**Fundamental of data analytics**

**Unit-I**

Introduction to data analytics:

Data analytics converts raw data into actionable insights. It includes a range of tools, technologies, and processes used to find trends and solve problems by using data. Data analytics can shape business processes, improve decision-making, and foster business growth.

Data analytics is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision making.

5 Types of Data Analytics to Drive Your Business

* Descriptive Analytics. Business intelligence and data analysis rely heavily on descriptive analytics. ...
* Diagnostic Analytics. ...
* Predictive Analytics. ...
* Prescriptive Analytics. ...
* Cognitive Analytics.

Introduction – Importance of Data

“Data is the new oil.” Today data is everywhere in every field. Whether you are a data scientist, marketer, businessman, data analyst, researcher, or you are in any other profession, you need to play or experiment with raw or structured data. This data is so important for us that it becomes important to handle and store it properly, without any error. While working on these data, it is important to know the types of data to process them and get the right results. There are two types of data: Qualitative and Quantitative data, which are further classified into:

The data is classified into four categories:

* Nominal data.
* Ordinal data.
* Discrete data.
* Continuous data.

So there are 4 Types of Data: Nominal, Ordinal, Discrete, and Continuous.

Now business runs on data, and most companies use data for their insights to create and launch campaigns, [design strategies](https://www.mygreatlearning.com/academy/learn-for-free/courses/strategic-management#gl_blog_id=45853), launch products and services or try out different things. According to a report, today, at least 2.5 quintillion bytes of data are produced per day.



Also read: [22 Top Data Science Books – Learn Data Science Like an Expert](https://www.mygreatlearning.com/blog/data-science-books/)

Types of Data

Qualitative or Categorical Data

Qualitative or Categorical Data is data that can’t be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That’s why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.

Qualitative data tells about the perception of people. This data helps market researchers understand the customers’ tastes and then design their ideas and strategies accordingly.

The other examples of qualitative data are :

* What language do you speak
* Favorite holiday destination
* Opinion on something (agree, disagree, or neutral)
* Colors

The Qualitative data are further classified into two parts :

Nominal Data

Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can’t be compared with another color.

The name “nominal” comes from the Latin name “nomen,” which means “name.” With the help of nominal data, we can’t do any numerical tasks or can’t give any order to sort the data. These data don’t have any meaningful order; their values are distributed into distinct categories.

Examples of Nominal Data :

* Colour of hair (Blonde, red, Brown, Black, etc.)
* Marital status (Single, Widowed, Married)
* Nationality (Indian, German, American)
* Gender (Male, Female, Others)
* Eye Color (Black, Brown, etc.)

Ordinal Data

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can’t do any arithmetical tasks on them.

Ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered “in-between” qualitative and quantitative data. The ordinal data only shows the sequences

and cannot use for statistical analysis. Compared to nominal data, ordinal data have some kind of order that is not present in nominal data.

Examples of Ordinal Data :

* When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
* Letter grades in the exam (A, B, C, D, etc.)
* Ranking of people in a competition (First, Second, Third, etc.)
* Economic Status (High, Medium, and Low)
* Education Level (Higher, Secondary, Primary)

Volume. Volume refers to the amount of data that exists. Volume is like the base of big data, as it's the initial size and amount of data that's collected. If the volume of data is large enough, it can be considered big data.

## What are the 3 V's of big data?

The 3 V's (volume, velocity and variety) are three defining properties or dimensions of [big data](https://www.techtarget.com/searchdatamanagement/definition/big-data). *Volume* refers to the amount of data, *velocity* refers to the speed of data processing, and *variety* refers to the number of types of data.

According to the 3 V's model, the challenges of [big data management](https://www.techtarget.com/searchdatamanagement/definition/big-data-management) result from the expansion of all three properties, rather than just the volume alone or the sheer amount of data to be managed. An organization can be better equipped to deal with big data challenges through understanding the 3 V's of big data management.

MEASUREMENT OF DATA UNITS:



Velocity. Velocity in big data refers to how fast data can be generated, gathered and analyzed. Big data does not always have to be used imminently, but in some fields, there is a great advantage to receiving up to the second information about rates and being able to act accordingly.

Difference between supervised and unsupervised learning in machine learning:



Importance of real-time processing in handling data velocity:

Real-time processing plays a crucial role in handling data velocity, especially in scenarios where large volumes of data are generated, collected, and analyzed in near real-time. Here are some key reasons highlighting the importance of real-time processing in handling data velocity:

1. Timely Decision-Making:

 - Real-time processing enables organizations to make decisions promptly based on the most up-to-date information. This is essential in various industries such as finance, healthcare, manufacturing, and e-commerce, where timely decisions can have a significant impact.

2. Competitive Advantage:

 - The ability to process and analyze data in real-time provides a competitive edge. Businesses can respond quickly to market changes, customer preferences, and emerging trends, allowing them to adapt faster than competitors relying on batch processing or slower methods.

3. Improved Customer Experience:

 - Real-time processing allows companies to offer a more personalized and responsive customer experience. For example, in e-commerce, real-time data processing can be used to recommend products, personalize content, and optimize pricing based on customer behavior.

4. Fraud Detection and Security:

 - In industries such as finance, real-time processing is crucial for detecting and preventing fraudulent activities. Analyzing transactions and events in real-time allows for immediate identification of anomalies and potential security threats.

5. IoT and Sensor Data:

 - With the proliferation of IoT devices and sensors, there is a massive influx of data from various sources. Real-time processing is essential to handle the continuous stream of data generated by these devices, enabling quick responses to changing conditions, monitoring equipment health, and ensuring efficient operations.

6. Supply Chain Optimization:

 - Real-time processing is vital for optimizing supply chain operations. It helps in tracking inventory levels, monitoring shipping routes, and adjusting production schedules based on real-time demand, contributing to overall efficiency and cost savings.

7. Streaming Analytics:

 - Real-time processing is fundamental for streaming analytics, where data is processed as it is generated. This is particularly useful for monitoring live events, social media feeds, and other dynamic data sources, allowing organizations to extract valuable insights in real time.

8. Preventing Data Overload:

 - In high-velocity data environments, the sheer volume of incoming data can be overwhelming. Real-time processing allows organizations to filter and prioritize data, ensuring that only relevant and actionable information is processed, reducing the risk of data overload.

9. Adaptive Systems:

 - Real-time processing facilitates the development of adaptive systems that can adjust their behavior dynamically based on changing conditions. This is valuable in applications such as smart grids, traffic management, and autonomous vehicles.

how data veracity is different from data quality in the context of analytics:

Data veracity and data quality are related concepts, but they address different aspects of data in the context of analytics. Let's explore the differences between the two:

1. Data Quality:

 -Definition: Data quality refers to the overall reliability, accuracy, completeness, consistency, and relevance of data. It assesses the degree to which data meets the requirements and expectations of users. Quality issues can include errors, inconsistencies, missing values, and inaccuracies.

 - Focus: The focus of data quality is on ensuring that the data is fit for its intended purpose. This involves measures to clean, standardize, and validate the data to enhance its trustworthiness.

2. Data Veracity:

 -Definition: Data veracity specifically deals with the trustworthiness and reliability of the data. It addresses the uncertainty and potential biases that may exist in the data. Veracity considers the degree to which data can be trusted for making decisions and drawing accurate conclusions.

 Focus: Veracity is concerned with assessing the authenticity, reliability, and credibility of data. It acknowledges that data can be imperfect, and efforts are made to understand, quantify, and mitigate uncertainties.

Automation:

The term "automation" refers to the use of technology and software tools to streamline and perform tasks related to data processing, analysis, and decision-making without manual intervention. Automation in data analytics is aimed at improving efficiency, reducing human error, and accelerating the overall data-driven decision-making process. Here are some key aspects of automation in data analytics:

1. Data Collection and Ingestion: - Automated processes can be set up to collect and ingest data from various sources. This includes data extraction from databases, APIs, flat files, streaming sources, and other relevant data repositories.

2. Data Cleaning and Preprocessing:- Automation tools can be employed to clean and preprocess raw data, addressing issues such as missing values, outliers, and inconsistencies. This ensures that the data is in a suitable format for analysis.

3. Data Transformation:- Automation facilitates the transformation of data into a format that is conducive to analysis. This involves tasks such as normalization, aggregation, and creating derived features.

4. Data Analysis and Modeling:- Automated analytics tools can perform statistical analysis and build machine learning models. This includes tasks such as regression analysis, clustering, classification, and predictive modeling without requiring constant manual intervention.

5. Report Generation and Visualization: - Automation can be applied to generate reports and visualize data insights. Dashboards and reports can be automatically updated with the latest data, providing stakeholders with real-time information.

6. Alerts and Notifications:- Automated systems can be set up to monitor data in real-time and trigger alerts or notifications based on predefined conditions. This proactive approach allows for quick responses to changes or anomalies in the data.

7. Deployment of Models: - In machine learning and predictive analytics, automation plays a role in deploying models into production environments. Automated deployment ensures that models are seamlessly integrated into operational systems.

8.Data Governance and Compliance:- Automation can enforce data governance policies and compliance measures. This includes ensuring data quality, privacy, and security standards are met throughout the analytics process.

The use of automation in data analytics is essential for handling large volumes of data efficiently, reducing time-to-insight, and enabling data scientists and analysts to focus on higher-value tasks such as interpretation, strategy, and decision-making. It contributes to the agility and scalability of analytics processes in today's data-intensive environments.

Internet and social media has contributed to the generation of Big Data:

The growth of the internet and the widespread adoption of social media platforms have significantly contributed to the generation of Big Data. Several factors highlight how these developments have led to the massive accumulation of data:

1. Increased Digital Connectivity:- The expansion of the internet has brought about increased connectivity worldwide. More people are accessing the internet through various devices, leading to a surge in digital interactions, transactions, and communications.

2. Social Media Usage:- The rise of social media platforms, such as Facebook, Twitter, Instagram, LinkedIn, and others, has resulted in billions of users sharing vast amounts of content daily. Users create and share text, images, videos, and other multimedia content, contributing to a continuous stream of data.

3. User-Generated Content:- Social media relies heavily on user-generated content. Individuals share their thoughts, opinions, experiences, and preferences in real-time. This content, when aggregated, forms a rich source of information for analysis.

4. E-commerce and Online Transactions:- The growth of e-commerce has led to an increase in online transactions. Each transaction generates data, including purchase history, user preferences, and demographic information, contributing to the volume of Big Data.

5.Mobile Devices and Apps:- The widespread use of smartphones and mobile apps has enabled people to access the internet and social media on the go. Mobile devices generate data such as location information, app usage patterns, and sensor data, further contributing to the pool of Big Data.

6. IoT Devices:- The Internet of Things (IoT) involves the connection of various devices to the internet, ranging from smart home appliances to industrial sensors. These devices continuously generate data, adding to the overall volume and variety of Big Data.

7. Streaming Data: - Social media platforms and other online services generate a continuous stream of real-time data. This includes updates, comments, likes, and other interactions, requiring the ability to process and analyze data in real-time, which is a characteristic of Big Data.

8. Data Variety:- The data generated through the internet and social media is diverse in nature. It includes structured data (e.g., user profiles), semi-structured data (e.g., JSON or XML files), and unstructured data (e.g., text in posts, images, and videos), contributing to the variety aspect of Big Data.

9. Data Velocity:- The speed at which data is generated and shared on the internet and social media is high. Real-time updates, instant messaging, and rapid content sharing contribute to the velocity of data, requiring efficient processing and analysis.

10.Data Accessibility: - The ease of access to online platforms encourages more people to contribute data. Social media platforms are designed to be user-friendly, encouraging users to share and engage with content regularly.

The combined effect of these factors has led to the creation of vast and complex datasets, commonly referred to as Big Data. Analyzing this data provides valuable insights into user behavior, market trends, sentiment analysis, and more, offering businesses and researchers unprecedented opportunities for informed decision-making and innovation.

**"Big Data" and list its three primary characteristics:**

The term "Big Data" refers to large and complex datasets that cannot be effectively managed, processed, and analyzed using traditional data processing tools and methods. Big Data is characterized by three primary dimensions, often referred to as the "3Vs":

1. Volume:

 - Definition: Volume refers to the sheer size of the data generated or collected. Big Data involves massive amounts of information that exceed the capacity of traditional databases and processing systems.

 - Example:Social media platforms, e-commerce websites, sensor networks, and scientific research generate and store enormous volumes of data on a daily basis.

2. Velocity:

 - Definition: Velocity represents the speed at which data is generated, processed, and made available for analysis. Big Data is often characterized by the rapid and continuous flow of data that requires real-time or near-real-time processing.

 - Example: Streaming data from social media updates, online transactions, sensor readings, and clickstream data from websites are instances where data velocity is high.

3. Variety:

 - Definition: Variety refers to the diversity of data types and sources. Big Data includes a wide range of data formats, including structured, semi-structured, and unstructured data. It may originate from different sources, such as text, images, videos, and sensor readings.

 - Example: Structured data may include databases and spreadsheets, while semi-structured data may come in the form of JSON or XML files. Unstructured data encompasses text in social media posts, images, and videos.

These three characteristics, Volume, Velocity, and Variety, collectively define the challenges and opportunities presented by Big Data. As the volume of data continues to grow at an unprecedented rate, the need for advanced tools and technologies to handle the velocity and variety of data becomes crucial. In addition to the 3Vs, other characteristics like Veracity (data accuracy and reliability) and Value (the ability to derive meaningful insights) are sometimes considered to provide a more comprehensive understanding of Big Data.

**The key components that contribute to the generation of large volumes of data:**

Several key components contribute to the generation of large volumes of data in today's interconnected and digital world. The increasing use of technology, the proliferation of devices, and the growth of online activities have played a significant role in the exponential increase in data generation. Here are some key components:

1. Internet Usage:- The widespread adoption of the internet has led to a surge in data generation. Activities such as web browsing, online searches, and content consumption contribute to the continuous creation of data.

2. Social Media: - Social media platforms, including Facebook, Twitter, Instagram, LinkedIn, and others, have billions of active users who generate vast amounts of data through posts, comments, likes, shares, and multimedia content.

3. Mobile Devices:- The ubiquity of smartphones and other mobile devices has facilitated on-the-go internet access and app usage. Mobile devices generate data related to location, app usage, communication, and sensor readings.

4. E-commerce Transactions:- The growth of online shopping and e-commerce platforms has resulted in a significant increase in transactional data. Each purchase, online payment, and interaction on e-commerce websites generates data.

5. IoT Devices:- The Internet of Things (IoT) involves the connection of various devices to the internet, such as smart appliances, wearable devices, industrial sensors, and more. These devices continuously generate data, contributing to the overall data volume.

6. Streaming Services:- The popularity of streaming services for music, videos, and other media has led to the continuous generation of streaming data. This includes real-time content consumption and user interactions.

7. Scientific Research and Simulations: - Scientific experiments, simulations, and research activities in fields such as astronomy, genomics, climate modeling, and high-energy physics generate massive datasets.

8. Business Operations:- Day-to-day business operations, including customer interactions, financial transactions, supply chain activities, and human resources data, contribute to the large volumes of enterprise data.

9. Healthcare Data:- Electronic health records, medical imaging, wearable health devices, and health monitoring systems generate substantial amounts of data in the healthcare sector.

10. Government and Public Sector:- Government agencies and public sector organizations collect and generate large volumes of data related to demographics, public services, transportation, and security.

11. Educational Systems:- Educational institutions generate data related to student enrollment, academic performance, administrative operations, and online learning platforms.

12. Search Engines and Online Platforms:- Search engines process massive amounts of search queries, and online platforms host a vast array of content, both of which contribute significantly to data generation.

The collective impact of these components results in an ever-growing and diverse landscape of data. As technology continues to advance, new components and sources of data are likely to emerge, further contributing to the expansion of large volumes of data.

The needs of sophisticated consumers in data analytics differ from those of casual users based on the depth of analysis, technical proficiency, and specific requirements. Here are some key distinctions between the two groups:

 Sophisticated Consumers in Data Analytics:

1. Advanced Analysis:

 - Sophisticated consumers:Require advanced analytical capabilities, including statistical modeling, machine learning, predictive analytics, and complex data mining techniques.

 -Activities: Engage in in-depth exploratory data analysis, hypothesis testing, and building sophisticated models to uncover intricate patterns and insights.

2. Technical Proficiency

 - Sophisticated consumers:Typically possess a higher level of technical proficiency, including knowledge of programming languages (e.g., Python, R), database systems, and statistical methodologies.

 Tools:Are comfortable using advanced analytics tools and platforms that offer a wide range of functionalities.

3. Customization and Flexibility:

 - Sophisticated consumers: Require the ability to customize analyses, algorithms, and workflows to meet specific business or research requirements.

 - Data Manipulation: May engage in complex data manipulation and transformation to derive meaningful insights from raw data.

4. Data Integration:

 - Sophisticated consumers: Often work with data from multiple sources, requiring advanced data integration and merging capabilities to create comprehensive datasets.

 - Complex Ecosystems:Deal with complex data ecosystems involving structured, semi-structured, and unstructured data.

5. Real-time and Streaming Analytics:

 - Sophisticated consumers: May require real-time or streaming analytics capabilities for immediate insights and decision-making.

 - Advanced Use Cases:Use cases may involve monitoring and analyzing data in real-time for applications like fraud detection, IoT, and financial trading.

6. Data Governance and Compliance:

 - Sophisticated consumers: Place a high emphasis on data governance, compliance, and security measures, especially when dealing with sensitive or regulated data.

 - Ethical Considerations: Are often concerned with ethical considerations related to data privacy and responsible data usage.

Sophisticated consumers in data analytics differ from those of casual users

 Casual Users in Data Analytics:

1. User-Friendly Interfaces:

 - Casual users: Prefer user-friendly interfaces and tools with intuitive dashboards, visualizations, and easy-to-understand features.

 - Accessibility: Seek tools that don't require extensive training and are accessible to users with varying levels of technical expertise.

2. Pre-built Templates and Reports:

 - Casual users: Benefit from pre-built templates, reports, and predefined analytics workflows that simplify data analysis.

 - Guided Analytics: Appreciate guided analytics features that help them step through the analysis process.

3. Basic Descriptive Analytics:

 -Casual users: Primarily engage in basic descriptive analytics, such as generating summary statistics, charts, and simple visualizations.

 -Common Metrics: Focus on commonly used metrics and KPIs to monitor performance and trends.

4. Decision Support:

 - Casual users: Rely on data analytics tools to support day-to-day decision-making rather than complex strategic planning.

 - Quick Insights: Appreciate tools that quickly provide insights without delving into intricate data analysis.

5.Data Exploration:

 - Casual users: Engage in exploratory data analysis to gain a general understanding of the data but may not delve into deeper analytical techniques.

 - Ad Hoc Queries: Often perform ad hoc queries to answer specific questions without requiring advanced analytical methodologies.

6. Time Efficiency:

 - Casual users:Value tools that allow them to quickly analyze data and obtain insights without investing a significant amount of time.

 - Efficient Workflows: Seek streamlined workflows for routine tasks and analyses.

The needs of sophisticated consumers in data analytics are characterized by a demand for advanced analysis, technical depth, customization, and real-time capabilities. On the other hand, casual users prioritize user-friendly interfaces, simplicity, pre-built templates, and basic descriptive analytics to support their more straightforward analytical needs. The distinction between these two user groups underscores the importance of providing diverse tools and solutions that cater to a broad spectrum of analytics users within an organization.