential pitfalls of ignoring it) and fostering the skills of using data correctly, managers engage team members as essential pieces of the information process.

These and other challenges stand between the old way of doing business and initiatives that harness the power of data for business intelligence. But with proper planning, practices, and partners, technologies like accelerated machine learning can turn pinch points into gateways for deeper business insights and better customer experience.

❖ **Data management best practices**
Though specific data needs are unique to every organization's data strategy and data systems, preparing a framework will smooth the path to easier, more effective data management solutions. Best practices like the three below are key to a successful strategy.
1. Make a plan
2. Store your data
3. Share your data

1. Make a plan
- Develop and write a data management plan (DMP). This document charts estimated data usage, accessibility guidelines, archiving approaches, ownership, and more. A DMP serves as both a reference and a living record and will be revised as circumstances change.

- Additionally, DMPs present the organization's overarching strategy for data management to investors, auditors, and other involved parties — which is an important insight into a company's preparedness for the rigors of the modern market.

The best DMPs define granular details, including:

- Preferred file formats
- Naming conventions
- Access parameters for various stakeholders
- Backup and archiving processes
- Defined partners and the terms and services they provide
- Thorough documentation
- There are online services that can help create DMPs by providing step-by-step guidance to creating plans from templates.

2. Store your data

- Among the granular details mentioned above, a solid data storage approach is central to good data management. It begins by determining if your storage needs best suit a data warehouse or a data lake (or both), and whether the company's data belongs on-premises or in the cloud.
- Then outline a consistent, and consistently enforced, agreement for naming files, folders, directories, users, and more. This is a foundational piece of data management, as these parameters will determine how to store all future data, and inconsistencies will result in errors and incomplete intelligence.

1. **Security and backups.** Insecure data is dangerous, so security must be considered at every layer. Some organizations come under special regulatory burdens like HIPAA, CIPA, GDPR, and others, which add additional security requirements like periodic audits. When security fails, the backup plan can be the difference between business life and death. Traditional models called for three copies of all important data: the original, the locally stored copy, and a remote copy. But emerging cloud models include decentralized data duplication, with even more backup options available at an increasingly affordable cost for storage and transfer.

2. **Documentation is key.** If it's important, document it. If the entire team splits the lottery and runs off to Jamaica, thorough, readable documentation outlining security and backup procedures will give the next team a fighting chance to pick up where they left off. Without it, knowledge resides exclusively with holders who may or may not be part of a long-term data management approach.

Data storage needs to be able to change as fast as the technology demands, so any approach should be flexible and have a reasonable archiving approach to keep costs manageable.

3. Share your data

After all the plans are laid for storing, securing, and documenting your data, you should begin the process of sharing it with the appropriate people.

Here are some critical questions to answer before other people access potentially critical information:

- Who owns the data?
- Can it be copied?
- Has everyone contributing to the data consented to share it with others?
- Who can access it and at what times?

- Are there copyrights, corporate secrets, proprietary intellectual property, or other off-limits information in the data set?
- What else does the organization's data reveal about itself?

With those and other questions answered, it's time to find a place and means of sharing the data. Once called a repository, this role is increasingly filled by software and infrastructure as service models that are fine-tuned for big data management.

❖ **Big Data Management**

Big data consists of huge amounts of information that cannot be stored or processed using traditional data storage mechanisms or processing techniques. It generally consists of three different variations.

i. **Structured data** (as its name suggests) has a well-defined structure and follows a consistent order. This kind of information is designed so that it can be easily accessed and used by a person or computer. Structured data is usually stored in the well-defined rows and columns of a table (such as a spreadsheet) and databases — particularly relational database management systems, or RDBMS.

ii. **Semi-structured data** exhibits a few of the same properties as structured data, but for the most part, this kind of information has no definite structure and cannot conform to the formal rules of data models such as an RDBMS.

iii. **Unstructured data** possesses no consistent structure across its various forms and does not obey conventional data models' formal structural rules. In very few instances, it may have information related to date and time.
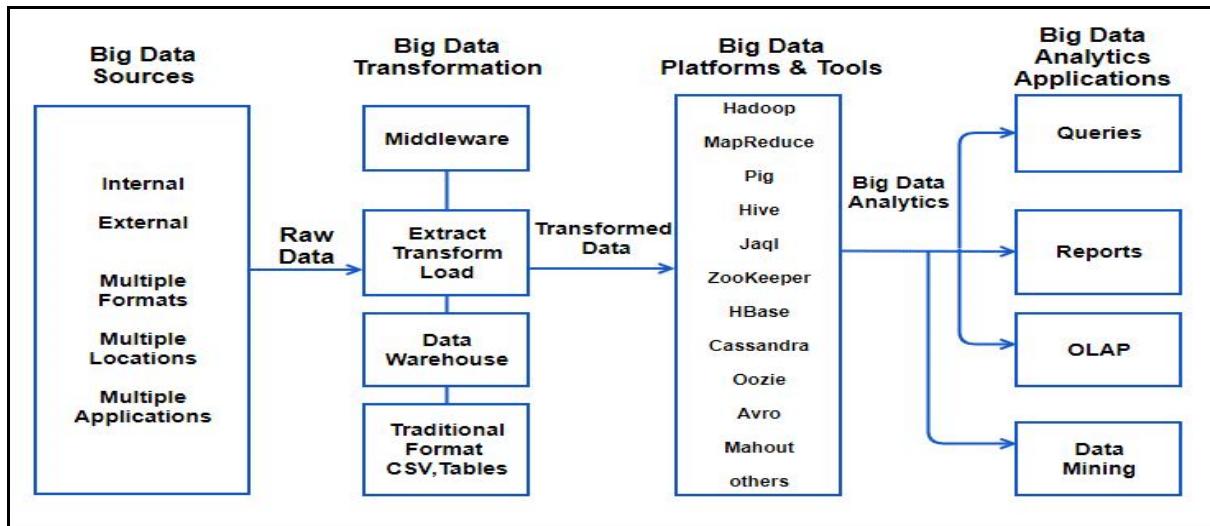
Characteristics of Big Data Management

In line with classical definitions of the concept, big data is generally associated with three core characteristics:

1. **Volume:** This trait refers to the immense amounts of information generated every second via social media, cell phones, cars, transactions, connected sensors, images, video, and text. In petabytes, terabytes, or even zettabytes, these volumes can only be managed by big data technologies.

2. **Variety:** To the existing landscape of transactional and demographic data such as phone numbers and addresses, information in the form of photographs, audio streams, video, and a host of other formats now contributes to a multiplicity of data types — about 80% of which are completely unstructured.

3. **Velocity:** Information is streaming into data repositories at a prodigious rate, and this characteristic alludes to the speed of data accumulation. It also refers to the speed with which big data can be processed and analyzed to extract the insights and patterns it contains. These days, that speed is often real-time.

Beyond "the Three Vs," current descriptions of big data management also include two other characteristics, namely:

- **Veracity:** This is the degree of reliability and truth that big data has to offer in terms of its relevance, cleanliness, and accuracy.

- **Value:** Since the primary aim of big data gathering and analysis is to discover insights that can inform decision-making and other processes, this characteristic explores the benefit or otherwise that information and analytics can ultimately produce.

Big Data Management Services

When it comes to technology, organizations have many different types of big data management solutions to choose from. Vendors offer a variety of standalone or multi-featured big data management tools, and many organizations use multiple tools. Some of the most common types of big data management capabilities include the following:

- **Data cleansing:** finding and fixing errors in data sets

- **Data integration:** combining data from two or more sources

- **Data migration:** moving data from one environment to another, such as moving data from in-house data centres to the cloud

- **Data preparation:** readying data to be using in analytics or other applications

- **Data enrichment:** improving the quality of data by adding new data sets, correcting small errors or extrapolating new information from raw data

- **Data analytics:** analysing data with a variety of algorithms in order to gain insights

- **Data quality:** making sure data is accurate and reliable

- **Master data management (MDM) :**linking critical enterprise data to one master set that serves as the single source of truth for the organization

- **Data governance:** ensuring the availability, usability, integrity and accuracy of data

- **Extract transform load (ETL):** moving data from an existing repository into a database or data warehouse.

❖ **Organization/Sources of Data**

Data organization is the practice of categorizing and classifying data to make it more usable. Similar to a file folder, where we keep important documents, you'll need to arrange your data in the most logical and orderly fashion, so you — and anyone else who accesses it — can easily find what they're looking for.

DATA IS BEING COLLECTED

- The big data includes information produced by humans and devices.
- Device-driven data is largely clean and organized,
- But of far greater interest is human-driven data that exist in various formats and need more exquisite tools for proper processing and management.

The big data collection is focused on the following types of data:

➢ **Network data.** This type of data is gathered on all kinds of networks, including social media, information and technological networks, the Internet and mobile networks, etc.

➢ **Real-time data.** They are produced on online streaming media, such as YouTube, Twitch, Skype, or Netflix.

➢ **Transactional data.** They are gathered when a user makes an online purchase (information on the product, time of purchase, payment methods, etc.)

➢ **Geographic data.** Location data of everything, humans, vehicles, building, natural reserves, and other objects are continuously supplied with satellites.

➢ **Natural language data.** These data are gathered mostly from voice searches that can be made on different devices accessing the Internet.

➢ **Time series data.** This type of data is related to the observation of trends and phenomena taking place at this very moment and over a period of time, for instance, global temperatures, mortality rates, pollution levels, etc.

➢ **Linked data.** They are based on HTTP, RDF, SPARQL, and URIs web technologies and meant to enable semantic connections between various databases so that computers could read and perform semantic queries correctly.

HOW IS BIG DATA COLLECTED?

There are different ways of how to collect big data from users. These are the most popular ones.

➢ **1. Asking for it**
the majority of firms prefer asking users directly to share their personal information. They give these data when creating website accounts or buying online. The minimum information to be collected includes a username and an email address, but some profiles require more details.

➢ **2. Cookies and Web Beacons**

Cookies and web beacons are two widely used methods to gather the data on users, namely, what web pages they visit and when. They provide basic statistics about how a website is used. Cookies and web beacons in no way compromise your privacy but just serve to personalize your experience with one or another web source.

➢ **3. Email tracking**

Email trackers are meant to give more information on the user actions in the mailbox. In particular, an email tracker allows detecting when an email was opened. Both Google and Yahoo use this method to learn their users' behavioural patterns and provide personalized advertising.

❖ **Importance of Data Quality**

Data quality is defined as:
"The degree to which data meets a company's expectations of accuracy, validity, completeness, and consistency"

By tracking <u>data quality</u>, a business can pinpoint potential issues harming quality, and ensure that shared data is fit to be used for a given purpose.

When collected data fails to meet the company expectations of accuracy, validity, completeness, and consistency, it can have massive negative impacts on customer service, employee productivity, and key strategies.

Quality data is key to making accurate, informed decisions. And while all data has some level of "quality," a variety of characteristics and factors determines the degree of data quality (high-quality versus low-quality). Furthermore, different data quality characteristics will likely be more important to various stakeholders across the organization.
A list of popular data quality characteristics and dimensions include:

1. Completeness: Completeness is defined as a measure of the percentage of data that is missing within a dataset.
2. Timeliness: Timeliness measures how up-to-date or antiquated the data is at any given moment.
3. Validity: Validity refers to information that fails to follow specific company formats, rules, or processes.
4. Integrity: Integrity of data refers to the level at which the information is reliable and trustworthy.
5. Uniqueness: Uniqueness is a data quality characteristic most often associated with customer profiles.
6. Consistency: It ensures that the source of the information collection is capturing the correct data based on the unique objectives of the department or company.

❖ **Dealing with Missing or incomplete Data**

The concept of missing data is implied in the name: its data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results.
Fortunately, there are proven techniques to deal with missing data.

Imputation vs. Removing Data

When dealing with missing data, <u>data scientists</u> can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Before deciding which approach to employ, data scientists must understand why the data is missing.

Missing at Random (MAR)

Missing at Random means the data is missing relative to the observed data. It is not related to the specific missing values. The data is not missing across all observations but only within sub-samples of the data. It is not known if the data should be there; instead, it is missing given the observed data. The missing data can be predicted based on the complete observed data.

Missing Completely at Random (MCAR)

In the MCAR situation, the data is missing across all observations regardless of the expected value or other variables. Data scientists can compare two sets of data, one with missing observations and one without. Using a t-test, if there is no difference between the two data sets, the data is characterized as MCAR.

Data may be missing due to test design, failure in the observations or failure in recording observations. This type of data is seen as MCAR because the reasons for its absence are external and not related to the value of the observation.

It is typically safe to remove MCAR data because the results will be unbiased. The test may not be as powerful, but the results will be reliable.

Missing Not at Random (MNAR)

The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing. In a survey, perhaps a specific group of people – say women ages 45 to 55 – did not answer a question. Like MAR, the data cannot be determined by the observed data, because the missing information is unknown. Data scientists must model the missing data to develop an unbiased estimate. Simply removing observations with missing data could result in a model with bias.

Deletion

There are two primary methods for deleting data when dealing with missing data: list wise and dropping variables.

List wise

In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, it may be the most efficient method to eliminate those cases from the analysis. However, in most cases, the data are not missing completely at random (MCAR). Deleting the instances with missing observations can result in biased parameters and estimates and reduce the statistical power of the analysis.

Pair wise

Pair wise deletion assumes data are missing completely at random (MCAR), but all the cases with data, even those with missing data, are used in the analysis. Pairwise deletion allows data scientists to use more of the data. However, the resulting statistics may vary because they are based on different data sets. The results may be impossible to duplicate with a complete set of data.

Dropping Variables

If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

❖ **Imputation**

When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.

Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

Mean, Median and Mode

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

Time-Series Specific Methods

Another option is to use time-series specific methods when appropriate to impute data. There are four types of time-series data:

- No trend or seasonality.
- Trend, but no seasonality.
- Seasonality, but no trend.
- Both trend and seasonality.

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these

methods won't always produce reasonable results, particularly in the case of strong seasonality.

Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)

These options are used to analyze longitudinal repeated measures data, in which follow-up observations may be missing. In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

Linear Interpolation

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

When dealing with missing data, you should use this method in a time series that exhibits a trend line, but it's not appropriate for seasonal data.

Seasonal Adjustment with Linear Interpolation

When dealing with data that exhibits both trend and seasonality characteristics, use seasonal adjustment with linear interpolation. First you would perform the seasonal adjustment by computing a centered moving average or taking the average of multiple averages – say, two one-year averages – that are offset by one period relative to another. You can then complete data smoothing with linear interpolation as discussed above.

Multiple Imputations

Multiple imputations is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.

The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

K Nearest Neighbours

In this method, data scientists choose a distance measure for k neighbours, and the average is used to impute an estimate. The data scientist must select the number of nearest neighbours and the distance metric. KNN can identify the most frequent value among the neighbours and the mean among the nearest neighbours.

- ❖ **Data Visualization**
  - 🞣 Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.

- The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.
- Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modelled, it must be visualized for conclusions to be made.
- Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.
- Data visualization is important for almost every career. It can be used by teachers to display student test results, by computer scientists exploring advancements in artificial intelligence (AI) or by executives looking to share information with stakeholders.
- It also plays an important role in big data projects. As businesses accumulated massive collections of data during the early years of the big data trend, they needed a way to quickly and easily get an overview of their data. Visualization tools were a natural fit.
- Visualization is central to advanced analytics for similar reasons. When a data scientist is writing advanced predictive analytics or machine learning (ML) algorithms, it becomes important to visualize the outputs to monitor results and ensure that models are performing as intended. This is because visualizations of complex algorithms are generally easier to interpret than numerical outputs.

Why is data visualization important?

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information. The practice can also help businesses identify which factors affect customer behaviour; pinpoint areas that need to be improved or need more attention; make data more memorable for stakeholders; understand when and where to place specific products; and predict sales volumes.

Other benefits of data visualization include the following:

- the ability to absorb information quickly, improve insights and make faster decisions;
- an increased understanding of the next steps that must be taken to improve the organization;
- an improved ability to maintain the audience's interest with information they can understand;
- an easy distribution of information that increases the opportunity to share insights with everyone involved;
- eliminate the need for data scientists since data is more accessible and understandable; and
- An increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes.

Data visualization and big data

- The increased popularity of big data and data analysis projects has made visualization more important than ever.
- Companies are increasingly using machine learning to gather massive amounts of data that can be difficult and slow to sort through, comprehend and explain.
- Visualization offers a means to speed this up and present information to business owners and stakeholders in ways they can understand.

- o Big data visualization often goes beyond the typical techniques used in normal visualization, such as pie charts, histograms and corporate graphs. It instead uses more complex representations, such as heat maps and fever charts.
- o Big data visualization requires powerful computer systems to collect raw data, process it and turn it into graphical representations that humans can use to quickly draw insights.

❖ **Examples of data visualization**

In the early days of visualization, the most common visualization technique was using a Microsoft Excel spreadsheet to transform the information into a table, bar graph or pie chart. While these visualization methods are still commonly used, more intricate techniques are now available, including the following:

- info graphics
- bubble clouds
- bullet graphs
- heat maps
- fever charts
- time series charts

Some other popular techniques are as follows.

Line charts. This is one of the most basic and common techniques used. Line charts display how variables can change over time.

Area charts. This visualization method is a variation of a line chart; it displays multiple values in a time series -- or a sequence of data collected at consecutive, equally spaced points in time.

Scatter plots. This technique displays the relationship between two variables. A scatter plot takes the form of an x- and y-axis with dots to represent data points.

Tree maps. This method shows hierarchical data in a nested format. The size of the rectangles used for each category is proportional to its percentage of the whole. Treemaps are best used when multiple categories are present, and the goal is to compare different parts of a whole.

Population pyramids. This technique uses a stacked bar graph to display the complex social narrative of a population. It is best used when trying to display the distribution of a population.

Data Visualization Applications

Common use cases for data visualization include the following:

Sales and Marketing: Research from the media agency Magna predicts that half of all global advertising dollars will be spent online by 2020. As a result, marketing teams must pay close attention to their sources of web traffic and how their web properties generate revenue. Data visualization makes it easy to see traffic trends over time as a result of marketing efforts.

Politics: A common use of data visualization in politics is a geographic map that displays the party each state or district voted for.

Healthcare: Healthcare professionals frequently use choropleth maps to visualize important health data. A choropleth map displays divided geographical areas or regions that are

assigned a certain color in relation to a numeric variable. Choropleth maps allow professionals to see how a variable, such as the mortality rate of heart disease, changes across specific territories.

Scientists: Scientific visualization, sometimes referred to in shorthand as SciVis, allows scientists and researchers to gain greater insight from their experimental data than ever before.

Finance: Finance professionals must track the performance of their investment decisions when choosing to buy or sell an asset. Candlestick charts are used as trading tools and help finance professionals analyze price movements over time, displaying important information, such as securities, derivatives, currencies, stocks, bonds and commodities. By analyzing how the price has changed over time, data analysts and finance professionals can detect trends.

Logistics: Shipping companies can use visualization tools to determine the best global shipping routes.

❖ **Data visualization tools and vendors**

Data visualization tools can be used in a variety of ways. The most common use today is as business intelligence (BI) reporting tool. Users can set up visualization tools to generate automatic dashboards that track company performance across key performance indicators (KPIs) and visually interpret the results.

The generated images may also include interactive capabilities, enabling users to manipulate them or look more closely into the data for questioning and analysis. Indicators designed to alert users when data has been updated or when predefined conditions occur can also be integrated.

Many business departments implement data visualization software to track their own initiatives. For example, a marketing team might implement the software to monitor the performance of an email campaign, tracking metrics like open rate, click-through rate and conversion rate.

As data visualization vendors extend the functionality of these tools, they are increasingly being used as front ends for more sophisticated big data environments. In this setting, data visualization software helps data engineers and scientists keep track of data sources and do basic exploratory analysis of data sets prior to or after more detailed advanced analyses.

The biggest names in the big data tools marketplace include Microsoft, IBM, SAP and SAS.

Some other vendors offer specialized big data visualization software; popular names in this market include **Tableau, Qlik and Tibco.**

While **Microsoft Excel continues to be a popular tool for data visualization,** others have been created that provide more sophisticated abilities:

- IBM Cognos Analytics
- Qlik Sense and QlikView
- Microsoft Power BI
- Oracle Visual Analyzer
- SAP Lumira
- SAS Visual Analytics
- Tibco Spotfire
- Zoho Analytics
- D3.js
- Jupyter

- MicroStrategy
- Google Charts

❖ Data Classification

- ✓ Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. On a basic level, the classification process makes data easier to locate and retrieve.
- ✓ Data classification is of particular importance when it comes to risk management, compliance, and data security.
- ✓ Data classification involves tagging data to make it easily searchable and traceable.
- ✓ It also eliminates multiple duplications of data, which can reduce storage and backup costs while speeding up the search process. Though the classification process may sound highly technical, it is a topic that should be understood by your organization's leadership.

Importance of Data Classification:

Data classification is a regulatory requirement, as data must be searchable and retrievable within specified timeframes.

For the purposes of data security, data classification is a useful tactic that facilitates proper security responses based on the type of data being retrieved, transmitted, or copied.

❖ Types of Data Classification

Data classification involves the use of tags and labels to define the data type, its confidentiality, and its integrity. There are three main types of data classification that are considered the industry standard:

- **Content-based classification** – inspects and interprets files, looking for sensitive information
- **Context-based classification** – looks to the application, location, metadata, or creator (among other variables) as indirect indicators of sensitive information
- **User-based classification** – requires a manual, end-user selection for each document. User-based classification takes advantage of the user knowledge of the sensitivity of the document, and can be applied or updated upon creation, edit, review, or dissemination.

DETERMINING DATA RISK

In addition to the types of classification, it's wise for an organization to determine the relative risk associated with the types of data, how that data is handled and where it is stored/sent (endpoints). A common practice is to separate data and systems into three levels of risk

Low risk: If data is public and it's not easy to permanently lose (e.g. recovery is easy), this data collection and the systems surrounding it are likely a lower risk than others.

Moderate risk: Essentially, this is data that isn't public or is used internally (by your organization and/or partners). However, it's also not likely too critical to operations or sensitive to be "high risk." Proprietary operating procedures cost of goods and some company documentation may fall into the moderate category.

High risk: Anything remotely sensitive or crucial to operational security goes into the high risk category. Also, pieces of data those are extremely hard to recover (if lost). All confidential, sensitive and necessary data falls into a high risk category.

❖ **Data Sensitivity Levels**

While we've looked at mapping data out by type, you should also look to segment your organization's data in terms of the level of sensitivity – high, moderate, or low.

- **High sensitivity data (Confidential)** – data that if compromised or destroyed would be expected to have a severe or catastrophic effect on organizational operations, assets, or individuals. Examples can include financial data, medical records, and intellectual property.
- **Moderate sensitivity data (Restricted)** – data that if compromised or destroyed would be expected to have a serious effect on organizational operations, assets, or individuals. Examples can include unpublished research results, information strictly for internal use, and operational documents.
- **Low sensitivity data (Public)** – data that if compromised or destroyed would be expected to have a limited effect on organizational operations, assets, or individuals. Examples can include press releases, job advertisements, and published research.

The following shows common examples of organizational data which may be classified into each sensitivity level:
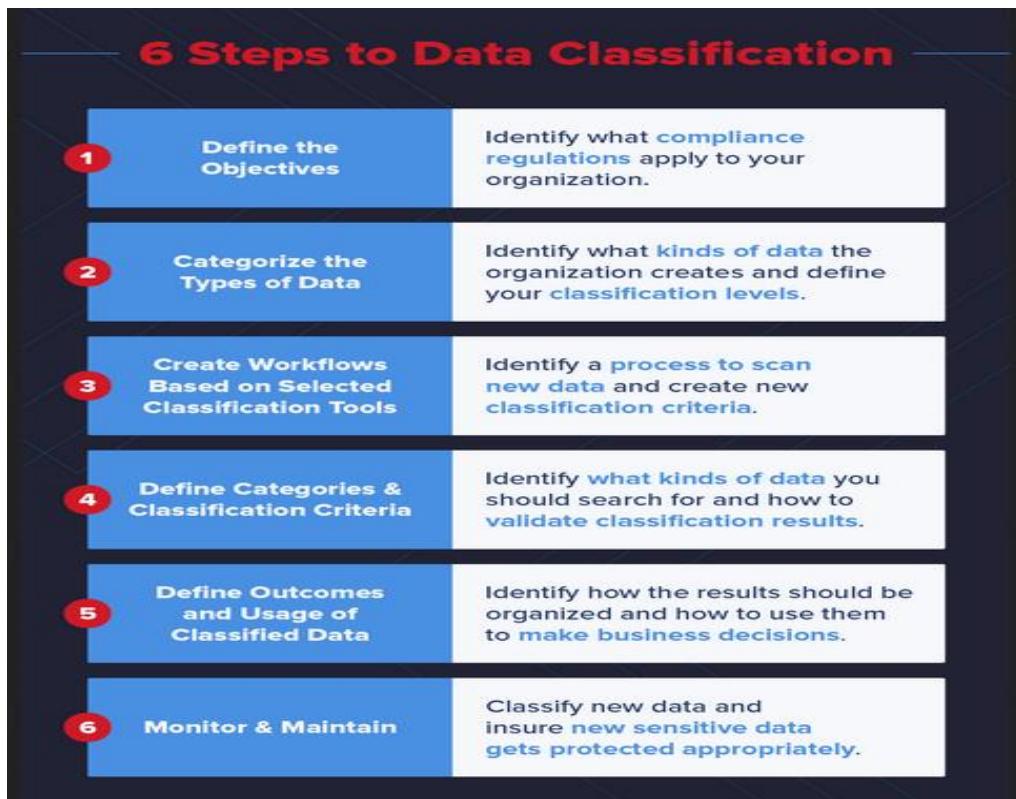
High:
- Personally identifiable information (PII)
- Credit card details (PCI)
- Intellectual property (IP)
- Protected healthcare information (including HIPAA regulated data)
- Financial information
- Employee records
- ITAR materials
- Internal correspondence including confidential data

Moderate:
- Student education records
- Unpublished research data
- Operational data
- Information security information
- Supplier contact information
- Internal correspondence not containing confidential data

Low:
- Public websites
- Public directory data
- Publicly available research
- Press releases
- Job advertisements
- Marketing materials

❖ Data Science Project Life Cycle :

Data Science is a multidisciplinary field that uses scientific methods to extract insights from structured and unstructured data. Data science is such a huge field and concept that's often intermingled with other disciplines, but generally, DS unifies statistics, data analysis, machine learning, and related fields.

Data Science life cycle provides the structure to the development of a data science project. The lifecycle outlines the major steps, from start to finish, that projects usually follow. Now, there are various approaches to managing DS projects, amongst which are **Cross-industry standard process for data mining (aka CRISP-DM),** process of knowledge discovery in databases (aka KDD), any proprietary-based custom procedures conjured up by a company, and a few other simplified processes.

CRISP-DM

CRISP-DM is an open standard process model that describes common approaches used by data mining scientists. In 2015, it was refined and extended by IBM, which released a new methodology called Analytics Solutions Unified Method for Data Mining/Predictive Analytics (aka ASUM-DM).

The CRISP-DM model steps are:
1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation and
6. Deployment

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data*	**Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*		**Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*					

Knowledge discovery in databases **(KDD)**

KDD is commonly defined with the following stages:
- ✓ Selection
- ✓ Pre-processing
- ✓ Transformation
- ✓ Data mining
- ✓ Interpretation/evaluation
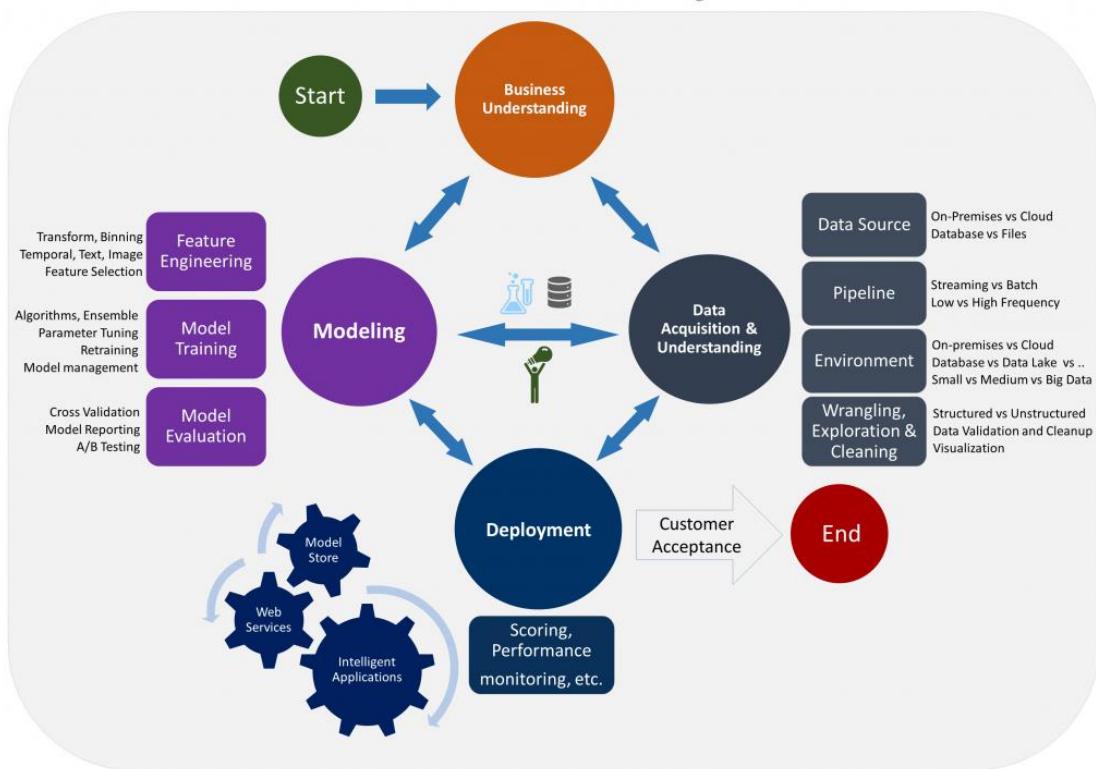
The **simplified process** looks as follows: (1) Pre-processing, (2) Data Mining, and (3) Results Validation.

Suppose, we have a standard DS project (without any industry-specific peculiarities), then the lifecycle would typically include:
- ✓ **Business understanding**
- ✓ **Data acquisition and understanding**
- ✓ **Modelling**
- ✓ **Deployment**
- ✓ **Customer acceptance**

Data Science Lifecycle

The DS project life cycle is an iterative process of research and discovery that provides guidance on the tasks needed to use predictive models. The goal of this process is to move a DS project to an engagement end-point by providing means for easier and clearer communication between teams and customers with a well-defined set of artifacts and standardized templates to homogenize procedures and avoid misunderstandings.

Each stage has the following information:

- Goals and specific objectives of the stage
- A clear outline of specific tasks and instructions on how to complete them
- The expected deliverables (artifact)

Business understanding

Before you even embark on a DS project, you need to understand the problem you're trying to solve and define the central objectives of your project by identifying the variables to predict.

Goals:
- ✓ Identify key variables that will serve as model targets and serve as the metrics for defining the success of the project
- ✓ Identify data sources that the business has already access to or need to obtain such access

Guidelines:
Work with customers and stakeholders to define business problems and formulate questions that data science needs to answer.
The goal here is to identify the key business variables (aka model targets) that your analysis needs to predict and the project's success would be assessed against. For example, the sales

forecasts. This is what needs to be predicted, and at the end of your project, you'll compare your predictions to the actual volume of sales.

Define project goals by asking specific questions related to data science, such as:

- How much/many? (regression)
- Which category? (classification)
- Which group? (clustering)
- Does this make sense? (anomaly detection)
- Which option should be taken? (recommendation)

❖ **Business Requirements**

✓ The purpose of business requirements is to define a project's business need, as well as the criteria of its success.

✓ Business requirements describe why a project is needed, whom it will benefit, when and where it will take place, and what standards will be used to evaluate it.

✓ Business requirement generally do not define *how* a project is to be implemented; the requirements of the business need do not encompass a project's implementation details.

✓ "Business requirements are higher-level statements of the goals, objectives, or needs of the enterprise."

✓ "They describe the reasons why a project has been initiated, the objectives that the project will achieve, and the metrics that will be used to measure its success."

✓ In short, business requirements chart where a project is going, not how it's going to get there.

The business requirements the analyst creates for this project would include (but not be limited to):

- **Identification of the business problem** (key objectives of the project), i.e., "Declining ticket sales require a strategy to increase the number of customers at our theatres."
- **Why the solution has been proposed** (its benefits; why it will produce the desired outcome of returning ticket sales to higher levels), i.e., "Customers have overwhelmingly cited the inconvenience of standing in line as the primary reason they no longer attend our theatre. We will remove this impediment by enabling customers to buy and print their theatre tickets at home with just a few clicks."
- **The scope of the project**. A few examples might be: "1. while the plan is to bring this project to all 400 theatres eventually, we will start with 50 theatres in the most populated metropolitan areas."
- **Rules, policies, and regulations**. For example, "We will design our web site and commerce so that all other relevant governmental regulations are properly adhered to."
- **Key features of the service** (without details as to how they will be implemented). A few examples might include: "1. we will provide a secure site for the user to select the number of tickets and showing they wish, and to enter their payment information. 2. We will give the user the option to store his or her card information in our system so that they do not have to re-enter it in a later session. 3. The system will accommodate credit, debit, or PayPal payment methods only."

- **Key performance features** (without details as to how they will be implemented), i.e., "1. The system will be designed so that it can recover within 30 seconds of any downtime. 2. Because our peak audience has been 25,000 customers in all of our theatres on one night, the system will accommodate at least 10 times that many users at any given time without any impact on system performance."
- **Key security features** (again without details), i.e., "We will devise a unique identifier for each ticket that will prohibit photocopies or counterfeits."
- Criteria to measure the project's success, such as: "This project will be deemed successful if ticket sales return to 2008 levels within 12 months of its launch."

This project's resulting business requirements would not include:

- A description of how to adhere to governmental or regulatory requirements.
- A description of how performance requirements will be implemented, such as: "The XYZ server on which customer information is stored will be backed up every five minutes using XYZ program."
- Any description of how the unique ticket identifier would be implemented.
- Any details or specifics related to the service's features, such as: "1. The credit card number text box will be 20 characters long and accommodate simple text. 2. If the user selects Yes (01), the information will be loaded to our XYZ storage server called."

While the above examples accompanying selected bullet points are textual, business requirements may include graphs, models, or any combination of these that best serves the project. Effective business requirements require strong strategic thinking, significant input from a project's business owners, and the ability to clearly state the needs of a project at a high level.

As with all <u>requirements</u>, business requirements should be:

- **Verifiable.** Just because business requirements state business needs rather than technical specifications doesn't mean they mustn't be demonstrable.
- Verifiable requirements are specific and objective. A quality control expert must be able to check, for example, that the system accommodates the debit, credit, and PayPal methods specified in the business requirements. (S)he could not do so if the requirements were more vague, i.e., "The system will accommodate appropriate payment methods." (*Appropriate* is subject to interpretation.)
- **Unambiguous**, stating precisely what problem is being solved. For example, "This project will be deemed successful if ticket sales increase sufficiently," is probably too vague for all stakeholders to agree on its meaning at the project's end.
- **Comprehensive,** covering every aspect of the business need. Business requirements are indeed big picture, but they are very thorough big picture. In the aforementioned example, if the analyst assumed that the developers would know to design a system that could accommodate many times the number of customers the theatre chain had seen at one time in the past, but did not explicitly state so in the requirements, the developers might design a system that could accommodate only 10,000 patrons at any one time without performance issues.

Remember that business requirements answer the what's, not the how's, but they are meticulously thorough in describing those's. No business point is overlooked. At a project's

end, the business requirements should serve as a methodical record of the initial business problem and the scope of its solution.

Understanding the project objectives and requirements from a domain perspective and then converting this knowledge into a data science problem definition with a preliminary plan designed to achieve the objectives. Data science projects are often structured around the specific needs of an industry sector (as shown below) or even tailored and built for a single organization. A successful data science project starts from a well defined question or need.

- ❖ **Data Acquisition**

  - ✓ **Data acquisition** (commonly abbreviated as **DAQ or DAS**) is the process of sampling signals that measure real-world physical phenomena and converting them into a digital form that can be manipulated by a computer and software.

  - ✓ Data Acquisition is generally accepted to be distinct from earlier forms of recording to tape recorders or paper charts. Unlike those methods, the signals are converted from the analog domain to the digital domain and then recorded to a digital medium such as ROM, flash media, or hard disk drives.

❖ **The Purposes of Data Acquisition**

The primary purpose of a data acquisition system is to acquire and store the data. But they are also intended to provide real-time and post-recording visualization and analysis of the data. Furthermore, most data acquisition systems have some analytical and report generation capability built-in.

Engineers in different applications have various requirements, of course, but these key capabilities are present in varying proportion:

- o Data recording
- o Data storing
- o Real-time data visualization
- o Post-recording data review
- o Data analysis using various mathematical and statistical calculations
- o Report generation

❖ **Data Preparation**
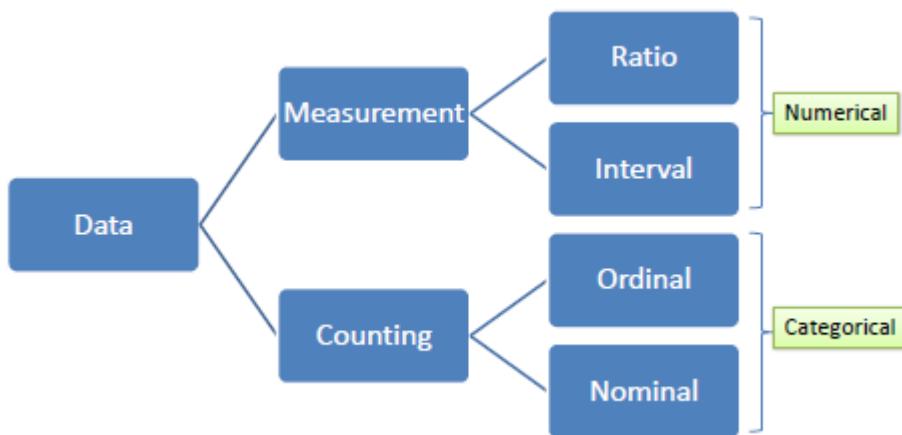
Data preparation is about constructing a dataset from one or more data sources to be used for exploration and modeling. It is a solid practice to start with an initial dataset to get familiar with the data, to discover first insights into the data and have a good understanding of any possible data quality issues. Data preparation is often a time consuming process and heavily prone to errors. The old saying "garbage-in-garbage-out" is particularly applicable to those data science projects where data gathered with many invalid, out-of-range and missing values. Analyzing data that has not been carefully screened for such problems can produce highly misleading results. Then, the success of data science projects heavily depends on the quality of the prepared data.

Data

Data is information typically the results of measurement (numerical) or counting (categorical). **Variables** serve as placeholders for data. There are two types of variables, *numerical* and *categorical*.

A **numerical** or **continuous variable** is one that can accept any value within a finite or infinite interval (e.g., height, weight, temperature, blood glucose,). There are two types of numerical data, *interval* and *ratio*. Data on an interval scale can be added and subtracted but cannot be meaningfully multiplied or divided because there is no true zero. For example, we cannot say that one day is twice as hot as another day. On the other hand, data on a ratio scale has true zero and can be added, subtracted, multiplied or divided (e.g., weight).

A **categorical** or **discrete variable** is one that can accept two or more values (categories). There are two types of categorical data, *nominal* and *ordinal*. Nominal data does not have an intrinsic ordering in the categories. For example, "gender" with two categories, male and female. In contrast, ordinal data does have an intrinsic ordering in the categories. For example, "level of energy" with three orderly categories (low, medium and high).

Dataset

Dataset is a collection of data, usually presented in a tabular form. Each column represents a particular variable, and each row corresponds to a given member of the data.

ID	Outlook	Temp	Humidity	Windy	Play Golf
1	Rainy	85	92	False	No
2	Rainy	80	88	True	No
3	Overcast	83	86	False	Yes
4	Sunny	70	80	False	Yes
5	Sunny	68	?	False	Yes
6	Sunny	65	58	True	No
7	Overcast	64	62	True	Yes
8	Rainy	72	95	?	No
9	Rainy	?	70	False	Yes
10	Sunny	75	72	False	Yes
11	Rainy	75	74	True	Yes
12	?	72	78	True	Yes
13	Overcast	81	66	False	Yes
14	Sunny	71	79	True	No

There are some alternatives for columns, rows and values.

- Columns, Fields, Attributes, Variables
- Rows, Records, Objects, Cases, Instances, Examples, Vectors

- Values, Data

In predictive modeling, **predictors** or **attributes** are the input variables and **target** or **class attribute** is the output variable whose value is determined by the values of the predictors and function of the predictive model.

Database

Database collects, stores and manages information so users can retrieve, add, update or remove such information. It presents information in tables with rows and columns. A table is referred to as a relation in the sense that it is a collection of objects of the same type (rows). Data in a table can be related according to common keys or concepts, and the ability to retrieve related data from related tables is the basis for the term relational database. A Database Management System (**DBMS**) handles the way data is stored, maintained, and retrieved. Most data science toolboxes connect to databases through **ODBC** (Open Database Connectivity) or **JDBC** (Java Database Connectivity) interfaces.

SQL (Structured Query Language) is a database computer language for managing and manipulating data in relational database management systems (RDBMS).

SQL Data Definition Language (**DDL**) permits database tables to be created, altered or deleted. We can also define indexes (keys), specify links between tables, and impose constraints between database tables.

- CREATE TABLE : creates a new table
- ALTER TABLE : alters a table
- DROP TABLE : deletes a table
- CREATE INDEX : creates an index
- DROP INDEX : deletes an index

SQL Data Manipulation Language (**DML**) is a language which enables users to access and manipulate data.

- SELECT : retrieval of data from the database
- INSERT INTO : insertion of new data into the database
- UPDATE : modification of data in the database
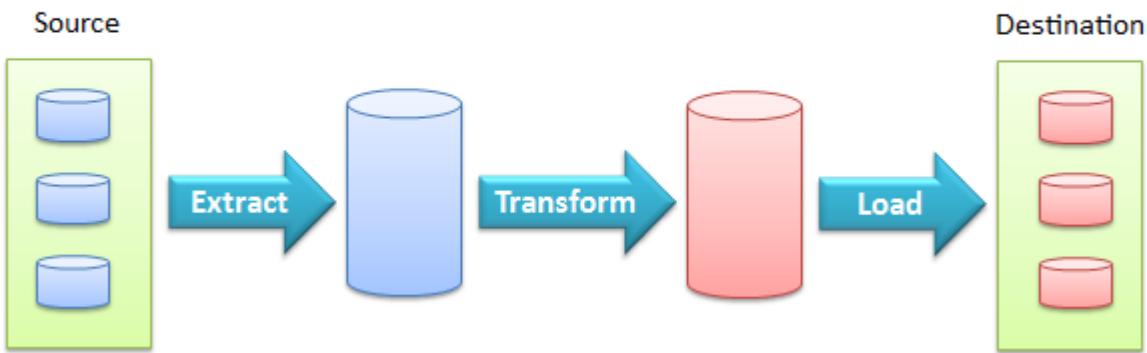
- DELETE : deletion of data in the database

ETL (Extraction, Transformation and Loading)

ETL extracts data from data sources and loads it into data destinations using a set of transformation functions.

- **Data extraction** provides the ability to extract data from a variety of data sources, such as flat files, relational databases, streaming data, XML files, and ODBC/JDBC data sources.
- **Data transformation** provides the ability to cleanse, convert, aggregate, merge, and split data.
- **Data loading** provides the ability to load data into destination databases via update, insert or delete statements, or in bulk.
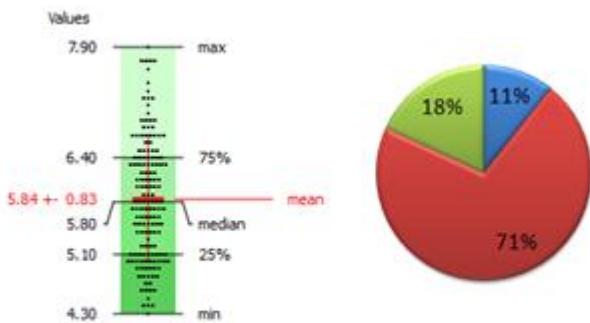
Credit Default Datasets
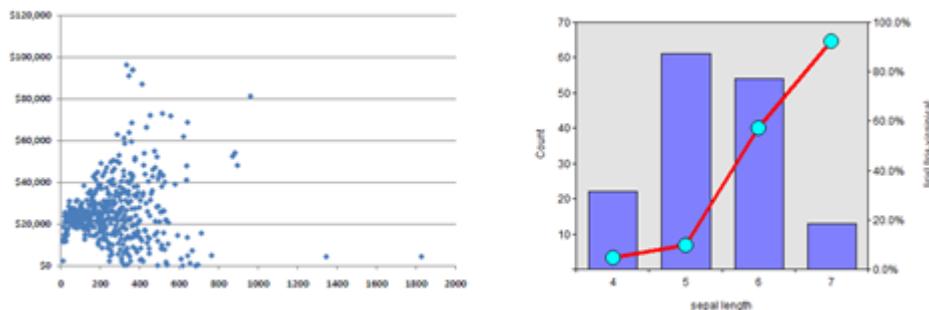
❖ **Data Exploration**

Data Exploration is about describing the data by means of statistical and visualization techniques. We explore data in order to bring important aspects of that data into focus for further analysis.

1. Univariate Analysis
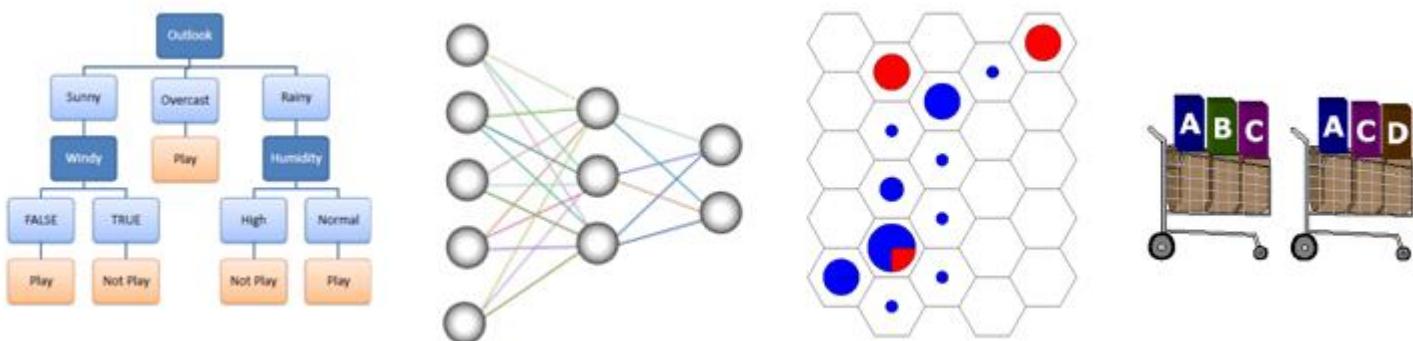
2. Bivariate Analysis

Modeling

Predictive modeling is the process by which a model is created to predict an outcome. If the outcome is categorical it is called classification and if the outcome is numerical it is called regression. Descriptive modeling or clustering is the assignment of observations into clusters so that observations in the same cluster are similar. Finally, association rules can find interesting associations amongst observations.

Model Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set (not

seen by the model) to evaluate model performance.

Hold-Out

In this method, the mostly large dataset is *randomly* divided to three subsets:

1. **Training set** is a subset of the dataset used to build predictive models.
2. **Validation set** is a subset of the dataset used to assess the performance of model built in the training phase. It provides a test platform for fine tuning model's parameters and selecting the best-performing model. Not all modeling algorithms need a validation set.
3. **Test set** or unseen examples are a subset of the dataset to assess the likely future performance of a model. If a model fit to the training set much better than it fits the test set, overfitting is probably the cause.

Cross-Validation

When only a limited amount of data is available, to achieve an unbiased estimate of the model performance we use k-fold cross-validation. In k-fold cross-validation, we divide the data into k subsets of equal size. We build models k times, each time leaving out one of the subsets from training and use it as the test set. If k equals the sample size, this is called "leave-one-out".

Model evaluation can be divided to two sections:
- Classification Evaluation
- Regression Evaluation

Model Deployment
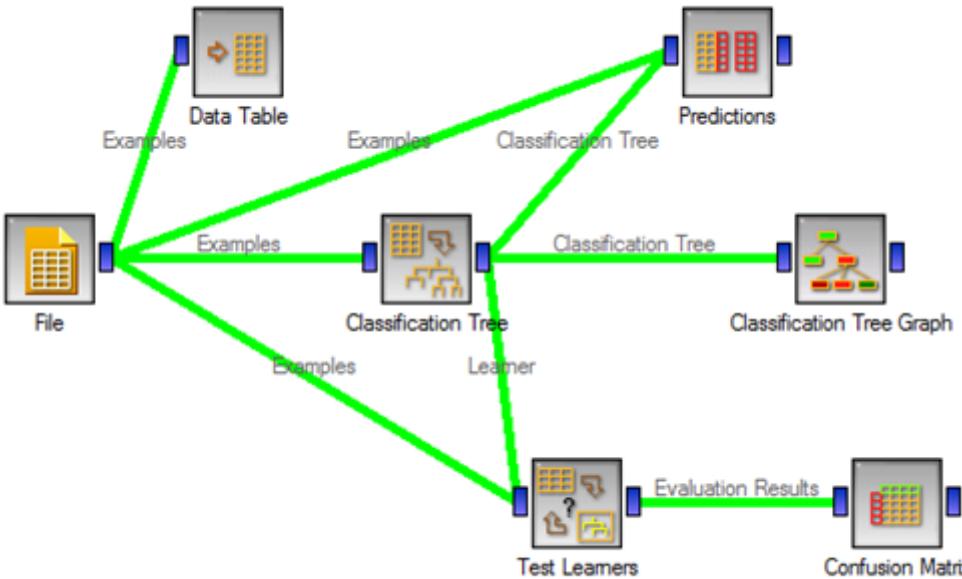
The concept of deployment in data science refers to the application of a model for prediction using a new data. Building a model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process. In many cases, it will be the customer, not the data analyst, who will carry out the deployment steps. For example, a credit card company may want to deploy a trained model or set of models (e.g., neural networks, meta-learner) to quickly identify transactions, which have a high probability of being fraudulent. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models.

Model deployment methods:

In general, there is four way of deploying the models in data science.

1. Data science tools (or cloud)
2. Programming language (Java, C, VB, …)
3. Database and SQL script (TSQL, PL-SQL, …)
4. PMML (Predictive Model Markup Language)
5.

An example of using a data mining tool (Orange) to deploy a decision tree model.

❖ **Operations Research**

Generally, OR is concerned with obtaining extreme values of some real-world objective functions; maximum (profit, performance, utility, or yield), minimum (loss, risk, distance, or cost). It incorporates techniques from mathematical modelling, optimization, and statistical analysis while emphasising the human-technology interface. However, one of the difficulties in answering this question is that there is a lot of overlap in scientific terminology — and sometimes terms become extremely popular, affecting the landscape of the terminology. E.g. the popularity of vague, broad terms such as AI and Big Data that works good for marketing but does nothing for the discussion on the research. Therefore, I have tried illustrating it in terms of ORs related fields, subfields, and the addressed problems

Process optimization is an exercise that aims to streamline operations within a project process, maximizing resource use and improving overall output. It is a significant element of business decision-making and is used in many different project management areas.

Process optimization methods and techniques

There are many process optimization techniques you can use to get you started. Here are three examples:

Process mining: This is a group of techniques with a data science approach. Data is taken from event logs to analyze what team members are doing in a company and what steps they

take to complete a <u>task</u>. This data can then be turned into insights, helping project managers to spot any roadblocks and optimize their processes.

DMAIC: DMAIC is a data-focused method used in Six Sigma to improve processes. It stands for Define, Measure, Analyze, Improve, and Control. These five stages combine to form a cycle. First, customers are defined. Then, performance is measured, and the data is analyzed. Finally, improvements are implemented and controlled to ensure the process remains in optimal condition.

PDSA: PDSA is an acronym for Plan, Do, Study, Act. It uses a four-stage cyclical model to improve quality and optimize business processes. Project managers will start by mapping what achievements they want to accomplish. Next, they will test proposed changes on a small scale. After this, they will study the results and determine if these changes were effective. If so, they will implement the changes across the entire business process.

It's good practice for a project manager to take some time to research various process optimization methods before deciding which one is most suited to their business

Major Applications of Data Science

Data Science is the deep study of a large quantity of data, which involves extracting some meaningful from the raw, structured, and unstructured data. The extracting out meaningful data from large amounts use processing of data and this processing can be done using statistical techniques and algorithm, scientific techniques, different technologies, etc. It uses various tools and techniques to extract meaningful data from raw data. Data Science is also known as the **Future of Artificial Intelligence**.

1. In Search Engines
The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.

For Example, When we search something suppose "Data Structure and algorithm courses " then at that time on the Internet Explorer we get the first link of GeeksforGeeks Courses. This happens because the GeeksforGeeks website is visited most in order to get information regarding Data Structure courses and Computer related subjects. So this analysis is Done using Data Science, and we get the Topmost visited Web Links.
2. In Transport
Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

For Example, In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what is the speed limit in Highway, Busy Streets, Narrow Roads, etc. And how to handle different situations while driving etc.

3. In Finance

Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

For Example, In Stock Market, Data Science is the main part. In the Stock Market, Data Science is used to examine past behavior with past data and their goal is to examine the future outcome. Data is analyzed in such a way that it makes it possible to predict future stock prices over a set timetable.

4. In E-Commerce

E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.

For Example, When we search for something on the E-commerce websites we get suggestions similar to choices according to our past data and also we get recommendations according to most buy the product, most rated, most searched, etc. This is all done with the help of Data Science.

5. In Health Care

In the Healthcare Industry data science act as a boon. Data Science is used for:

- Detecting Tumor.
- Drug discoveries.
- Medical Image Analysis.
- Virtual Medical Bots.
- Genetics and Genomics.
- Predictive Modeling for Diagnosis etc.

6. Image Recognition

currently, Data Science is also used in Image Recognition. **For Example,** When we upload our image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science. When an Image is Recognized, the data analysis is done on one's Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.

7. Targeting Recommendation

Targeting Recommendation is the most important application of Data Science. Whatever the user searches on the Internet, he/she will see numerous posts everywhere. This can be explained properly with an example: Suppose I want a mobile phone, so I just Google search it and after that, I changed my mind to buy offline. Data Science helps those companies who are paying for Advertisements for their mobile. So everywhere on the internet in the social media, in the websites, in the apps everywhere I will see the recommendation of that mobile phone which I searched for. So this will force me to buy online.

8. Airline Routing Planning

With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.

9. Data Science in Gaming

In most of the games where a user will play with an opponent i.e. a Computer Opponent, data science concepts are used with machine learning where with the help of past data the Computer will improve its performance. There are many games like Chess, EA Sports, etc. will use Data Science concepts.

10. Medicine and Drug Development

The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors. The algorithms based on data science will forecast how this will react to the human body without lab experiments.

11. In Delivery Logistics

Various Logistics companies like DHL, FedEx, etc. make use of Data Science. Data Science helps these companies to find the best route for the Shipment of their Products, the best time suited for delivery, the best mode of transport to reach the destination, etc.

12. Auto complete

AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line. In Google Mail, when we are writing formal mail to someone so at that time data science concept of Auto complete feature is used where he/she is an efficient choice to auto-complete the whole line. Also in Search Engines in social media, in various apps, AutoComplete feature is widely used.
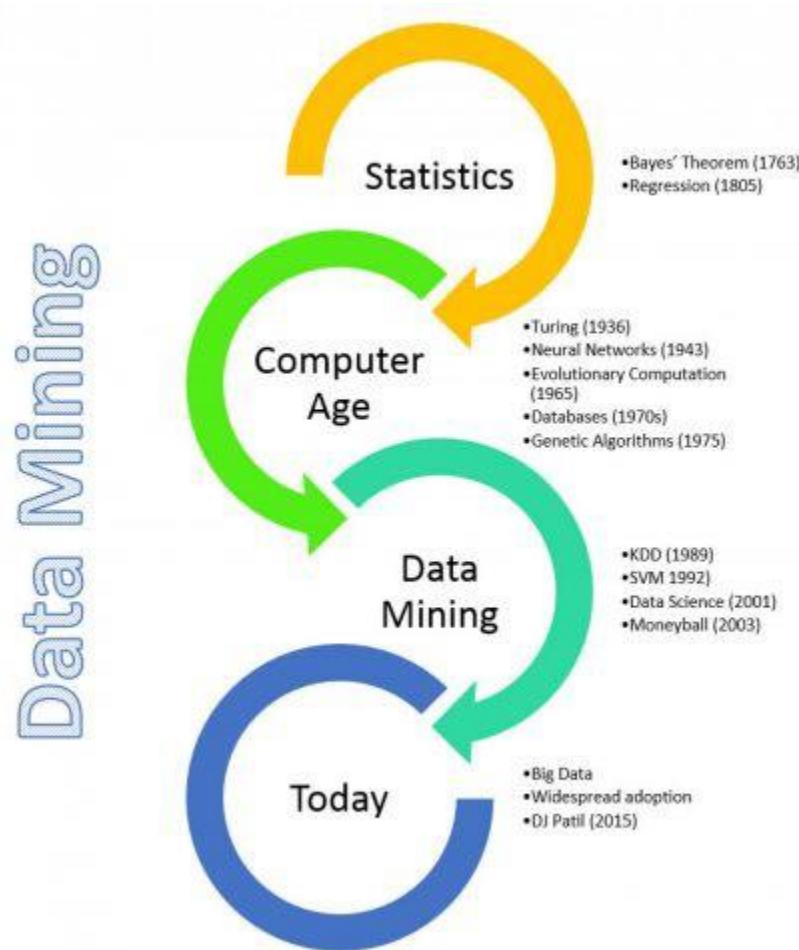
Data Mining: The origins of Data Mining- Data Mining Tasks- OLAP and Multidimensional-Data Analysis-Basic Concept of Association Analysis and Cluster Analysis.
Machine Learning: History and Evolution – AI Evolution- Statistics vs. Data Mining vs. Data Analytics vs. Data Science – Supervised Learning- Unsupervised Learning- Reinforcement Learning- Frameworks for Building Machine Learning Systems.

Data Mining:
The origins of Data Mining

- ✓ Data mining is a discipline with a long history. It starts with the early Data Mining methods Bayes' Theorem (1700`s) and Regression analysis (1800`s) which were mostly identifying patterns in data.
- ✓ Data mining is the process of analyzing large data sets (Big Data) from different perspectives and uncovering correlations and patterns to summarize them into useful information.
- ✓ Nowadays it is blended with many techniques such as artificial intelligence, statistics, data science, database theory and machine learning.
- ✓ Increasing power of technology and complexity of data sets has lead Data Mining to evolve from static data delivery to more dynamic and proactive information deliveries; from tapes and disks to advanced algorithms and massive databases.

- ✓ In the late 80`s Data Mining term began to be known and used within the research community by statisticians, data analysts, and the management information systems (MIS) communities.
- ✓ By the early 1990`s, data mining was recognized as a sub-process or a step within a larger process called Knowledge Discovery in Databases (KDD) .The most commonly used definition of KDD is "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad, 1996).

The sub-processes that form part of the KDD process are;

1. Understanding of the application and identifying the goal of the KDD process

2. Creating a target data set

3. Data cleaning and pre-processing

4. Matching the goals of the KDD process (step 1) to a particular data-mining method.

5. Research analysis and hypothesis selection

6. Data mining: Searching for patterns of interest in a particular form , including classification rules, regression, and clustering

7. Interpreting mined patterns

8. Acting on the discovered analysis

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What`s likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Data Mining Tasks:

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. Given the evolution of data warehousing technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades, assisting companies by transforming their raw data into useful knowledge.

Data mining functionalities are to perceive the various forms of patterns to be identified in data mining activities. To define the type of patterns to be discovered in data mining activities, data mining features are used. Data mining has a wide application for forecasting and characterizing data in big data.

Data mining tasks are majorly categorized into two categories: descriptive and predictive.

1. **Descriptive data mining**:
   Descriptive data mining offers a detailed description of the data, for example- it gives insight into what's going on inside the data without any prior idea. This demonstrates the common characteristics in the results. It includes any information to grasp what's going on in the data without a prior idea.
2. **Predictive Data Mining**:
   This allows users to consider features that are not specifically available. For example, the projection of the market analysis in the next quarters with the output of the previous quarters, In general, the predictive analysis forecasts or infers the features of the data previously available. For an instance: judging by the outcomes of medical records of a patient who suffers from some real illness.

Key Data Mining Tasks

1) Characterization and Discrimination

- **Data Characterization**: The characterization of data is a description of the general characteristics of objects in a target class which creates what are called characteristic rules.
  A database query usually computes the data applicable to a user-specified class and runs through a description component to retrieve the meaning of the data at various abstraction   levels.
  Eg;-Bar maps, curves, and pie charts.
- **Data Discrimination**: Data discrimination creates a series of rules called discriminate rules that is simply a distinction between the two classes aligned with the goal class and the opposite class of the general characteristics of objects.

2) Prediction

To detect the inaccessible data, it uses regression analysis and detects the missing numeric values in the data. If the class mark is absent, so classification is used to render the prediction. Due to its relevance in business intelligence, the prediction is common. If the class mark is absent, so the prediction is performed using classification. There are two methods of predicting data. Due to its relevance in business intelligence, a prediction is common. The prediction of the class mark using the previously developed class model and the prediction of incomplete or incomplete data using prediction analysis are two ways of predicting data.

3) Classification

Classification is used to create data structures of predefined classes, as the model is used to classify new instances whose classification is not understood. The instances used to produce the model are known as data from preparation. A decision tree or set of classification rules is based on such a form of classification process that can be collected to identify future details, for example by classifying the possible compensation of the employee based on the classification of salaries of related employees in the company.

4) Association Analysis

The link between the data and the rules that bind them is discovered. And two or more data attributes are associated. It associates qualities that are transacted together regularly. They work out what are called the rules of partnerships that are commonly used in the study of stock baskets. To link the attributes, there are two elements. One is the trust that suggests the possibility of both associated together, and another helps, which informs of associations' past occurrence.

5) Outlier Analysis

Data components that cannot be clustered into a given class or cluster are outliers. They are often referred to as anomalies or surprises and are also very important to remember.

Although in some contexts, outliers can be called noise and discarded, they can disclose useful information in other areas, and hence can be very important and beneficial for their study.

6) Cluster Analysis

Clustering is the arrangement of data in groups. Unlike classification, however, class labels are undefined in clustering and it is up to the clustering algorithm to find suitable classes. Clustering is often called unsupervised classification since provided class labels do not execute the classification. Many clustering methods are all based on the concept of maximizing the similarity (intra-class similarity) between objects of the same class and decreasing the similarity between objects in different classes (inter-class similarity).

7) Evolution & Deviation Analysis

We may uncover patterns and shifts in actions over time, with such distinct analysis, we can find features such as time-series results, periodicity, and similarities in patterns. Many technologies from space science to retail marketing can be found holistically in data processing and features.

❖ **OLAP (online analytical processing)**

- OLAP (online analytical processing) is a computing method that enables users to easily and selectively extract and query data in order to analyze it from different points of view. OLAP business intelligence queries often aid in trends analysis, financial reporting, sales forecasting, budgeting and other planning purposes.

- For example, a user can request that data be analyzed to display a spreadsheet showing all of a company's beach ball products sold in Florida in the month of July, compare revenue figures with those for the same products in September and then see a comparison of other product sales in Florida in the same time period.

How OLAP systems work

✓ To facilitate this kind of analysis, data is collected from multiple data sources and stored in <u>data warehouses</u> then cleansed and organized into data cubes.

✓ Each <u>OLAP cube</u> contains data categorized by dimensions (such as customers, geographic sales region and time period) derived by <u>dimensional tables</u> in the data warehouses.

✓ Dimensions are then populated by members (such as customer names, countries and months) that are organized <u>hierarchically</u>. OLAP cubes are often pre-summarized across dimensions to drastically improve query time over <u>relational databases</u>.

❖ **Types of OLAP Servers**

We have four types of OLAP servers −

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

1. Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following −

- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

2. Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

3. Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

4. Specialized SQL Servers

Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

❖ **OLAP Operations**

Since OLAP servers are based on multidimensional view of data, OLAP operations in **multidimensional data.**

Here is the list of OLAP operations −
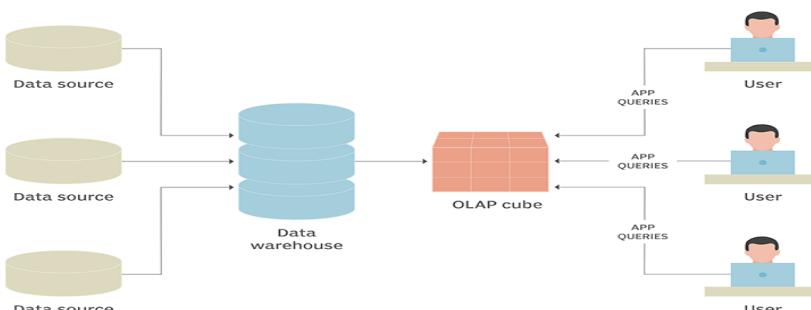
- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

- **Roll-up.** Also known as *consolidation*, or *drill-up*, this operation summarizes the data along the dimension.

- **Drill-down.** This allows analysts to navigate deeper among the dimensions of data, for example drilling down from "time period" to "years" and "months" to chart sales growth for a product.

- **Slice.** This enables an analyst to take one level of information for display, such as "sales in 2017."

- **Dice.** This allows an analyst to select data from multiple dimensions to analyze, such as "sales of blue beach balls in Iowa in 2017."

- **Pivot.** Analysts can gain a new view of data by rotating the data axes of the cube.

OLAP software then locates the intersection of dimensions, such as all products sold in the Eastern region above a certain price during a certain time period, and displays them. The result is the "measure"; each OLAP cube has at least one to perhaps hundreds of measures, which are derived from information stored in fact tables in the data warehouse.

OLAP begins with data accumulated from multiple sources and stored in a data warehouse. The data is then cleansed and stored in OLAP cubes, which users run queries against.

The OLAP process

How data is prepared for online analytical processing (OLAP)

❖ Association Rule Mining

Association analysis is useful for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of **association rules** or sets of frequent items.

Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market- Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Implication means co-occurrence, not causality!
Example of Association Rules
{Beer}→{Diaper}
{Milk, Bread}→{Eggs, Coke}
{Beer, Bread}→{Milk}

Support Count (   ) – Frequency of occurrence of a item set.
   Here   ({Milk, Bread, Diaper})=2

Frequent Item set – An item set whose support is greater than or equal to minsup threshold.
Association Rule – An implication expression of the form X -> Y, where X and Y are any 2 item sets.
Example: {Milk, Diaper}->{Beer}

Rule Evaluation Metrics –
   Support(s)   –
   The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction. It is a measure of how frequently the collection of items occurs together as a percentage of all transactions.
   Support   =   (X+Y) ÷   total   –
   It is interpreted as fraction of transactions that contain both X and Y.
   Confidence(c)   –
   It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
   Conf(X=>Y)   =   Supp(X Y) ÷   Supp(X)   –
   It measures how often each item in Y appears in transactions that contains items in X also.
   - **Lift   (l)   –**
     The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the item sets X and Y are independent of each other. The expected confidence is the confidence divided by the frequency of {Y}.
   - **Lift(X=>Y)   =   Conf(X=>Y)   ÷   Supp(Y)   –**
     Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1

means they appear less than expected. Greater lift values indicate stronger association.

- **Example** – From the above table, {Milk, Diaper}=>{Beer}

  s=   ({Milk, Diaper, Beer}) ÷|T|

  = 2/5

  = 0.4

- c=   (Milk, Diaper, Beer) ÷(Milk, Diaper)

  = 2/3

  = 0.67

- l= Supp({Milk, Diaper, Beer}) ÷Supp({Milk, Diaper})*Supp({Beer})

  = 0.4/ (0.6*0.6)

  = 1.11

The Association rule is very useful in analyzing datasets. The data is collected using bar-code scanners in supermarkets. Such databases consist of a large number of transaction records which list all items bought by a customer on a single purchase. So the manager could know if certain groups of items are consistently purchased together and use this data for adjusting store layouts, cross-selling, promotions based on statistics.

❖ Data Mining – Cluster Analysis

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

Cluster:

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles is given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

Now our task is to convert the unlabelled data to labelled data and it can be done using clusters.

The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc.

Simply it is partitioning of similar objects which are applied on unlabelled data.

Properties of Clustering:

1. **Clustering Scalability:** Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable if it is not scalable, then we can't get the appropriate result and would lead to wrong results.

2. **High Dimensionality:** The algorithm should be able to handle high dimensional space along with the data of small size.

3. **Algorithm Usability with multiple data kinds:** Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

4. **Dealing with unstructured data:** These would be some databases that contain missing values, noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle

❖ **Clustering Methods:**

The clustering methods can be classified into the following categories:

- o Partitioning Method
- o Hierarchical Method
- o Density-based Method
- o Grid-Based Method
- o Model-Based Method
- o Constraint-based Method

1. **Partitioning Method:** It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

- o One objective should only belong to only one group.
- o There should be no group without even a single purpose.

In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning.

2. **Hierarchical Method:** In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

- ▪ **Agglomerative Approach:** The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided in which the objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.

- **Divisive Approach:** The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

- One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.
- One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into micro clusters, macro clustering is performed on the micro cluster.

3. **Density-Based Method:** The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e, for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.

4. **Grid-Based Method:** In the Grid-Based method a grid is formed using the object together,i.e, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

5. **Model-Based Method:** In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore it yields robust clustering methods.

6. **Constraint-Based Method:** The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

Applications of Cluster Analysis:

- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies, identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

❖ Machine Learning

- ✓ **Machine learning** (**ML**) is the study of computer [algorithms](#) that can improve automatically through experience and by the use of data. It is seen as a part of [artificial intelligence](#). Machine learning algorithms build a model based on sample data, known as [training data](#), in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, [email filtering](#), [speech recognition](#), and [computer vision](#), where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

- ✓ A subset of machine learning is closely related to [computational statistics](#), which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of [mathematical optimization](#) delivers methods, theory and application domains to the field of machine learning. [Data mining](#) is a related field of study, focusing on [exploratory data analysis](#) through [unsupervised learning](#). Some implementations of machine learning use data and [neural networks](#) in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as [predictive analytics](#).

❖ How machine learning works

1. **A Decision Process**: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabeled, your algorithm will produce an estimate about a pattern in the data.

2. **An Error Function**: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.

3. **A Model Optimization Process**: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

❖ Machine learning methods

Machine learning classifiers fall into three primary categories.

1. Supervised machine learning

Supervised learning, also known as supervised machine learning, is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids over fitting or under fitting. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes,

linear regression, logistic regression, random forest, support vector machine (SVM), and more.

2. Unsupervised machine learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, image and pattern recognition. It's also used to reduce the number of features in a model through the process of dimensionality reduction; principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more.

3. Semi-supervised learning

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labelled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of having not enough labelled data (or not being able to afford to label enough data) to train a supervised learning algorithm.

❖ Reinforcement machine learning

Reinforcement machine learning is a behavioural machine learning model that is similar to supervised learning, but the algorithm isn't trained using sample data. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem.

Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data and information. In machine learning computers don't have to be explicitly programmed but can change and improve their algorithms by themselves.

Machine learning algorithms enable computers to communicate with humans, autonomously drive cars, write and publish sport match reports, and find terrorist suspects.
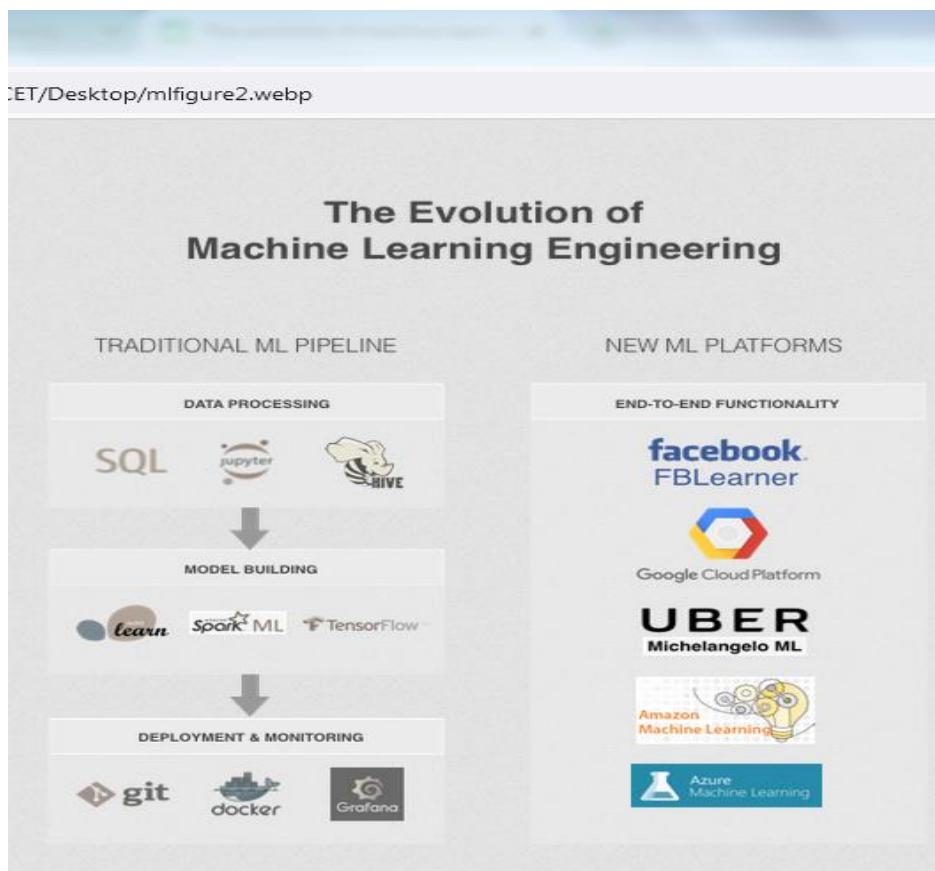
❖ The origins of machine learning

- ❖ The term *machine learning* was coined in 1959 by Arthur Samuel, an American IBMer and pioneer in the field of computer gaming and artificial intelligence. Also the synonym *self-teaching computers* was used in this time period. A representative book of the machine learning research during the 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973. In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal.

- ❖ Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms.

This follows <u>Alan Turing</u>'s proposal in his paper "<u>Computing Machinery and Intelligence</u>", in which the question "Can machines think?" is replaced with the question "Can machines do what we (as thinking entities) can do?".

❖ Modern day machine learning has two objectives, one is to classify data based on models which have been developed, and the other purpose is to make predictions for future outcomes based on these models. A hypothetical algorithm specific to classifying data may use computer vision of moles coupled with supervised learning in order to train it to classify the cancerous moles. A machine learning algorithm for stock trading may inform the trader of future potential predictions.

EVOLUTION OF MACHINE LEARNING:

Machine Learning Frameworks

A Machine Learning Framework is an interface, library or tool which allows developers to build <u>machine learning models</u> easily, without getting into the depth of the underlying algorithms. Let's discuss the Top 10 Machine Learning Frameworks in detail:

TensorFlow

Google's <u>Tensorflow</u> is one of the most popular frameworks today. It is an open-source software library for numerical computation using data flow graphs. TensorFlow implements data flow graphs, where batches of data or tensors can be processed by a series of algorithms described by a graph.

Theano

Theano is wonderfully folded over Keras, an abnormal state neural systems library, that runs nearly in parallel with the Theano library. Keras' fundamental favorable position is that it is a moderate Python library for profound discovering that can keep running over Theano or TensorFlow.

It was created to make actualizing profound learning models as quick and simple as feasible for innovative work. Discharged under the tolerant MIT permit, it keeps running on Python 2.7 or 3.5 and can consistently execute on GPUs and CPUs given the basic structures.

Sci-Kit Learn

Scikit-learn is one of the most well-known ML libraries. It is preferable for administered and unsupervised learning calculations. Precedents implement direct and calculated relapses, choice trees, bunching, k-implies, etc.

This framework involves a lot of calculations for regular AI and data mining assignments, including bunching, relapse, and order.

Caffe

Caffe is another popular learning structure made with articulation, speed, and measured quality as the utmost priority. It is created by the Berkeley Vision and Learning Center (BVLC) and by network donors.

Google's DeepDream depends on Caffe Framework. This structure is a BSD-authorized C++ library with Python Interface.

H20

H20 is an open-source machine learning platform. It is an artificial intelligence tool which is business-oriented and helps in making a decision based on data and enables the user to draw insights. It is mostly used for predictive modeling, risk and fraud analysis, insurance analytics, advertising technology, healthcare, and customer intelligence.

Amazon Machine Learning

Amazon Machine Learning provides visualization tools that help you go through the process of creating machine learning (ML) models without having to learn complex ML algorithms and technology.

It is a service that makes it easy for developers of all skill levels to use machine learning technology. It connects to data stored in Amazon S3, Redshift, or RDS, and can run binary classification, multiclass categorization, or regression on the data to build a model.

Torch

This framework provides wide support for machine learning algorithms to GPUs first. It is easy to use and efficient because of the easy and fast scripting language, **LuaJIT**, and an underlying **C/CUDA** implementation.

The goal of Torch is to have maximum flexibility and speed in building your scientific algorithms along with an extremely simple process.

Google Cloud ML Engine

Cloud Machine Learning Engine is a managed service that helps developers and data scientists in building and running superior machine learning models in production.

It offers training and prediction services that can be used together or individually. It is used by enterprises to solve problems like ensuring food safety, clouds in satellite images, responding four times faster to customer emails, etc.

Azure ML Studio

This Framework allows Microsoft Azure users to create and train models, then turn them into APIs that can be consumed by other services. Also, you can connect your own Azure storage to the service for larger models.

To use the Azure ML Studio, you don't even need an account to try out the service. You can log in anonymously and use Azure ML Studio for up to eight hours.

Spark ML Lib

This is Apache Spark's machine learning library. The goal of this framework is to make practical machine learning scalable and easy.

❖ **A Brief History of Artificial Intelligence**

The beginnings of modern AI can be traced to classical philosophers' attempts to describe human thinking as a symbolic system. But the field of AI wasn't formally founded until 1956, at a conference at Dartmouth College, in Hanover, New Hampshire, where the term "artificial intelligence" was coined.

The beginnings of modern AI can be traced to classical philosophers' attempts to describe human thinking as a symbolic system. But the field of AI wasn't formally founded until 1956, at a conference at Dartmouth College, in Hanover, New Hampshire, where the term "artificial intelligence" was coined.

Despite artificial intelligence has been present for millennia, it was not until the 1950s that its real potential was investigated. A generation of scientists, physicists, and intellectuals had the idea of AI, but it wasn't until Alan Turing, a British polymath, proposed that people solve problems and make decisions using available information and also a reason.

The difficulty of computers was the major stumbling block to expansion. They needed to adapt fundamentally before they could expand any further. Machines could execute orders but not store them. Until 1974, financing was also a problem.

By 1974, computers had become extremely popular. They were now quicker, less expensive, and capable of storing more data.

❖ **AI Research Today**
AI research is ongoing and expanding in today's world. AI research has grown at a pace of 12.9 percent annually over the last five years, as per Alice Bonasio, a technology journalist.

China is expected to overtake the United States as the world's leading source of AI technology in the next 4 years, having overtaken the United States' second position in 2004 and is rapidly closing in on Europe's top rank.

In the area of artificial intelligence development, Europe is the largest and most diverse continent, with significant levels of international collaboration. India is the 3rd largest country in AI research output, behind China and the USA.

❖ AI in the Present

Artificial intelligence is being utilized for so many things and has so much promise that it's difficult to imagine our future without it, related to business.

Artificial intelligence technologies are boosting productivity like never seen before, from workflow management solutions to trend forecasts and even the way companies buy advertisements.

Artificial Intelligence can gather and organize vast volumes of data in order to draw inferences and estimates that are outside of the human ability to comprehend manually. It also improves organizational efficiency while lowering the risk of a mistake, and it identifies unusual patterns, such as spam and frauds, instantaneously to alert organizations about suspicious behaviour, among other things. AI has grown in importance and sophistication to the point that a Japanese investment firm became the first to propose an AI Board Member for its ability to forecast market trends faster than humans.

Artificial intelligence will indeed be and is already being used in many aspects of life, such as self-driving cars in the coming years, more precise weather forecasting, and earlier health diagnoses, to mention a few.

❖ AI in The Future

It has been suggested that we are on the verge of the 4th Industrial Revolution, which will be unlike any of the previous three. From steam and water power through electricity and manufacturing process, computerization, and now, the question of what it is to be human is being challenged.

Smarter technology in our factories and workplaces, as well as linked equipment that will communicate, view the entire production process, and make autonomous choices, are just a few of the methods the Industrial Revolution will lead to business improvements. One of the most significant benefits of the 4th Industrial Revolution is the ability to improve the world's populace's quality of life and increase income levels. As robots, humans, and smart devices work on improving supply chains and warehousing, our businesses and organizations are becoming "smarter" and more productive.

❖ AI in Different Industries

Artificial intelligence (AI) may help you enhance the value of your company in a variety of ways. It may help you optimize your operations, increase total revenue, and focus your staff on more essential duties if applied correctly. As a result, AI is being utilized in a variety of industries throughout the world, including health care, finance, manufacturing, and others.

Health Care

AI is proven to be uplift in the healthcare business. It's enhancing nearly every area of the industry, from data security to robot-assisted operations. AI is finally providing this sector, which has been harmed by inefficient procedures and growing prices, a much-needed facelift.

Automotive

Self-driving vehicles are certainly something you've heard of, and they're a hint that the future is almost here. It's no longer science fiction; the autonomous car is already a reality. As per recent projections, by 2040, roughly 33 million automobiles with self-driving capability are projected to be on the road.

Finance

According to experts, the banking industry and AI are a perfect combination. Real-time data transmission, accuracy, and large-scale data processing are the most important elements driving the financial sector. Because AI is ideal for these tasks, the banking industry is recognizing its effectiveness and precision and incorporating machine learning, statistical arbitrage, adaptive cognition, chatbots, and automation into its business operations.

Transportation and Travel

From recommending the best route for drivers to arranging travel reservations remotely, AI has now become a gigantic trend in this business. End consumers are finding it easier to navigate thanks to artificial intelligence. Furthermore, travel businesses that integrate AI into their systems profit from Smartphone usage.

E-Commerce

Have you ever come upon a picture of clothing that you were hunting for on one website but couldn't find on another? Well, that is done by AI. It's due to the machine learning techniques that businesses employ to develop strong client connections. These technologies not only personalize customers' experiences but also assist businesses in increasing sales.

Conclusion

In the early twenty-first century, no place has had a larger influence on AI than the workplace. Machine-learning techniques are resulting in productivity gains that have never been observed before. AI is transforming the way we do business, from workflow
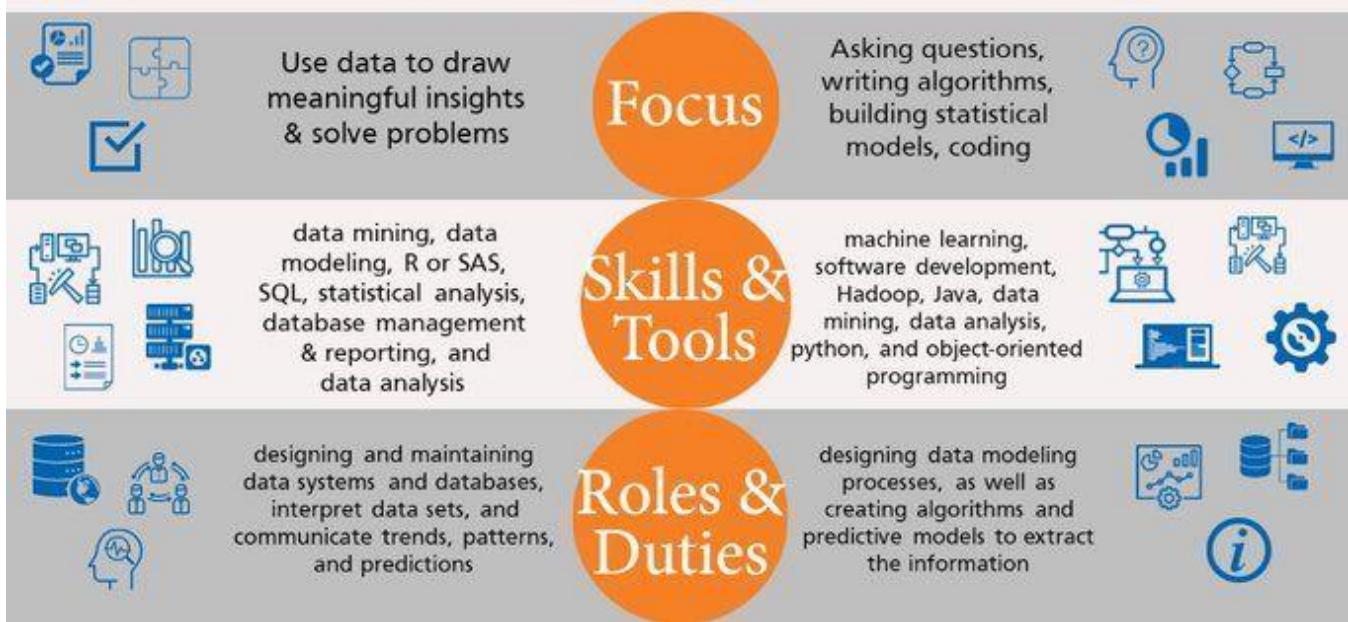
management solutions to trend forecasts and even the way businesses buy advertising. AI research has so much promise that it's becoming difficult to envisage a world without it. Be its self-driving vehicles, more precise weather predictions, or space travel, AI will be prevalent in everyday life by 2030.

❖ `Statistics vs. Data Mining

Data Mining	Statistics
Data mining is a process of extracting useful information, pattern, and trends from huge data sets and utilizes them to make a data-driven decision.	Statistics refers to the analysis and presentation of numeric data, and it is the major part of all data mining algorithm.
The data used in data mining is numeric or non-numeric.	The data used in the statistic is numeric only.
In data mining, data collection is not more important.	In statistics, data collection is more important.
The types of data mining are clustering, classification, association, neural network, sequence-based analysis, visualization, etc.	The types of statistics are descriptive statistical and Inferential statistical.
It is suitable for huge data sets.	It is suitable for smaller data set.
Data mining is an inductive process. It means the generation of new theory from data.	Statistics is the deductive process. It does not indulge in making any predictions.
Data cleaning is a part of data mining.	In statistics, clean data is used to implement the statistical method.
It requires less user interaction to validate the model, so it is easy to automate.	It requires user interaction to validate the model, so it is complex automate.
Data mining applications include financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Certain Scientific Applications, etc.	The application of statistics includes biostatistics, quality control, demography, operational research, etc.

❖ Data Science vs. Data Analytics

Data Analytics *VS.* Data Science

Data Analytics	Focus	Data Science
Use data to draw meaningful insights & solve problems	**Focus**	Asking questions, writing algorithms, building statistical models, coding
data mining, data modeling, R or SAS, SQL, statistical analysis, database management & reporting, and data analysis	**Skills & Tools**	machine learning, software development, Hadoop, Java, data mining, data analysis, python, and object-oriented programming
designing and maintaining data systems and databases, interpret data sets, and communicate trends, patterns, and predictions	**Roles & Duties**	designing data modeling processes, as well as creating algorithms and predictive models to extract the information

Overview of Business Analytics Application : Financial Analytics- Marketing Analytics- HR Analytics – Supply Chain Analytics- Retail Industry- Sales Analytics- Web & Social Media Analytics- Healthcare Analytics- Energy Analytics- Transportation Analytics- Lending Analytics- Sports Analytics- Future of Business Analytics.

Financial Analytics

Financial analytics is a concept that provides different views on the business' financial data. It helps give in-depth knowledge and take strategic actions against them to improve your business' overall performance. Financial analytics is a subset of <u>BI & EPM</u> and has an impact on every aspect of your business. It plays a crucial role in calculating your business' profit. It helps you answer every business question related to your business while letting your forecast the future of your business.

So why is financial analytics important?

- Today's businesses require timely information for decision-making purposes
- Every company needs prudent financial planning and forecasting
- The diverse needs of the traditional financial department, and advancements in technology, all point to the need for financial analytics.
- Financial analytics can help shape up the business' future goals. It can help you improve the decision-making strategies for your business.
- Financial analytics can help you focus on measuring and managing your business' tangible assets such as cash and equipment.
- It provides an in-depth insight into the organization's financial status and improves the cash flow, profitability, and business value.

Important financial analytics you need to know

In today's data-driven world, analytics is critical for any business that wants to remain competitive. Financial analytics can help you understand your business' past and present performance and make strategic decisions. Here are some of the critical financial analytics that any company, size notwithstanding, should be implementing.

1. Predictive sales analytics

Sales revenue is critical for every business. As such, accurate sales projection has essential strategic and technical implications for the organization. A predictive sales analytics involves coming up with an informed sales forecast. There are many approaches to predicting sales, such as the use of correlation analysis or use of past trends to forecast your sales. Predictive sales analytics can help you plan and manage your business' peaks and troughs.

2. Client profitability analytics

Every business needs to differentiate between clients that make them money and clients that lose them money. Customer profitability typically falls within the 80/20 rule, where 20 percent of the clients account for 80 percent of the profits, and 20 percent of the clients account for 80 percent of customer-related expenses. Understanding of which is vital.

By understanding your customers' profitability, you will be able to analyze every client group and gains useful insight. However, the greatest challenge to customer profitability analytics comes in when you fail to analyze the client's contribution to the organization.

3. Product profitability analytics

For organizations to remain competitive within an industry, organizations need to know where they are making, and losing money. Product profitability analytics can help you establish the profitability of every product rather than analyzing the business as a whole. To do this, you need to assess each product individually. Product profitability analytics can also help you establish profitability insights across the product range so you can make better decisions and protect your profit and growth over time.

4. Cash flow analytics

You need a certain amount of cash to run the organization on a day-to-day basis. Cash flow is the lifeblood of your business. Understanding cash flow is crucial for gauging the health of the business. Cash flow analytics involves the use of real-time indicators like the Working Capital Ratio and Cash Conversion Cycle. You can also predict cash flow using tools like regression analysis. Besides helping with cash flow management and ensuring that you have enough money for day-to-day operations, cash flow analytics can also help you support a range of business functions.

5. Value-driven analytics

Most organizations have a sense of where they are going to and what they are hoping to achieve. These goals can be formal and listed on a strategy map that pinpoints the business' value drivers. These value drivers are the vital drivers that the organization needs to pull to realize its strategic goals. Value driver analytics assesses these levers to ensure that they can deliver the expected outcome.

6. Shareholder value analytics

The profits and losses, and their interpretation by analysts, investors, and the media can influence your business' performance on the stock market. Shareholder value analytics

calculates the value of the company by looking at the returns it is providing to shareholders. In other words, it measures the financial repercussions of a strategy and reports how much value the strategy in question is delivering to the shareholders. Shareholder value analytics is used concurrently with profit and revenue analytics. You can use tools like Economic Value Added (EVA) to measure the shareholder value analytics.

Conclusion

Financial analytics is a valuable tool that every organization, small and large, should use to manage and measure its progress. Done right, it can help the organization adapt to the trends that affect its operations.

❖ **Marketing Analytics**

Modern marketing is a data-driven process fueled by analytics. Without analyzing relevant key performance indicators (KPI), businesses can't tell whether their marketing efforts are providing the expected return on investment (ROI). Marketing analytics is the key to evaluating past performance and determining how to improve it going forward.

Marketing analytics is a set of technologies and methods used to transform raw data into marketing insights. The goal of marketing analytics is to maximize ROI from an enterprise's marketing initiatives. Marketing analytics encompasses tools for planning, managing, and evaluating these efforts across every channel.

Marketers use established business metrics, and sometimes create new KPIs, to measure the success of their organizations' marketing initiatives. These metrics include:

- Profitability segmented by demographic
- Churn rate
- Customer lifetime value
- Customer satisfaction
  - Public perception

Businesses can analyze performance indicators alongside other data, such as customer profiles or demographic trends, to reveal the causal links between marketing decisions and actual sales.

❖ **Importance of marketing analytics**

Marketing analytics makes advertising more effective and automates many rote tasks:

- Marketing analytics helps stakeholders achieve a comprehensive view across all marketing channels, such as pay-per-click (PPC) advertising, email marketing, and social media. Analytics can clarify the big picture, as well as dig down into more granular marketing trends.

- Marketing analytics tools improve lead generation by providing the insights needed to optimize advertising efforts and target the most profitable consumers. Better leads generate more sales and improved ROI.

- Marketing analytics provides insights into customer behavior and preferences. Businesses can then tailor their marketing initiatives to meet the needs of individual consumers.

- Marketing analytics enables real-time decision support as well as proactive management. Modern analytics tools make it easy for stakeholders to analyze data as it comes in, so marketing can be adjusted as required by changing trends, and they also allow businesses to use predictive analytics to anticipate those trends rather than react to them.

❖ Benefits of Marketing Analytics

Marketing analytics can benefit organizations' marketing initiatives across all channels. Enterprises should consider the many applications of marketing analytics and determine which may be valuable to them.

Understand search marketing

Many organizations access huge markets through search engines like Google, where consumers often begin their purchasing journeys. Search engine marketing (SEM) promotes businesses and raises online visibility through advertising on search engine results pages (SERP). Revenue from digital advertising in the U.S. breaks new records every year, and search advertising accounts for **almost half of this revenue**. Businesses can use services like Google Ads and Bing Ads to expand their reach.

Organizations need marketing analytics to track and optimize the performance of their SEM efforts. One application of SEM analytics might involve serving different versions of the same ad to a randomized set of browsers and then comparing the performance of these ads in real time.

Search engine optimization (SEO) involves adjusting web content and structure to improve organic search engine rankings. An enterprise can use SEO to reach more consumers and enhance its brand. Tools like **Google Analytics** allow businesses to track relevant KPIs and analyze how their SEO initiatives are progressing and how to improve them.

Analyze social media engagement

More than **a third of the world's population** — including **98% of digital consumers** — spend time on social media, averaging almost two and a half hours per day on these platforms. While SEM drives sales from customers who are searching for specific products, social media marketing can generate interest and demand from new groups of consumers. Social media is now the primary or sole marketing channel for many businesses and organizations, such as crafts sellers on Pinterest, fashion brands on Instagram, and nonprofits on Facebook. Many social media platforms offer their own analytics tools, such as Facebook Insights or Twitter Analytics, and third-party options are available as well.

Analyzing data obtained through social media platforms can provide valuable insights for building business or customer relationships. For example, marketers can set up an account to automatically post information about new products or features as they come out, use an analytics tool to evaluate consumer sentiment through comments or reactions without manually sifting through the data, and then rework their social media marketing as necessary.

Optimize email marketing

Though businesses can use email promotions to reach new audiences, email marketing is more often concerned with existing customers who have opted in to mailing lists or have already purchased products and services. Email provides a more direct gauge of consumer sentiment than other channels, because existing customers are more likely to respond to

surveys or interact with advertised material. Popular email marketing tools include Salesforce Marketing Cloud, Mailjet, and Autopilot.

Enterprises can use analytics to optimize and personalize email marketing efforts. Analyzing how customers interact with different email promotions can help businesses target their email marketing and tailor their messages to meet customer expectations and needs. Enterprises can use marketing analytics to determine, for instance, whether customers respond well to certain keywords, emails sent at particular times of day, or links to content on specific topics.

Take advantage of predictive scoring
Predictive lead scoring models leverage marketing data from all channels, as well as internal data, to create a full picture of customer behavior, advertising potential, and marketing opportunities. These models use machine learning to build consumer profiles, which organizations can use to predict how consumers may react to different types of advertising and outreach. Campaigns can then target individual customers to maximize efficiency. For example, a predictive scoring system could rank individuals by likelihood of retention and risk of churning, which could help prioritize outreach to an organization's customer base.

❖   **HR Analytics**

HR Analytics brings "analysis" and "statistics" together to find the application of the data pool created by HR. Plainly enough, it is a data-driven approach to manage the employees.

The purpose of HR analytics examples can be divided into two subgroups:
(i) Reaching Business Goals

This dimension of HR analytics aims to provide the organization insight into the current state of operations. Such insights expedite the fulfilment of business goals according to the set timeframes.
(ii) Data-Driven Strategy Building

It helps build prediction models to identify the strategy that could lead to the optimum return on investment (ROI) for its human resources.

HR analytics makes it easy for the HR professional to create job offers that can procure them the best talent in the market, manage and retain them to give a boost to ROI.
Data Points for HR Analytics?

The data is collected across multiple points in the organization to create a pool of employee data. HR analytics tools collect data like:

(i) Employee performance

(ii) Employee attendance

(iii) Multi-rater or 360-degree reviews

(iv) Salary data

(v) Promotion data

(vi) Employee work history

(vii) Demographic details

(viii) Employee temperament data

HR analytics tools help in a close alignment of employee data and HR initiatives to direct them towards the achievement of the organization's goals.

Once the employee data is gathered, the analysts feed the same into sophisticated data modelling programs, run them through the algorithms, and predictive tools to gain insights that can be acted upon.

The insights could be represented in the reports, dashboards or visualizations. The HR analytics examples involve the below steps:

(i) Benchmark analysis

(ii) Relevant Data gathering

(iii) Data Cleansing

(iv) Data Analysis

(v) Evaluation of goals

(vi) Strategy building based on analysis

(vii) Plan execution

Practical Application of HR Analytics

The HR analytics examples are slowly edging towards practicality and it finds its use in below-mentioned areas:

1. Employee Retention

According to the stats from the US employers, the average employee replacement cost is 200% of the annual salary they draw.

When an employee leaves the organization, the cost of on boarding cost of the recruitment process and lost productivity adds to the total loss to the organization.

It becomes critical for the HR department to contain the attrition rate and it is achievable only when the HR department adapts to a data-driven approach.

Some of the aspects that drive impact iteration analysis are:

(i) Churn Rates

(ii) Per Department Attrition Rates

(iii) On boarding Experience

(iv) Employee Interview Data

(v) Employee Performance Data

HR analytics helps identify the reasons behind attrition, and develop policies and training programs to dampen the impact of attrition.

2. Employee Performance

Around 45% of HR professionals believe that performance reviews are not accurate which makes it difficult to retain the talented employees.

Thus accurate performance evaluation is critical for retention. HR analytics tools are intelligent and leverage the employee data to identify the key players based on multiple performance parameters.

As the performance measurement and career progression are interdependent, an accurate <u>data mining</u> of the two could help HR professionals identify the employee expectations.

For larger organizations, HR leaders can analyze the promotion expectation and average promotion wait time to determine how the employees can be motivated to perform better and stay.

3. Employee recruitment

The talent shortage is the biggest nightmare for enterprises. Almost 42% of the employers recruiting today are worried about the selection of not so appropriate candidate.

The HR recruitment team is primarily responsible for finding out the right CV's from the pool but before that, they need to develop a powerful ideal candidate portfolio.

This is where what is HR analytics is answered and its role is identified. The data collected from hiring managers and the performance data of previous hires for the same role are fed to the HR analytics tools to create an optimized skill set which is desired.

The data which is considered is:

(i) Identify the average number of the applicant after analyzing the applicant pool

(ii) Number of interview rounds

(iii) Offer acceptance statistics

4. Employee Development

HR plays a critical in employee development as a skill gap always exists with new recruitment.

According to American Employers, 40% of the recruited resource is not ideal for the job but with HR's employee development programs, the skill gap can be covered substantially.

The HR Analytics tools help human resource management assess the skill needs, train the employees accordingly and allocate the right resources to the teams. This increases the agility of the organization as well as enhances employee satisfaction.

Employee development programs are running currently also but digging out the right requirement is still a challenge. In light of HR analytics, more refined employee development programs can be led.

5. Employee Engagement

Attracting the best talent to the organization is an art and HR strive hard to achieve it. Having appropriate employee engagement is critical for an organization to attract and retain the employees.

As critical as it is to identify the factors that drive employee engagement, it is equally difficult to find the right metrics.

The HR needs to do Statistical analysis of employee engagement data surveys to identify the data which leads to better employee engagement.

6. Developing Compensation Programs

Employee compensation is known to contribute 33% to employee retention and performance.

As it still remains the biggest investment by the business expense, its appropriation and accuracy needs go without saying.

Both internal and external factors impact the compensation plans that is why this area requires more precise automation.

The HR needs to analyze what the competitors are offering their resources and what kind of compensation is inducing higher retention.

Overall Benefits of HR Analytics

After thoroughly analyzing the role of HR analytics it can be deduced that it can directly lead to the operational enhancement and more strategic hiring.

As the organizations are adapting HR analytics tools, the ROI is increasing. Let us take a look at what are the changes when HR analytics examples are adopted.

(i) Decreased attrition

(ii) Manual task automation

(iii) HR Process improvement

(iv) Refined hiring practices

(v) Enhancement in employee productivity

❖ **HR analytics challenges**

HR analytics requires a lot of data handling and this could prove to potential challenge enterprises need to overcome.

The practical implementation of HR analytics is not easily achievable and some of the challenges that retard the adoption rates are:

(i) Access to the right skill set who can work with the HR analytics tools

(ii) Aggressive Data Cleansing

(iii) Maintenance of high-quality Data quality

(iv) Data privacy

(v) Data compliance

(vi) Proving the value of reports to the leadership for implementation

(vii) Identifying the impact of strategies on ROI

❖ **Supply Chain Management**

- Supply chain analytics refers to the processes organizations use to gain insight and extract value from the large amounts of data associated with the procurement, processing and distribution of goods. Supply chain analytics is an essential element of supply chain management (SCM).
- Supply chains typically generate massive amounts of data. Supply chain analytics helps to make sense of all this data — uncovering patterns and generating insights.

Different types of supply chain analytics include:

1. Descriptive analytics
Provides visibility and a single source of truth across the supply chain, for both internal and external systems and data.

2. Predictive analytics
Helps an organization understand the most likely outcome or future scenario and its business implications. For example, by using predictive analytics, you can project and mitigate disruptions and risks.

3. Prescriptive analytics
Helps organizations solve problems and collaborate for maximum business value. Helps businesses collaborate with logistic partners to reduce time and effort in mitigating disruptions.

4. Cognitive analytics
Helps an organization answer complex questions in natural language — in the way a person or team of people might respond to a question. It assists companies to think through a complex problem or issue, such as "How might we improve or optimize X?"

❖ **Applying cognitive technologies**

Supply chain analytics is also the foundation for applying cognitive technologies, such as artificial intelligence (AI), to the supply chain process. Cognitive technologies understand, reason, learn and interact like a human, but at enormous capacity and speed.

This advanced form of supply chain analytics is ushering in a new era of supply chain optimization. It can automatically sift through large amounts of data to help an organization improve forecasting, identify inefficiencies, respond better to customer needs, drive innovation and pursue breakthrough ideas.

Important supply chain analytics: Supply chain analytics can help an organization make smarter, quicker and more efficient decisions. Benefits include the ability to:

Reduce costs and improve margins: Access comprehensive data to gain a continuous integrated planning approach and real-time visibility into the disparate data that drives operational efficiency and actionable insights.

Better understand risks: Supply chain analytics can identify known risks and help to predict future risks by spotting patterns and trends throughout the supply chain.

Increase accuracy in planning: By analyzing customer data, supply chain analytics can help a business better predict future demand. It helps an organization decide what products can be minimized when they become less profitable or understand what customer needs will be after the initial order.

Achieve the lean supply chain: Companies can use supply chain analytics to monitor warehouse, partner responses and customer needs for better-informed decisions.

Prepare for the future: Companies are now offering advanced analytics for supply chain management. Advanced analytics can process both structured and unstructured data, to give organizations an edge by making sure alerts arrive on time, so they can make optimal decisions. Advanced analytics can also build correlation and patterns among different sources to provide alerts that minimize risks at little costs and less sustainability impact.

As technologies such as AI become more commonplace in supply chain analytics, companies may see an explosion of further benefits. Information not previously processed because of the limitations of analyzing natural language data can now be analyzed in real time. AI can rapidly and comprehensively read, understand and correlate data from disparate sources, silos and systems.

It can then provide real-time analysis based on interpretation of the data. Companies will have far broader supply chain intelligence. They can become more efficient and avoid disruptions — while supporting new business models.

Key features of effective supply chain analytics

The supply chain is the most obvious face of the business for customers and consumers. The better a company can perform supply chain analytics, the better it protects its business reputation and long-term sustainability.

Key features of effective supply chain optimization include:

Connected: Being able to access unstructured data from social media, structured data from the Internet of Things (IoT) and more traditional data sets available through traditional ERP and B2B integration tools.

Collaborative: Improving collaboration with suppliers increasingly means the use of cloud-based commerce networks to enable multi-enterprise collaboration and engagement.

Cyber-aware: The supply chain must harden its systems from cyber-intrusions and hacks, which should be an enterprise-wide concern.

Cognitively enabled: The AI platform becomes the modern supply chain's control tower by collating, coordinating and conducting decisions and actions across the chain. Most of the supply chain is automated and self-learning.

Comprehensive: Analytics capabilities must be scaled with data in real time. Insights will be comprehensive and fast. Latency is unacceptable in the supply chain of the future.

❖ **Retail Analytics**

Retail analytics is the process of tracking business data, such as inventory levels, consumer behavior, sales numbers, and more, to make more informed, strategic decisions. This includes providing insights to understand and optimize the retail business's supply chain, consumer behavior, sales trends, operational processes, and overall performance. With today's high customer expectations for retail, companies must meet those rising needs with personalized omni channel offers, efficient processes, and quick adjustments to upcoming trends—all of which require retail analytics.

Retail analytics translate real-world business activity into quantifiable data to drive better business decision-making. This data can cover consumer behavior patterns, supply chain information, inventory updates, and more. Retailers can operationalize these insights in numerous ways, including:

- Optimizing store layout and design
- Iterating on product displays
- Improving pricing strategy
- Enhancing promotional campaigns
- Building comprehensive customer personas
- Driving personalized product recommendations

While the sources of this data are endless, not all analytics are created equal. Retail analytics derived from real-time consumer data collected via mobile app usage are particularly high-yield.

Retail analytics revolutionizes business decision-making in four main steps:

Data gathering

A team of experts collect data from all retailers, in whichever form they hold it. This is entered into an analytics platform in an intuitive dashboard, saving time and resources compared with internal sales teams doing it manually.

Data cleansing and validation

Retail analytics tools – and a team of experts – comb through the data, removing false entries and making sure all remaining data is accurate and consistent.

Data analyzing

The analytics platform puts clean data into dashboards, tables and graphs – making it effortless for businesses to analyze performance and inform campaigns. This data can also be entered into data warehouse tools to inform stock and inventory decisions.

Drive adoption

Retail analytics tools – including automated reports, intuitive dashboards and convenient data files – make it simple for everyone across the supply chain to access and learn from the business' data.

5 high-value retail analytics use

Behavioural analytics

Properly optimizing the in-store experience requires an in-depth understanding of how consumers move through the retail space, and customer flow analytics provides exactly these insights. Provided through tech platforms like Vera, retailers can discover how traffic density varies across floor space, identify points of interest, and spot shoppers' navigation patterns. In the case of Vera, this data can be collected and analyzed in real-time, ensuring retailers have the most current information possible.

Behavioural analytics like these enable retailers to make relevant, data-based decisions on in-store experience design. Instead of relying on best practices, retailers can tailor floor layouts to match the particular behavioural patterns of their consumers. Essentially, this technology brings a level of experiential optimization once only possible on the web to real-world commercial environments, significantly enhancing the efficacy of commercial spaces.

Recommendation engines

For years, online retailers have leveraged the power of analytics to offer their customers personalized product recommendations. This innovation has been integral to the success of ecommerce giants like Amazon. But, emerging technologies like augmented reality (AR) are helping level the playing field for brick and mortar businesses by providing a natural platform for deploying this analysis to drive purchasing behaviour.

By analyzing purchasing behaviour and predicting future demand, retailers can deliver personalized product recommendations through AR-enhanced in-store shopping experiences. Like online shopping, these custom notifications can be pushed to consumers according to their proximity and behaviour during the natural points in browsing. Personalized recommendations can dramatically increase customer

engagement, as research by Accenture has revealed that 65% of consumers prefer retailers that know their shopping history. Global retailer Auchan deploys a simplified version of this approach to boost foot traffic at their locations.

Inventory optimization

The goal of inventory management is to optimize the relationship between supply and demand. While this process was once the domain of educated guesses, managers can now leverage a deep set of data and analytical tools to make stocking decisions. For example, America's largest grocer — Kroger — has been using an in-house analytics team to analyze everything from economic trends to shopper behavior to accurately forecast demand for years.

One of the most exciting innovations in this area is the use of real-time customer flow statistics to anticipate inventory needs. Instead of relying on relevant historical data to project demand, retailers can also pull data from across the entire retail enterprise to make stocking decisions. While this inventory management model is preferable in almost any circumstance, it's even more essential during periods of unprecedented consumer behavior like has been seen throughout the pandemic. Having comprehensive and current data empowers inventory managers to identify trends and respond appropriately.

Predictive pricing

Price is one of the most powerful levers in commerce, and the increasing availability of data and analytical tools gives retailers an even better grasp of it. Now, retailers can use a multitude of inputs to drive their pricing strategies. These include everything from fundamentals like the cost of goods sold and competitor pricing to advanced analytics like weather forecasts and real-time customer behavioral data. With this data, retailers can use analytics to predict the ideal sale duration, identify customer price tolerance, and determine other critical elements of their pricing strategy.

Smart merchandising

From promotions to display optimization, merchandising is the art and science of selling retail goods, and few retail analytics use cases have a more direct impact on business performance. Access to store-level data and customer behavior analytics — combined with machine learning-driven analysis — are transforming how retailers are running their in-store campaigns.

The growing use of AR in shopping experiences adds another layer of actionable data to this trove, allowing retailers to get feedback on their display strategies faster than ever before. AR experiences serve as a testing ground for product placements, displays, signage, and other promotional collateral. These applications can then relay real-time data on the effectiveness of these campaigns, allowing retailers to quickly and cost-effectively iterate on campaign concepts before committing significant resources.

How to implement retail analytics

Information is power — always has been and always will be. While 21st-century innovations have largely favoured online retailers, emerging technologies like AR have changed the game for brick and mortar businesses. Now, they too can have access to real-time data about their consumers to optimize their store layouts, test campaigns, improve inventory decisions, and more.

For forward-thinking retailers, this is all excellent news. If you're interested in learning more about some of the most exciting retail analytics use cases, then reach out. Our AR platform is powering the next generation of in-store data, and we'd love to help you leverage it to grow your business.

❖ **Sales analytics**

- Sales analytics refers to the technology and processes used to gather sales data and gauge sales performance. Sales leaders use these metrics to set goals, improve internal processes, and forecast future sales and revenue more accurately.
- The goal of sales analytics is always to simplify the information available to you. It should help you clearly understand your team's performance, sales trends, and opportunities.

Generally, sales analytics is divided into four categories:

Descriptive: What happened? Descriptive analytics entails tracking historical sales data—revenue, number of users, etc.—so you can make comparisons and better understand what's currently happening.

Diagnostic: Why did it happen? Diagnostic analytics is examining and drilling down into the data to determine exactly why something occurred.

Predictive: What's going to happen? Predictive analytics is taking what you've learned about past sales and using it to gauge patterns and trends. This allows you to make educated predictions.

Prescriptive: What's the best solution or action? Prescriptive analytics involves assessing all the data and recommending the best plan of action.

Sales analysis should be a priority if you want your business to stand out in a highly competitive world, especially during decision-making scenarios.

Here are several **benefits of sales analytics for businesses:**

1. Promotes Better Decision-making

Access to data from sales analytics provides a company with the capability to make accurate decisions that can be beneficial in the long run. Companies can utilize sales analytics when they share the discussion with the workforce.

Working collaboratively allows better analysis and decisions for the benefit of all, especially when deciding on a marketing or sales strategy to implement.

2. Helps Achieve the Mission Statement

❖ With quantified values, it'll promote the growth of the business, especially the analytical process, since it defines a common objective that the workforce follows. Once these values are quantified, they'll undergo evaluation by employees to better understand their expectations of them. With well-informed employees, they're likely to be more productive.

3. Keeps Your Business Updated

Today, consumers readily change their minds as fads come and go. Most get easily swayed by seemingly good offers. With sales analytics, it can provide a company insight on the latest flow in the target market.

Always remember that the fluctuations in the industry occur rapidly. In recent years, you might see big-scale companies succumb to promising startups. Make it a priority to protect your business from the unpredictable nature of the industry with sales analytics. With the data, you can make the right move by innovating according to the current needs and preferences of the consumers.

4. Boosts Efficiency

The availability of business analytics has made it possible for businesses to improve in terms of efficiency. Since analytics can rapidly gather large amounts of data and present it appealingly, companies can decide on suitable plans to reach their objectives.

Remember that analytics can encourage a company culture that values efficiency and teamwork. It creates an environment where the workforce can readily express insights and share in the decision-making process.

Additionally, analytics allows businesses to develop better choices, such as the direction to take or figuring out the necessary steps to reach new objectives.

5. Provides Better Insights via Data Visualization

In recent years, the versions of sales analytics are easy to decipher and even presentable. Businesses can check out highly comprehensive graphs and charts to aid with the decision-making process.

With the visual representations of the data, any business will gain beneficial insights more straightforwardly. The presentation of the data visualization is organized and visually appealing.

6. Better Accessibility

With the help of sales analytics software, businesses can readily access data and produce accurate reports. Depending on the sales analytics software your company uses, it generally allows quick access on any device with Internet connectivity.

The sales team can easily access the progress at any time. The convenient access provides better efficiency and flexibility, which are both crucial to a fast-paced business environment. When the sales analytics software you're using has a quick and easy reporting interface, it can simplify the decision-making process, eventually resulting in better sales.

7. Transparency of Sales Data

The sales data can provide businesses with complete transparency and serves as a tool for mentoring the sales team. With the data, companies can utilize the necessary tactics, such as a cross-sell campaign.

The business has a clear view of the sales team's progress, including strengths, priorities, challenges, and possible weak points. In the long run, it directly affects sales since it provides you with control over variables to boost efficiency and productivity levels.

8. Helps Pinpoint Profitable and Slow-Moving Products

The data from sales analysis will allow a business to pinpoint both profitable and slow-moving products and services.

Depending on the business you're managing, you can make the appropriate modifications to the sales and marketing tactics basing on the data and distributing resources for products that show the highest chances of future growth.

Aside from identifying the profitable assets, the data can also help you identify slow-moving products or services. It'll allow you to allocate resources efficiently, cut holding expenses, and prevent over-stocking. The data from the analytics will serve as your basis for implementing changes in the prices or providing discounts.

Conclusion

Sales analytics is an indispensable tool for businesses all over the globe. Without this must-have element, your business won't last long in a highly competitive industry. Depending on the company you're managing, finding the right sales analytics software is crucial. With the benefits that sales analytics provides, making the most out of the tool will keep your business running efficiently and maintain superior productivity for years to come.

❖ Web & Social Media Analytics

Web Analytics, by definition "Web analytics is the measurement of data, the collection of information, analysis, and reporting of Internet data for the purposes of optimizing and understanding Web usage.

Web analytics uses the data collected directly from a particular business website and Social media analytics uses the data collected from social media networks.

Web analytics Gathers information from the Business website only. In general, Web Analytics tells you about your traffic levels, referral sources, bounce rate, and user behaviour on your website.

Web Analytics mainly used to improve the user experience and conversion rate. Below are the few important analysis can be done through web analytics,

- Where are your website visitors coming from?
- On which page of your website they are spending more time?
- How are they connecting to your website?
- Which part of the day is your website getting more traffic?
- How well do you retain users?

In other words, the four important key metrics can be analyzed from web analytics,

- 1. Total Traffic
- 2. Traffic Source
- 3. Bounce Rate
- 4. Conversion rate

Total Traffic

Total Traffic to your website gives insights about where are you getting more traffic, which helps you to understand your target market. In addition, you can analyze which hours of the day and days of the week, you are getting more visits to your websites. Based on this information you can run a campaign to optimize more conversions.

Traffic Source

Traffic source is about how you are getting most visitors to your website. Is it through social media, search engines or Referral Sites? Based on that information you can strategize your marketing campaign or write a blog or focusing on a particular social media network. For example, if most of your visitors are coming from social media networks, use that information to brand your business more on Facebook, Twitter or any other social media platforms to boost your website traffic.

Bounce Rate

Bounce rate is the percentage of visitors to a particular website who are leaving the site after viewing only one page without navigating other pages on the website.

This could be higher for any number of reasons may be,

- Irrelevant content

- Inappropriate designs
- Confusing navigation
- Frequent pop-ups
- Unnecessary ads
- Or, Annoying sounds

However, this metric helps you to improve your web design overall.

Conversion Rate

A conversion rate is the percentage of visitors who have taken some action on your website or complete the desired goal; it could be purchasing a product, Sign up for newsletters, etc.

❖ **Social Media Analytics**

Social media analytics is the practice of gathering data from social media websites or networks such as Facebook, Twitter, Google plus, etc., and analyzing those metrics to understand insights to make business decisions.

Social media analytics gathers information from social media networking sites and helps businesses better understand customer sentiment, users' attitudes, build rich consumer profiles, and, most importantly, build effective business strategies.

There are many tools available in the market to track and analyze social media data.

The most common use of social media analytics is to discover

- Sentiments
- Opinions
- Emotions
- Topics

Put together known as Sentiment Analysis.

❖ **Sentiment Analysis**

- It is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Opinion mining or sentiment analysis refers to the use of natural language processing.
- Importantly, the first step is to define which business goal we are trying to address through social media analytics.
- In general, the business objectives include increasing revenues, reducing customer service costs, getting feedback on products and services, and improving public opinion of a particular product or business division.
- A quick recap of information gathering

❖ **Importance of social media analytics**

There is a tremendous amount of information in social media data. In decades past, enterprises paid market research companies to poll consumers and conduct focus groups to get the kind of information that consumers now willingly post to public social media platforms.

In the past few years, businesses have rushed to use Web and Social Media widely. Almost 94% of all businesses now use some form of social media to promote their brand and engage with customers. While companies sprint to master web and social media marketing, the analysis of data available in web and social media more of a struggle.

The problem is this information is in the form of free text and natural language, the kind of unstructured data that analytics algorithms have traditionally. However, as machine learning and artificial intelligence have advanced, it has become easier for businesses to quantify in a scalable way the information in social media posts.

This allows enterprises to extract information about how the public perceives their brand, what kind of products consumers like and dislike and generally, where markets are going. Social media analytics makes it possible for businesses to quantify all this without using less reliable polling and focus groups.

What Information can be gathered from?

Web Analytics
Four important key metrics can be analysed

1 Total Traffic
Total Traffic to your website gives insights about from where are you getting more traffic, that helps you to understand your target market.

2 Traffic Source
Traffic source is about how you are getting most visitors to your website.
@gectouch

3 Bounce Rate
Bounce rate is the percentage of visitors to a particular website who are leaving the site after viewing only one page without navigating other pages on the website.

4 Conversion Rate
A conversion rate is the percentage of visitors who have taken some action in your website or complete the desired goal.

Social Media Analytics
Social media analytics is to discovery

1 Sentiments
It is the process of computationally identifying and categorizing opinions expressed in a piece of text.

2 Opinions
Opinion about the product or services and review will be shared by the consumers in social media networks.

3 Emotions
Mostly consumer will share their emotion about the product only if they are extremely satisfied or dissatisfied.

4 Topics
Analysing the topic in the social media will give you good insight about market trends and expectation.

© gecdesigns.com

❖ **Health Care Analytics**

Healthcare analytics refers to the use of data that offers comprehensive insight into patients and the conditions that affect or have the potential to affect them. Analytics in healthcare can also provide insight into the effectiveness of healthcare providers themselves in terms of productivity, cost and general performance. Essentially, it is all about gathering and leveraging information to improve quality and efficiency of healthcare services.

The benefits of data analytics in healthcare

The advantages that advanced analytics in healthcare can bring are vast and well-recognised by those working within the industry. Let's dive into some of the specific ways that data and analytics can support healthcare.

- **Predict risks** - As the saying goes, prevention is better than cure. By collecting mass data, hospitals can identify common symptoms and causes of conditions and diseases. This helps doctors spot when a patient may be at risk of developing a certain health problem and treat them as early as possible.

- **Make data-driven decisions** – With more data at their disposal in regard to patient medical history and the health of the wider population, healthcare professionals can make informed decisions about individual treatment and how likely it is to be successful.

- **Increase patient satisfaction** – Insights from data can help doctors to personalise treatment and improve how they care for patients. Software can even assess the performance of healthcare professionals and provide feedback.

- **Improve service delivery** – Hospitals can use software to predict busier times and appropriately plan to meet demand, for example by having more staff on rota. This can help reduce long waiting times and shortages of beds.

- **Electronic record-keeping** - Storing medical records electronically, as opposed to on paper, also improves productivity by mitigating the problem of having multiple records. It also enables different healthcare professionals to access the same records without transporting paperwork between facilities.
- **Reduce costs** – With data improving patient care and allowing healthcare facilities to run more productively, treatment costs and other hospital expenses can be minimised.

Healthcare analytics importance

- Despite the pandemic advancing how technology is used in healthcare, the industry is slow to adapt overall. For example, it has been reported that just over half of hospitals

have no strategies for data governance or analytics in their day to day practices, and 97% of data produced by hospitals is wasted.

- The Health Foundation similarly discovered that, although the NHS generates masses of data, they lack staff with the right analytical experience to interpret this data. Consequently, opportunities to improve services – such as by improving diagnoses and day-to-day care – are being missed. Similarly, leaders in the industry noted that the pandemic exposed many flaws within many health care systems with poor quality of data, time-consuming analytic processes, and staff lacking the training to use data properly being a common occurrence.

- With much of the industry failing to innovate when it comes to technology, it is important to both understand the benefits that analytics can bring and consider how it can be incorporated into your organisation. After all, technology is only going to develop further, and the industry will need to innovate in order to be ready for future challenges. As industry professionals state, "digital health solutions and technology will play a crucial role in the difficult work of optimizing processes and systems for greater efficiency, financial viability, and enhanced outcomes."

❖ **Energy Analytics**

Energy analytics generally describes the process of collecting electrical data and applying sophisticated analytical software and algorithms to deliver insights around consumption and time of use reductions.

Business benefit from energy analytics

1. Valuable insights into your energy data

Energy analytics can provide you with unique insights into your business energy data that would've been impossible to find manually or with Excel.

For example, energy analytics software can show you the periods where you spend the most on energy. It can also help you understand which areas of your organisation are inefficient and how energy consumption is affected by external factors such as the weather.

2. Improved energy efficiency and reduced energy costs

If you pay attention to the insights delivered by energy analytics and take appropriate actions, you can quickly reduce your energy costs. Before you know it, the software you're using will be paying for itself.

As an example, energy data analytics could highlight that one of your buildings uses a surprising amount of energy in non-operational hours. Further investigation could uncover that the heating controls for this building are faulty. Taking action and fixing this problem will likely save thousands over the course of a year.

Research suggest that most companies can save at least 10% on their energy bills with energy data analytics.

3. It can streamline your job as an energy manager

Energy managers are usually responsible for more than just electricity use. Chances are you're also held accountable for gas and water use, maybe even generation if you've installed renewable energy solutions onsite.

Energy analytics software can connect with all of this data and import it automatically. This means you no longer have to log into four or five different systems to build a complete picture of your organisation's utility use.

4. Automation of time-consuming activities

- Energy management is a role that sometimes involves doing the same things over and over again. This can include creating and sending out reports, conducting degree day analysis or analysing the success of a project.
- Energy analytics software can help with this by automating these tasks, allowing you to spend more time planning and running energy-saving projects and improving the energy efficiency of your business.

5. It simplifies data sharing and collaboration

- As businesses have begun to realise the financial and environmental benefits associated with energy efficiency, there's been an increase in the number of energy managers employed by companies.
- Some organisations are even hiring teams of energy managers to improve their energy efficiency.
- When working in a team, it's vitally important that data can be shared and communicated quickly and clearly.

- In other words, energy analytics software helps ensure that essential energy conversations don't get lost in a sea of emails. It also means that there's one dedicated place for people in your organisation to see, analyse and discuss energy data.

❖ **Transportation Analytics**
Transportation data analytics increasingly power mobility information and insights – transforming transportation planning by making it easier, faster, cheaper, and safer to collect and understand critical information.
While the transportation industry may not be in crisis, it is certainly being heavily disrupted by multiple forces, including the COVID-19 pandemic. As these changes unfold, transportation experts must:

- Prioritize projects accurately to guide effective resource investment and make the biggest impact.
- Make informed decisions based on recent, accurate data, not on guesses or input from a few vocal stakeholders.
- Maintain social equity and environmental justice, providing access and support for outlying areas and the underserved.
- Foster public engagement, so that residents, constituents, and public officials understand, can respond to questions about, and support planned mobility efforts.
- Accurately and quickly measure results of transportation initiatives, enabling adjustment and optimization in real time.

Increasing numbers of cities, transit organizations, departments of transportation, and other localities are using transportation data analytics to solve problems, prioritize investments, and win stakeholder support.

❖ **Implementation of Multimodal Transport Segment-wise Analysis**

With the growth of multi-modal transport, the need for segment-wise analysis is essential. The primary modes of transportation include roadways, railways, waterways, and airways.

Let us explore the implementation of analysis in each segment:

Roadways

Using analytics for one of the most used modes of transportation, roadways, has several benefits:

Road Safety Management

Advanced data can be used to analyze where, why, and when accidents happen. With this data, they can create **Prognostication Crash Maps** (shown in the image) that analyze data to shortlist high-risk areas. These maps can help issue warnings to be extra careful at these locations and help authorities take precautionary measurements.

Road Traffic Management

Keeping a record of automobiles' moving patterns, velocity, and lane changing behaviour can help us understand how different road designs can influence driving. The insights are useful for smarter traffic control and identifying congestion in the road layout when planning future infrastructural developments. The graph depicts the same.

Rail Traffic Management

There is a whole range of possibilities that railways can explore in big data analytics. Applications in the railway industry include booking, improving security, automatic scheduling and planning, network enhancement and ticket management. The existing data from the passenger operating control, reservation system, CCTV, and maintenance depots can be used to our advantage to yield business benefits in the above areas. Real-time train information system (RTIS), the Nation Train Enquiry System (NTES), and the control office application (COA) are some examples where data analytics is used.

Air Traffic Management

Long queues are a top annoyance of air travellers. However, by accessing data of those travellers coming through the facility, advanced analytics can help airport workers easily visualize the busiest periods for their security checkpoints. Over time, machine learning-

powered by AI can generate predictive models that can allow the airport to strategize better and allocate resources.

Waterways

Ship Monitoring and Route Optimization
Ship monitoring is one of the most critical factors for seamless planning and execution. Various tools such as vessel's sensors, weather station reports, and satellite reports will increase ships' efficiency. The entire data array can be processed through machine learning, and the following questions can be answered using the same:
• When does the hull need cleaning to save fuel?
• When should the ship equipment be changed?
• Which is the best route in terms of weather, safety, and which route is fuel sustainable?
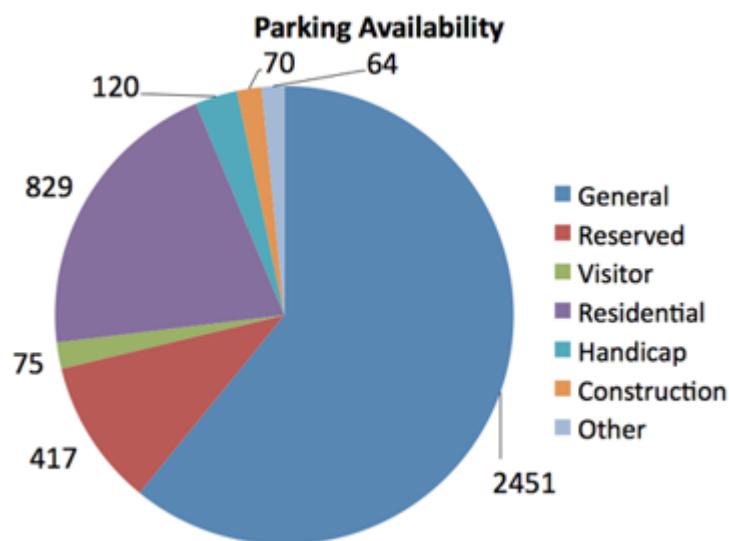
❖ **Analytics in Regular Day-to-day Transportation**

Ease Traffic Congestion:
Agencies can help ease traffic congestion by using a high occupancy toll (HOT) with real-time analytics. Based on the traffic, they can dynamically adjust prices and open HOT lanes at a much higher cost, reducing traffic.

Increased efficiency is finding Parking Slots:
In major cities, people who are trying to find parking slots cause 15%-30% of traffic. New technologies like cameras, sensors, and geo-tracking and analytics can help drivers find parking spots.

Parking Availability

	Value
General	2451
Reserved	417
Visitor	75
Residential	829
Handicap	120
Construction	70
Other	64

Analytics can help the transportation industry, especially the multi-modal transportation system, to be sustainable and efficient. However, with rapid advancements in technology and data flowing in and out, various factors such as privacy, regulations, and confidentiality must be taken care of to use data 100% effectively to provide fruitful analysis.

❖ **Lending Analytics**

Financial institutions like banks have been using predictive customer analysis for a long time. But as the complexity of loans increased, so did the need for more complex and accurate analysis.

Most of the past loan frauds are correlated to the prediction models that were being used for decades, which were not effective to detect bad loan potentials. But, as the days passed by, the financial risk factors started to plummet and it became almost impossible to approve bad loans. Today, thanks to data analytics, the default and fraud probability of loans have decreased significantly.

1. **Customer Selection**

   Customer selection is the fundamental part of any loan proceeding. While the data verification methods are still present, it's now possible to analytically predict the quality of applications with the help of data analytics. This kind of analysis is deemed better as it leaves no-to-minimal room for errors.

   Data analytics take account of the credit card purchases, subscriptions, and loyalty cards then categorize them into financial profiles. These are used to approve loans that have a better probability of being repaid.

   A customer selection model usually concludes if a customer will be inclined to pay the EMIs regularly and is safe to grant loans to. But, typically, the financial verifications are done by business experts with the help of data analytics to understand their financial behavior, spending pattern, and repayment credibility's for safer loan disbursement.

2. **Designing Custom Offers**
   Most of the loan offers are customized to the needs of individuals. If the applicant's credit score is poor, they might have to pay a higher interest rate with collateral. If they have faulty loan histories, they might not be able to secure a higher principal.

   In contrast, customers, who were efficient at paying previous loans are entitled to higher principal and lower interest rates.

   Although these elements were always present in loan processing, with data analytics and the availability of past data, the applicants can get to know the best offers within seconds. Immediate feedback helps customers make an informed decision quickly, in addition to ensuring the experts that no bad loans are being distributed.

   Data analytics are also being used by financial institutions to customize the promotional offers that a specific demographic is likely to avail. Small interest rates, longer repayment terms, and no-cost EMIs targeted to reach certain customers are such examples.

3. **Delinquency Detection**

It was almost impossible to predict the financial behavior of a lender with older prediction models. A customer, appearing to be the perfect candidate could pose as a bad loan with their erratic payments after the loan was approved.

As the problem grew to the extent of jeopardizing the business of the lenders, delinquency prediction models came into play. With extensive data of past loans, records of transactions, late payments, partial payments, and failed payments, the models are now able to predict the risky loans before they are approved.

Delinquency detection not only helps the banks but the borrowers too. It's possible that an individual somehow missed payments and has shown unusual financial behavior in the past, but is now trying to rework the errors. With ample data available on them, they can check their financial scores – which are usually reversible – to understand what went wrong and decide how it can be improved.

4. Strategizing Collection

Even after probability models are employed, some bad loans pass through the check. When that happens, the only way to recover the principal is by collection methods. In the past, customers were categorized by risk factors. And different contact strategies were used to extract the amount. Which resulted in failures more often than not.

But after data analytic models were introduced to the banking sector, even though bad loans still went through, the applications could be segmented into micro categories depending on demographics, financial activity, and risk ratings to employ more effective intervention techniques.

According to ScoreData, the customers are divided into three categories in collection analytics.

1. The ones who have defaulted for the first time
2. The lazy ones
3. Self-cure customers
4. Bad debt
5. Those who are beyond any redemption

The first-time defaulters are often the ones who are the safest of the customers but can be flagged as frauds if the account is new. Hence the collection efforts are marked as low.

The lazy customers forget to pay the bills but are the safest bet among the lenders. They usually pay after the period is over with late fines and interest. The collection efforts are marked as low.

Self-cure customers are occasional defaulters and usually are safe bets. Collection efforts for these customers are medium

Bad debt and point of no return customers are the ones who have a comparatively higher spending rate than their earnings but usually pay the minimum fees. Collection Efforts are considered High or non-viable for the borrowers of this category.

Fraud Detection

Credit card frauds, loan frauds, and deliberate payment delays have always been an issue for lenders. With data analytics and better customer selection, the problem had been handled quite effectively. Let's look at some of the ways data analytics help to detect fraudulent activities of a profile.

Data Analysis

Data analysis models provide an all-around view of alarming actions and suspicious transactions of the customers. It isolates attributes and identifies hidden threats to notify the banks about them.

Analytical Frameworks

By calculating statistical parameters like standard deviation and moving averages, analytical frameworks identify fraud patterns in customers. Excessive high or low numbers are also taken into consideration while screening fraudulent activities.

Cross Channel Monitoring

As the financial sectors became more centralized, all the transaction data of customers are available to the models. Which data analytics programs can monitor and analyze to prevent loan fraud.

Easy Loan Processing

The data analytical models were not designed to keep good loans at bay. Moreover, the data analytics capabilities now make the loan proceedings easier than ever if you have a good credit score.

As the financial institutions run their business on loans and interests, the faster and safer proceeding time helps them get the most benefit out of the customers as well.

The Bottom Line

Bad loans will always be there. What would the lenders do if a business goes bankrupt even after showing promise? But, it's now becoming possible to detect patterns of customers who are likely to fail loans and categorize them with the help of data analytics.

Delinquency detection, offer creation, fraud detection, and loan collection are also the benefits of data analytics in loan processing.

❖ **Sports Analytics**

Sports analytics is the process of plugging statistics into a mathematical model to predict the outcome of a given play or game. Coaches rely on analytics to scout opponents and optimize play calls in game, while front offices use it to prioritize player development. Analytics also play a major role off the field, providing fans with both sports betting and fantasy sports insights.

sports analytics is the practice of applying mathematical and statistical principles to sports and related peripheral activities. While there are many factors and priorities specific to the industry, sports analysts use the same basic methods and approach as any other kind of data analyst. Establishing parameters for measurement, like hit or fumble rate, and consistently collecting data from a broad sample is the basis of the analytics process. This data is then curated and optimized to improve the accuracy and usability of the results.

Sports analytics goes beyond traditional statistics to add accurate analysis to improve many factors in team performance.

On-Field Applications

Analytics has many on-field applications in a sports environment, including managing both individual and group performance. Coaches can use data to optimize exercise programs for their players and develop nutrition plans to maximize fitness. Analytics is also commonly used in developing tactics and team strategies. With thousands of games worth of data to study, analysts can look for patterns across a broad sample size regarding formation, counter strategies and other key variables.

Uses in Team Management

Practical data analysis has plenty of applications for the business side of sports as well. Since most professional sports teams function as businesses, they are always seeking ways to improve sales and reduce expenses across their organization. Some sports analysts specifically focus on issues regarding the marketing and sale of sports tickets and team merchandise. Modern marketing and fan outreach efforts also rely heavily on analytics to predict their consumer base and identify opportunities to increase brand engagement.

On-field Analytics
- Players performance
- Opponent's performance
- Game strategy

Off-field Analytics
- Fans engagement
- Ticket churn
- Merchandise Sales

Sport	Statistic	Definition
Baseball	Batting average	ratio of hits and number of at bats
	On-base percentage	times a player reaches base by hitting, walking, or by being hit by a pitch
	Slugging average	the ratio of number of bases earned and the number of at bats
	WHIP	Walks plus Hits allowed per Inning Pitched measure the number of baserunners the pitcher allows on hits and walks
Football	GAP ratings	Generalized Attacking Performance, introduced by Wheatcroft (2020), is a rating system for assessing the attacking and defensive capability of a team with respect to number of corners or shots in football.
	BA ratings	Bivariate attacking ratings is based on the parameter that minimises the mean absolute error (MAE) between estimated and observed match statistic.
Cricket: (Criclytics)	WASP	Winning and Scoring Prediction is a machine learning technique to predict final score of the first innings and probability of winning while chasing in second innings
	Win Probability Statistic	is a model that calculates the team's win probability in real-time using the team's historical data

Future of Business Analytics

The future of business analytics is bright. That's partly because people generate a massive amount of digital data every day. In 2020, people consumed a whopping 64.2 zettabytes of data, according to Statista. To put that number in perspective, just 1 zettabyte provides enough storage for 30 billion 4K movies, 60 billion video games, or 7.5 trillion songs in MP3 format. This number is poised for growth, too: It's projected that the data consumed globally will reach approximately 181 zettabytes in 2025.

It's therefore not surprising that students earning a degree in business analytics will enter a market with ample opportunity. According to the U.S. Bureau of Labor Statistics, the

employment of market research analysts is growing much faster than average — by a projected 22% between 2020 and 2030. The job outlook is also above average for related positions, such as management analyst (14%) and operations research analyst (25%). These numbers demonstrate the need for qualified professionals to manage all that digital information and convert it to meet business goals.

Business Analytics Trends

Technology is an ever-evolving process. Naturally, this evolution produces business analytics trends that professionals must know about to help companies leverage raw data to reach their goals. These pros must understand the following trends in data and business intelligence to optimize opportunities for their companies.

Search-Based Discovery Tools

Using raw data to answer specific questions or track trends isn't a foreign concept to many. After all, people do it routinely when they turn to tools such as Google and other search engines to find something. Search queries are second nature to many individuals, but not every company has user-friendly data discovery technology for this purpose.

These tools make it possible to sift through and find actionable data from disparate sources more efficiently. This can boost the capacity to find key insights for effective business strategies that may otherwise be lost in a sea of information. As the number of potential sources increases, the ability to cut through superfluous information and get to what matters becomes more essential.

Bringing the power of search-based discovery tools to the internal operations of enterprises isn't the only area where knowledge of business analytics trends shines. Online marketing is a key initiative for all types of businesses, and an understanding of the data mechanics of Google and other search engines is critical to its success. The future of business analytics in marketing is now. Time spent on-page, social shares, where page visits originate from, and other data can inform marketing decisions and drive relevant traffic.

AI and Machine Learning

Artificial intelligence (AI) and machine learning facilitate high efficiency at substantially reduced costs. Experts predict machine learning will complete a growing number of customer service tasks in the future, but the effectiveness of these technologies depends heavily on the people behind them. Even machines with the capacity to learn must be fed the right information, and business analysts are often the keepers of these solutions.

AI and machine learning offer widespread applications for businesses. The online publication Business 2 Community points out that the concepts provide the backbone for elements that are crucial for business effectiveness and efficiencies, such as personalized marketing, cybersecurity, talent recruitment, and customer relationship management. While these technology-driven elements are poised to impact the business environment, they still need the human touch of business analytics professionals to ensure that their functions translate to success.

Cloud Computing

Cloud computing, the process of using remote servers on the internet to store and manage data, provides many of the benefits businesses demand. These benefits include reduced reliance on physical resources that are often outdated as soon as they're installed; increased efficiencies, especially when working with teams in disparate locations; and plenty of options

for redundancy management and disaster planning. When utilized correctly, cloud computing can increase operational efficiency and reduce costs.

However, taking advantage of these benefits isn't always easy. Companies need to develop viable workflows to pool resources, share information efficiently but securely, oversee data access to protect consumers, and maintain the speed and ease with which data can be accessed. Business analysts manage data flow within cloud-based structures, help design and develop data processes, and analyze both the data and the performance of systems overall to ensure that business goals are supported.

Predictive Analytics Tools

Predictive analytics will likely play a large role in the future of business analytics. Predictive analytics in business often comes down to the need to anticipate the customer's or the client's next move. By analyzing historical data patterns in consumer behavior, market fluctuations, and even societal trends, businesses can prepare for certain outcomes and performances with increased confidence. This could not only keep them consistently relevant in their industry but also transform them into industry leaders.

Because the concept's success depends on the future, the information from the past must be properly interpreted for upcoming plans. If not, negative ramifications might follow. The work of a highly skilled business analytics professional can be critical for keeping the purpose behind predictive analytics beneficial and not something that can inadvertently cause business strategies to hit unexpected snags.

Data Automation

With the amount of data reaching the zettabytes, sorting, storing, and managing it can be an increasingly challenging and time-consuming process. This is what makes data automation such an important component for the future of business. Data automation can automatically take care of the mundane yet critical parts of data management, leaving those in business analytics roles more time to analyze and interpret gathered findings. It can also be a vital part of helping businesses overcome any scalability issues they may have.

The key to effectively using data automation is building an effective strategy that properly curates the automated system. This means the data is stored in a way that makes it easy to access and use at the appropriate time. Business analytics professionals can be instrumental in this key process, as they can use their knowledge and skills to build strategies that integrate sensibility and efficiency into the data automation process.

UNIT V
Ethical, Legal and Organizational Issues

Issues & Challenges- Business Analytics Implementation Challenges –Privacy and Anonymization- Hacking and Insider Threats- Making Customer Comfortable

---

The era of technology has given rise to a sea wave of change of how data is used. Business Analytics is an emerging field in data science that aims at using data to develop business insights that can be beneficial to the organization. You'll see implementation of BA in various areas. Let's take an example: Credit rating companies analyze the credit card transactions of its consumers and can predict the spending pattern of the consumer as well as his financial health. This information can be useful to companies to find their target audience. Another example can be a mobile company which extracts data about the customer's frequency of calls, recharge amounts, etc. This information can be classified and later used by the company as per its objective. In fact, Bharti Airtel, has successfully implemented BA as a part of their Analytics department, and it has led to positive outcomes.

Benefits of implementing BA in your organization

Apart from having applications in various arenas, following are the benefits of Business Analytics and its impact on business –

- Accurately transferring information
- Consequent improvement in efficiency
- Help portray Future Challenges
- Make Strategic decisions
- As a perfect blend of data science and analytics
- Reduction in Costs
- Improved Decisions
- Share information with a larger audience
- Ease in Sharing information with stakeholders

Moreover, any technology is subject to its own set of problems and challenges. Following are the common challenges in implementing business analytics in an organization.

- Lack of technical skills in employees
- Fuss over acceptance of BA by staff
- Data Security and Maintenance
- Integrity of Data
- Delivering relevant information in the given time
- Inability to address complex issues
- Costs involved in implementing BA
- Investment of staff time in implementation of BA
- Lack of a proper strategy to implement BA

The major challenges in Business Analytics are as follows:

Increase in number of Sources

When data sets become bigger and more complex, bringing them into an analytics framework poses a huge challenge in business analytics. If this is ignored, it creates gaps that result in incorrect communications and observations.

Shortage of Talent for Data Analytics

Data processing is essential in order to render usable this voluminous volume of data that is generated every minute. The immense demand for big data scientists and big data analysts has been generated on the market with the rapid growth in data. It is essential for business organizations, because the role of a data scientist is multidisciplinary, to employ a data scientist with qualifications that are versatile.

Data Secrecy

Gaining valuable insights from Big Data analytics is crucial for business enterprises and it is therefore critical that this intelligence is obtained only through the appropriate team. A major challenge in business analytics posed by the Big Data analytics firms is to successfully bridge this large void.

Handling Large Volumes of Data

It's not shocking that for every day that passes data is through. It clearly means that business entities need to manage a great regular amount of info. The volume and quality of data available these days will overpower every computer engineer and this is why it is deemed important for brand owners and managers to make data accessibility quick and convenient.

Changing technological Landscape

Each day new technology and businesses are being built with the growth of Big Data. Nonetheless, a major challenge in business analytics posed by big data analytics firms is to figure out which technologies can better match them without new challenges and future threats being added.

Quality of storage and retrieving data

Enterprise companies are increasing rapidly. With the exponential development of the companies and major business organizations, the volume of data generated increases. Storing a large volume of data is now a huge challenge in business analytics for us. Popular data storage solutions such as data lake/warehouses are widely used to capture and preserve vast volumes of unstructured, organized data in their format.

Considering the above challenges, there is a dearth of professionals who're well-equipped with the knowledge of Business Analytics. A BA professional can take up a corporate roles in various sectors: marketing, insurance, management, finance, health care & lifestyle, etc. In fact, there is still hesitation to use trends and statistics for making business decisions, and most of them still are comfortable trusting their gut feeling for making strategic decisions.

❖ **Data Privacy and Anonymization**

Data privacy, sometimes also referred to as information privacy, is an area of data protection that concerns the proper handling of sensitive data including, notably, *personal data* , but also other confidential data, such as certain financial data and intellectual property data, to meet regulatory requirements as well as protecting the confidentiality and immutability of the data.

Roughly speaking, data protection spans three broad categories, namely, traditional data protection (such as backup and restore copies), data security, and data privacy as shown in the Figure below. Ensuring the privacy of sensitive and personal data can be considered an outcome of best practice in data protection and security with the overall goal of achieving the continual availability and immutability of critical business data.

Please note that the term data privacy contains what the European Union (EU) refers to as "data protection."

Data Protection

Traditional Data Protection			Data Security			Data Privacy		
Backup/ Restore	RAID & Erasure Coding	Replication	Encryption	Threat Monitoring	Authentication	Legislation	Policies	Best Practice
Archiving	Physical Infrastructure	Data Retention	Access Control	Breach Access and recovery	Data Loss Prevention	3rd party contacts	Data Governance	Global variations

Timely availability of unspoiled data

Figure: The Three Categories of Data Protection

Security becomes an important element in protecting the data from external and internal threats but also when determining what digitally stored data can be shared and with whom. In a practical sense, data privacy deals with aspects of the control process around sharing data with third parties, how and where that data is stored, and the specific regulations that apply to those processes.

Almost all countries in the world have introduced some form of legislation concerning data privacy in response to the needs of a particular industry or section of the population.

Data Sovereignty

Data sovereignty refers to digital data that is subject to the laws of the country in which it is located.

The increasing adoption of cloud data services and a perceived lack of security has led many countries to introduce new legislation that requires data to be kept within the country in which the customer resides.

Current concerns surrounding data sovereignty are related to governments trying to prevent data from being stored outside the geographic boundaries of the originating country. Ensuring that data exists only in the host country can be complex and often relies on the detail provided in the Service Level Agreement with the Cloud Service Provider.

Data Privacy - Geographical variations in terms

In the European Union, privacy is recognised as an absolute fundamental right and in some parts of the world privacy has often been regarded as an element of liberty, the right to be free from intrusions by the state. In most geographies, privacy is a legal concept and not a technology, and so it is the term data protection that deals with the technical framework of keeping the data secure and available.

Why is Data Privacy important?

The answer to this question comes down to business imperatives:

1. **Business Asset Management:** Data is perhaps the most important asset a business owns. We live in a data economy where companies find enormous value in collecting, sharing and using data about customers or users, especially from social media. Transparency in how businesses request consent to keep *personal data,* abide by their privacy policies, and manage the data that they've collected, is vital to building trust with customers who naturally expect privacy as a human right.

2. **Regulatory Compliance:** Managing data to ensure regulatory compliance is arguably even more important. A business may have to meet legal responsibilities about how they collect, store, and process personal data, and non-compliance could lead to a huge fine. If the business becomes the victim to a hack or ransom ware, the consequences in terms of lost revenue and lost customer trust could be even worse.

Data Privacy is not Data Security

Businesses are sometimes confused by the terms and mistakenly believe that keeping personal and sensitive data secure from hackers means that they are automatically compliant with data privacy regulations. This is not the case. Data security protects data from compromise by external attackers and malicious insiders whereas data privacy governs how the data is collected, shared and used.

Differing legal definitions of Data Privacy

If there is agreement on the importance of data privacy to a business, then the legal definition can be extremely complex.

What is meant by data privacy - it is left to businesses to determine what they consider best practice in their own industry. The legislation often refers to what is considered 'reasonable' which may differ between laws, along with the respective fines.

In practice, this means that companies who work with sensitive and personal data should consider exceeding the legal parameters to ensure that their data practices are well above those outlined in the legislation.

❖ Data Anonymization

Data anonymization is the process of protecting private or sensitive information by erasing or encrypting identifiers that connect an individual to stored data. For example, you can run Personally Identifiable Information (PII) such as names, social security numbers, and addresses through a data anonymization process that retains the data but keeps the source anonymous.

However, even when you clear data of identifiers, attackers can use de-anonymization methods to retrace the data anonymization process. Since data usually passes through multiple sources—some available to the public—de-anonymization techniques can cross-reference the sources and reveal personal information.

The General Data Protection Regulation (GDPR) outlines a specific set of rules that protect user data and create transparency. While the GDPR is strict, it permits companies to collect

anonymized data without consent, use it for any purpose, and store it for an indefinite time—as long as companies remove all identifiers from the data.

Data Anonymization Techniques

- **Data masking**—hiding data with altered values. You can create a mirror version of a database and apply modification techniques such as character shuffling, encryption, and word or character substitution. For example, you can replace a value character with a symbol such as "*" or "x". Data masking makes reverse engineering or detection impossible.
- **Pseudonymization**—a data management and de-identification method that replaces private identifiers with fake identifiers or pseudonyms, for example replacing the identifier "John Smith" with "Mark Spencer". Pseudonymization preserves statistical accuracy and data integrity, allowing the modified data to be used for training, development, testing, and analytics while protecting data privacy.
- **Generalization**—deliberately removes some of the data to make it less identifiable. Data can be modified into a set of ranges or a broad area with appropriate boundaries. You can remove the house number in an address, but make sure you don't remove the road name. The purpose is to eliminate some of the identifiers while retaining a measure of data accuracy.
- **Data swapping**—also known as shuffling and permutation, a technique used to rearrange the dataset attribute values so they don't correspond with the original records. Swapping attributes (columns) that contain identifiers values such as date of birth, for example, may have more impact on anonymization than membership type values.
- **Data perturbation**—modifies the original dataset slightly by applying techniques that round numbers and add random noise. The range of values needs to be in proportion to the perturbation. A small base may lead to weak anonymization while a large base can reduce the utility of the dataset. For example, you can use a base of 5 for rounding values like age or house number because it's proportional to the original value. You can multiply a house number by 15 and the value may retain its credence. However, using higher bases like 15 can make the age values seem fake.
- **Synthetic data**—algorithmically manufactured information that has no connection to real events. Synthetic data is used to create artificial datasets instead of altering the original dataset or using it as is and risking privacy and security. The process involves creating statistical models based on patterns found in the original dataset. You can use standard deviations, medians, linear regression or other statistical techniques to generate the synthetic data.

❖ **Disadvantages of Data Anonymization**

The GDPR stipulates that websites must obtain consent from users to collect personal information such as IP addresses, device ID, and cookies. Collecting anonymous data and deleting identifiers from the database limit your ability to derive value and insight from your data. For example, anonymised data cannot be used for marketing efforts, or to personalize the user experience.

❖ **Hacking and Insider Threats**

A commonly used hacking definition is the act of compromising digital devices and networks through unauthorized access to an account or computer system. Hacking is not always a malicious act, but it is most commonly associated with illegal activity and data theft by cyber criminals.

❖ **Hacking**

Hacking refers to the misuse of devices like computers, smartphones, tablets, and networks to cause damage to or corrupt systems, gather information on users, steal data and documents, or disrupt data-related activity.

A traditional view of hackers is a lone rogue programmer who is highly skilled in coding and modifying computer software and hardware systems. But this narrow view does not cover the true technical nature of hacking. Hackers are increasingly growing in sophistication, using stealthy attack methods designed to go completely unnoticed by cybersecurity software and IT teams. They are also highly skilled in creating attack vectors that trick users into opening malicious attachments or links and freely giving up their sensitive personal data.

As a result, modern-day hacking involves far more than just an angry kid in their bedroom. It is a multibillion-dollar industry with extremely sophisticated and successful techniques.

History of Hacking/Hackers

Hacking first appeared as a term in the 1970s but became more popular through the next decade. An article in a 1980 edition of Psychology Today ran the headline "The Hacker Papers" in an exploration of computer usage's addictive nature. Two years later, two movies, Tron and WarGames, were released, in which the lead characters set about hacking into computer systems, which introduced the concept of hacking to a wide audience and as a potential national security risk.

Sure enough, later that year, a group of teenagers cracked the computer systems of major organizations like Los Alamos National Laboratory, Security Pacific Bank, and Sloan-Kettering Cancer Center. A Newsweek article covering the event became the first to use the word "hacker" in the negative light it now holds.

This event also led Congress to pass several bills around computer crimes, but that did not stop the number of high-profile attacks on corporate and government systems. Of course, the concept of hacking has spiraled with the release of the public internet, which has led to far more opportunities and more lucrative rewards for hacking activity. This saw techniques evolve and increase in sophistication and gave birth to a wide range of types of hacking and hackers.

Types of Hacking/Hackers

There are typically four key drivers that lead to bad actors hacking websites or systems:
(1) financial gain through the theft of credit card details or by defrauding financial services, (2) corporate espionage, (3) to gain notoriety or respect for their hacking talents, and (4) state-sponsored hacking that aims to steal business information and national intelligence. On top of that, there are politically motivated hackers—or hacktivists—who aim to raise public attention by leaking sensitive information, such as Anonymous, LulzSec, and WikiLeaks.

A few of the most common types of hackers that carry out these activities involve:

Black Hat Hackers

Black hat hackers are the "bad guys" of the hacking scene. They go out of their way to discover vulnerabilities in computer systems and software to exploit them for financial gain or for more malicious purposes, such as to gain reputation, carry out corporate espionage, or as part of a nation-state hacking campaign.

These individuals' actions can inflict serious damage on both computer users and the organizations they work for. They can steal sensitive personal information, compromise computer and financial systems, and alter or take down the functionality of websites and critical networks.

White Hat Hackers

White hat hackers can be seen as the "good guys" who attempt to prevent the success of black hat hackers through proactive hacking. They use their technical skills to break into systems to assess and test the level of network security, also known as ethical hacking. This helps expose vulnerabilities in systems before black hat hackers can detect and exploit them.

The techniques white hat hackers use are similar to or even identical to those of black hat hackers, but these individuals are hired by organizations to test and discover potential holes in their security defenses.

Grey Hat Hackers

Grey hat hackers sit somewhere between the good and the bad guys. Unlike black hat hackers, they attempt to violate standards and principles but without intending to do harm or gain financially. Their actions are typically carried out for the common good. For example, they may exploit a vulnerability to raise awareness that it exists, but unlike white hat hackers, they do so publicly. This alerts malicious actors to the existence of the vulnerability.

Other common hacker types include blue hat hackers, which are amateur hackers who carry out malicious acts like revenge attacks, red hat hackers, who search for black hat hackers to prevent their attacks, and green hat hackers, who want to learn about and observe hacking techniques on hacking forums.

Other common hacker types are cyber terrorists, hacktivists, state- or nation-sponsored hackers, script kiddies, malicious insiders, and elite hackers.

Devices Most Vulnerable To Hacking

Smart Devices

Smart devices, such as smartphones, are lucrative targets for hackers. Android devices, in particular, have a more open-source and inconsistent software development process than Apple devices, which puts them at risk of data theft or corruption. However, hackers are increasingly targeting the millions of devices connected to the Internet of Things (IoT).

Webcams

Webcams built into computers are a common hacking target, mainly because hacking them is a simple process. Hackers typically gain access to a computer using a Remote Access Trojan (RAT) in rootkit malware, which allows them to not only spy on users but also read their messages, see their browsing activity, take screenshots, and hijack their webcam.

Routers

Hacking routers enables an attacker to gain access to data sent and received across them and networks that are accessed on them. Hackers can also hijack a router to carry out wider malicious acts such as distributed denial-of-service (DDoS) attacks, Domain Name System (DNS) spoofing, or cryptomining.

Email

Email is one of the most common targets of cyber-attacks. It is used to spread malware and ransomware and as a tactic for phishing attacks, which enable attackers to target victims with malicious attachments or links.

Jail broken Phones

Jail breaking a phone means removing restrictions imposed on its operating system to enable the user to install applications or other software not available through its official app store. Aside from being a violation of the end-user's license agreement with the phone developer, jailbreaking exposes many vulnerabilities. Hackers can target jailbroken phones, which allows them to steal any data on the device but also extend their attack to connected networks and systems.

Prevention from Getting Hacked

There are several key steps and best practices that organizations and users can follow to ensure they limit their chances of getting hacked.

Software Update

Hackers are constantly on the lookout for vulnerabilities or holes in security that have not been seen or patched. Therefore, updating software and operating systems are both crucial to preventing users and organizations from getting hacked. They must enable automatic updates and ensure the latest software version is always installed on all of their devices and programs.

Use Unique Passwords for Different Accounts

Weak passwords or account credentials and poor password practices are the most common cause of data breaches and cyberattacks. It is vital to not only use strong passwords that are difficult for hackers to crack but also to never use the same password for different accounts. Using unique passwords is crucial to limiting hackers' effectiveness.

HTTPS Encryption

Spoofed websites are another common vehicle for data theft, when hackers create a scam website that looks legitimate but will actually steal the credentials that users enter. It is important to look for the Hypertext Transfer Protocol Secure (HTTPS) prefix at the start of a web address. For example: https://www.fortinet.com.

Avoid Clicking on Ads or Strange Links

Advertisements like pop-up ads are also widely used by hackers. When clicked, they lead the user to inadvertently download malware or spyware onto their device. Links should be treated carefully, and strange links within email messages or on social media, in particular, should never be clicked. These can be used by hackers to install malware on a device or lead users to spoofed websites.

Change the Default Username and Password on Your Router and Smart Devices

Routers and smart devices come with default usernames and passwords. However, as providers ship millions of devices, there is a risk that the credentials are not unique, which heightens the chances of hackers breaking into them. It is best practice to set a unique username and password combination for these types of devices.

❖**Measures against Hacking**

There are further steps that users and organizations can take to protect themselves against the threat of hacking.

Download from First-party Sources

Only download applications or software from trusted organizations and first-party sources. Downloading content from unknown sources means users do not fully know what they are accessing, and the software can be infected with malware, viruses, or Trojans.

Install Antivirus Software

Having antivirus software installed on devices is crucial to spotting potential malicious files, activity, and bad actors. A trusted antivirus tool protects users and organizations from the latest malware, spyware, and viruses and uses advanced detection engines to block and prevent new and evolving threats.

Use a VPN

Using a virtual private network (VPN) allows users to browse the internet securely. It hides their location and prevents hackers from intercepting their data or browsing activity.

Do Not Login as an Admin by Default

"Admin" is one of the most commonly used usernames by IT departments, and hackers use this information to target organizations. Signing in with this name makes you a hacking target, so do not log in with it by default.

Use a Password Manager

Creating strong, unique passwords is a security best practice, but remembering them is difficult. Password managers are useful tools for helping people use strong, hard-to-crack passwords without having to worry about remembering them.

Use Two-factor Authentication

Two-factor authentication (2FA) removes people's reliance on passwords and provides more certainty that the person accessing an account is who they say they are. When a user logs in to their account, they are then prompted to provide another piece of identity evidence, such as their fingerprint or a code sent to their device.

Brush Up on Anti-phishing Techniques

Users must understand the techniques that hackers deploy to target them. This is especially the case with ant phishing and ransom ware, which help users know the tell-tale signs of a phishing email or a ransomware attack or ransomware settlements.

What is Ethical Hacking? How Legal is Ethical Hacking?

Ethical hacking refers to the actions carried out by white hat security hackers. It involves gaining access to computer systems and networks to test for potential vulnerabilities, and then fixing any identified weaknesses. Using these technical skills for ethical hacking purposes is legal, provided the individual has written permission from the system or network owner, protects the organization's privacy, and reports all weaknesses they find to the organization and its vendors.

Examples

The biggest hack in history is thought to be the data breach against Yahoo! The 2013 attack compromised around 3 billion people, and the company revealed that every Yahoo! customer was affected by it.

China is believed to be the country with the highest number of dangerous hackers. Most of the major cyber attacks that occurred around the world can be traced back to China.

Insider Threat

An insider threat is a security risk that originates from within the targeted organization. It typically involves a current or former employee or business associate who has access to sensitive information or privileged accounts within the network of an organization, and who misuses this access.

Traditional security measures tend to focus on external threats and are not always capable of identifying an internal threat emanating from inside the organization.

Types of insider threats include:

- **Malicious insider**—also known as a Turncloak, someone who maliciously and intentionally abuses legitimate credentials, typically to steal information for financial or personal incentives. For example, an individual who holds a grudge against a former employer, or an opportunistic employee who sells secret information to a competitor. Turncloaks have an advantage over other attackers because they are familiar with the security policies and procedures of an organization, as well as its vulnerabilities.
- **Careless insider**—an innocent pawn who unknowingly exposes the system to outside threats. This is the most common type of insider threat, resulting from mistakes, such as leaving a device exposed or falling victim to a scam. For example, an employee who intends no harm may click on an insecure link, infecting the system with malware.
- **A mole**—an imposter who is technically an outsider but has managed to gain insider access to a privileged network. This is someone from outside the organization who poses as an employee or partner.

Three types of risky behavior explained

Malicious Insider Threat Indicators

Anomalous activity at the network level could indicate an inside threat. Likewise, if an employee appears to be dissatisfied or holds a grudge, or if an employee starts to take on

more tasks with excessive enthusiasm, this could be an indication of foul play. Trackable insider threat indicators include:

- o **Activity at unusual times**—signing in to the network at 3 am
- o **The volume of traffic**—transferring too much data via the network
- o **The type of activity**—accessing unusual resources

How to Protect Against an Insider Attack: Best Practices

You can take the following steps to help reduce the risk of insider threats:

- ❖ **Protect critical assets**—these can be physical or logical, including systems, technology, facilities, and people. Intellectual property, including customer data for vendors, proprietary software, schematics, and internal manufacturing processes, are also critical assets. Form a comprehensive understanding of your critical assets. Ask questions such as: What critical assets do we possess? Can we prioritize our assets? And, what do we understand about the current state of each asset?
- ❖ **Enforce policies**—clearly document organizational policies so you can enforce them and prevent misunderstandings. Everyone in the organization should be familiar with security procedures and should understand their rights in relation to intellectual property (IP) so they don't share privileged content that they have created.
- ❖ **Increase visibility**—deploy solutions to keep track of employee actions and correlate information from multiple data sources. For example, you can use deception technology to lure a malicious insider or imposter and gain visibility into their actions.
- ❖ **Promote culture changes**—ensuring security is not only about know-how but also about attitudes and beliefs. To combat negligence and address the drivers of malicious behavior, you should educate your employees regarding security issues and work to improve employee satisfaction.

- ❖ **Making Customer Comfortable**

Delivering excellent customer care and proactive customer support make clients feel valued. It is all about knowing what your customers' expectations are and offering the best of your ability.

1. Provide real time support
Businesses can align your customer's expectations with reality by deploying new technology to provide real time support to your customers. Great customer experience can be achieved by using live chat software and live engagement tools that boost customer satisfaction rates.

Live chat

Live chat is the most preferred channel over other communication channels such as phone and email. The real time support it delivers to customers makes it popular. **"79% of customers say they prefer to live chat because of the immediacy it offers compared to other channels."**

- Live chat instantly connects with your customers and assists them in real time.
- You can trigger proactive chat messages to guide customers in their buying journey and improve their experience.

Engagement tools

 In real time by using live customer engagement such as co-browsing, video & voice chat. These tools allow customers to communicate within minimal wait time and delight your customers.

- Video chat allows us to **identify the issue faster and deliver effective solutions**, which reduces the number of touch points and increases customer satisfaction.
- Co-browsing solution allows you to **collaborate with customers and guide them to complete a complex form fill up or application process.**

- Having **direct personalized conversations develops trust in customers** and delivers a delightful customer support experience. Streamline all the customer conversations under one platform and provide a cohesive experience.
- Identify the most preferred channels and be 24×7 active across those channels to reduce average response time.
- Make use of tools like live chat, chatbots, visual tools to gain faster details of the issue and deliver first contact resolution.

2. Deliver consistent omni-channel customer service

- Streamline all the customer conversations under one platform and provide a cohesive experience.
- Identify the most preferred channels and be 24×7 active across those channels to reduce average response time.
- Make use of tools like live chat, chatbots, visual tools to gain faster details of the issue and deliver first contact resolution.

Customer delight example – Bank of America

[Bank of America](#) is one of the biggest known brands following consistent omni-channel service to its customers. The bank allows for everything from depositing checks to scheduling an appointment to be handled by the company's mobile and desktop apps.

Empower your team to delight your customers

- **Freedom of decision making** – The employees hold the complete authority to handle customer's queries independently. It is their responsibility to amaze them by meeting and exceeding customer expectations.
- **Perform as a team** – Empowering your team allows them to perform together to take a move to deliver superior customer service that surpasses the customer delight index.
- **Employees feedback** – The feedback from employees are directly aligned with the company's objective. The mission of the company and the opinion of your team are linked that makes them valued.

Analyze customer feedback

"A satisfied customer is the best business strategy of all."

Customer feedback is crucial for the growth of all businesses. It provides valuable insights into what is working well about your products or services and what should be done to make the experience better.

Analyzing feedback involves **identifying customer needs and frustrations of customers so that businesses can improve customer satisfaction and reduce churn**. Feedback analysis needs to be done wisely by following the below steps.

- **Categorize all feedback into categories** – The feedback may include product delivery speed, after-sales services, customer service approach, etc. Once categorized you can divide further that deserves immediate attention.
- **Identify the nature of the feedback –** Customer feedback comes with negative and positive comments. The positive ones bring in concrete ideas on what can be extremely effective in building customer loyalty. On the other hand, the negative ones provide insights on improvement areas.
- **Consolidate results and plan your next move** – Finally, amalgamate all the results to make a plan of action as to how you intend to respond to each of the issues raised. Make a feasible and effective plan that would address all the problems your clients think your business is having while keeping the good services still functioning.

Note: You need to train your customer support team to ask feedback at the right time via the right channels and by tailoring feedback questionnaires.

Personalize your communication

- Use a tone that matches your customer personality. Some prefer short and direct communications and some like longer conversations and sharing opinions.
- Listen to your customers and empower them. Asking your customers about their preferences both personalizes the experience and builds their confidence in your brand.
- Understanding your customers' backgrounds by having authentic conversations helps to personalize every relationship.
- Make kind gestures by sending follow-up messages to customers after each purchase or service interaction to thank them and offer further assistance if required.
- Recommend products and services to your customers based on their purchase history. Personalized cart recommendations influence 92% of shoppers online.

Empower your customers with communities

Customers love to be part of a community or group. Building communities that benefit customers create a positive feeling and improve your brand image. Communities can be used as a resource for sharing useful information related to products and services.

Branded communities are 13% more likely to have an impact on customer experience than social media communities.

When you foster a special place for customers to interact with one another, your business is adding extra value to the customer experience both before and after the purchase. Generally, customers trust other customers, when they intend to purchase.

So, the sponsored community forum can be used to help to guide in their buying decisions.

7. Do not under promise and over deliver

Don't make a promise you can't keep and keep the ones you make.

Over delivering on customer expectations would raise customer satisfaction and be good for business. It develops trust and loyalty in customers and stays associated with your brand for a lifetime.

To keep customers highly satisfied, you must continue to deliver more value because their expectations will keep increasing.

Note: If you set the tone from the start of under-promising and over-delivering, then your customer is going to expect that same experience of getting more than promised with every interaction. You are setting yourself up to fail and for your customer to be disappointed. A better method might be to deliver on your promises.

8. Listen actively to delight your customers

- Understand your customer needs, expectations, and pain points and align your service to match accordingly to impress them.
- Welcome your customer complaints and feedback and adopt the right tools/process to deliver a better experience.

Actively listening to customers allows you to use the right empathy statements for customer service and deliver a delightful experience.

Stop trying to delight your customers

"To win customer loyalty, forget the bells and whistles and just fix their problems".

The role of motivational customer service can never be discounted as businesses create loyal customers mainly by resolving their issues faster. Customers resent having to contact customer support repeatedly, to get an issue fixed, having to repeat the complete information, and switching from one channel to another.

Final thoughts on customer delight

If you want to be on the right side of the customer service road then get together with your team today and plan different ways on how to impress and delight customers. Following the right strategies will help to deliver a positive customer experience.