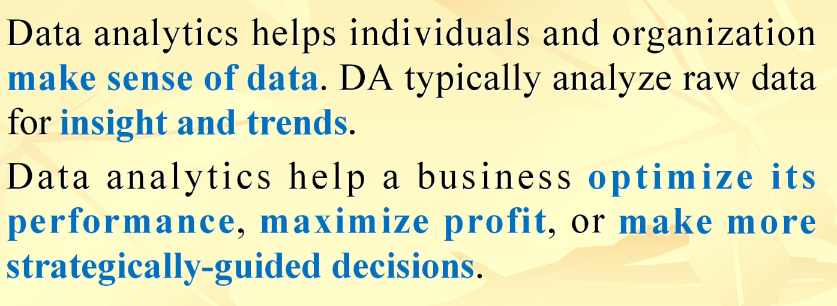
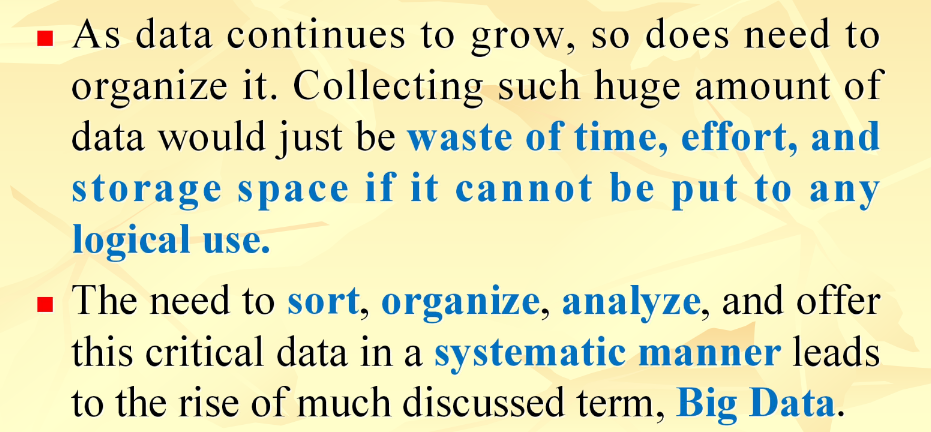
DA - Part - I

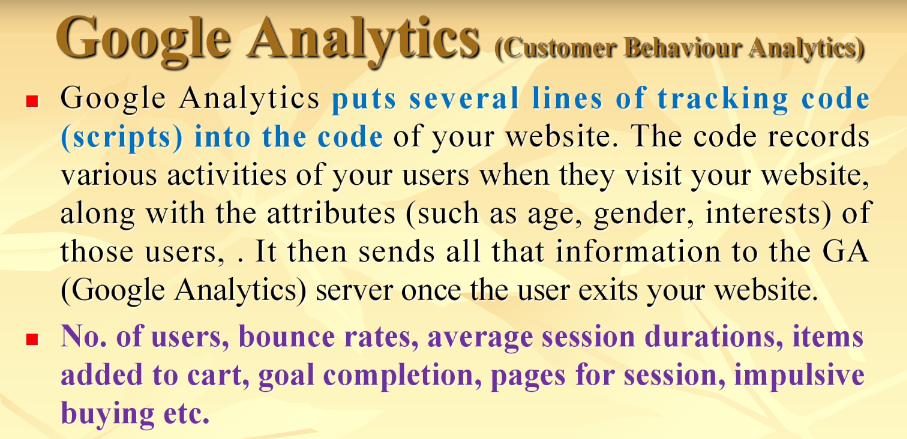
1. Data Analytics - Definiton : -



1. Why Big Data -



1. Real World Examples - pg - 7 to 10 give it a read
2. Google Analytics -   
   



[Unread]

1. The 5 V’s of Big Data -

The 5V's of Big Data are:

**Volume**:

Refers to the scale of data being generated or collected, which is typically massive and growing exponentially.

Big data often exceeds the capacity of traditional data storage and processing systems, requiring specialized tools and technologies.

Volume is a critical factor in big data analytics, as it affects the speed and accuracy of insights that can be derived.

**Velocity**:

Refers to the speed at which data is being generated, collected, and processed, often in real-time.

Big data analytics requires processing large amounts of data in a short amount of time to gain insights quickly.

Velocity is important for time-sensitive applications such as fraud detection, recommendation engines, and supply chain optimization.

**Variety**:

Refers to the diversity of data types and sources, including structured and unstructured data, such as text, images, audio, and video.

Big data often involves integrating data from multiple sources, including social media, sensors, and mobile devices, which can be in different formats and structures.

Variety poses a challenge for data integration and analysis, requiring new approaches to data modeling, storage, and processing.

**Veracity**:

Refers to the quality and accuracy of data, which can vary widely depending on the source and context.

Big data often includes noisy, incomplete, or conflicting data, which can lead to inaccurate or biased insights.

Veracity is critical for ensuring the reliability and validity of big data analytics, requiring data cleansing, quality assurance, and data governance.

**Value**:

Refers to the usefulness and relevance of data for solving business problems or creating new opportunities.

Big data analytics aims to extract valuable insights from data that can inform strategic decisions, optimize operations, and drive innovation.

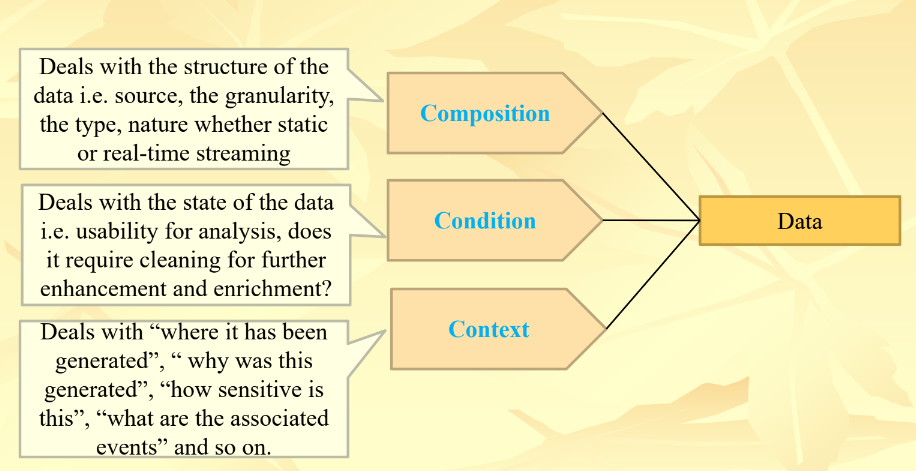
Value is the ultimate goal of big data, requiring a clear understanding of business objectives, data needs, and analytical techniques.

Common points that might be included in all -

1. Large amount of data is being generated, collected and processed in real time
2. Data analysis requires processing large amounts off data within a short period of time to get a quick view of the insights.

unread

1. Characteristics of Data -



unread

1. Big Data - defnition

High volume - high velocity - high variety - high veracity - data assets - **demand** - cost effective - innovative forms - information processing - enhanced insights - decision making

Try to re-approach it

1. Big Data = structured data + unstructured data + semi-structured data.
2. Why is big data is required?

A - Big data is required because it offers several benefits and advantages to organizations and individuals, including:

1. **Improved decision-making**: Big data provides a wealth of information and insights that can help organizations make better and more informed decisions. By analyzing large volumes of data, organizations can identify patterns, trends, and correlations that might otherwise go unnoticed.
2. **Increased efficiency**: Big data can help organizations streamline their operations and reduce costs. By analyzing data on their processes, organizations can identify areas of inefficiency and implement changes to improve productivity and reduce waste.
3. **Enhanced customer experience**: Big data can provide valuable insights into customer behavior and preferences, enabling organizations to tailor their products and services to meet their customers' requirements scaling their profit.
4. **Competitive advantage**: Big data can give organizations a competitive edge by providing insights into market trends, customer behavior, and industry dynamics. By leveraging big data, organizations can identify new opportunities and stay ahead of their competitors.
5. **Innovation**: Big data can provide a rich source of inspiration for innovation and new product development. By analyzing data on customer needs and preferences, organizations can identify unmet needs and develop new products and services to meet those needs.
6. Explain the Structured, semi-structured, and unstructured data of Big Data.

A -

**Structured Data**:

Structured data is **highly organized** and **follows a pre-defined format**. It is easy to store, process, and analyze because it is consistent and has a clear structure.

Advantages:

- Easy to process and analyze

- Enables quick data retrieval

- Easy to understand and use

Enables easy integration with other structured data sources

Provides accurate results for reporting and analysis

Disadvantages:

- Limited flexibility and adaptability

- Can be expensive to maintain and manage

Does not handle complex relationships well

Limited insights due to rigid structure

Examples:

Transaction data (sales orders, invoices, payments)

Inventory data (part numbers, quantities, locations)

Employee data (name, address, salary)

**Semi-Structured Data:**

Semi-structured data has some organization but does not follow a strict format. It may include elements that are organized into tables or fields, but also includes free-form data such as text.

Advantages:

Offers more flexibility than structured data

Enables the inclusion of unstructured data elements

Supports ad hoc queries and analysis

Provides more insights than structured data

Disadvantages:

Requires more processing power and storage than structured data

Can be more difficult to analyze than structured data

May have inconsistencies due to lack of structure

Examples:

XML files

JSON files

Log files

**Unstructured Data:**

Unstructured data has no organization or structure and may include text, images, audio, and video. It is the most challenging type of data to analyze, but it often provides valuable insights.

Advantages:

Provides rich insights into customer behavior and preferences <--

Enables sentiment analysis and other text mining techniques

Provides a wealth of information for machine learning and predictive analytics <---

Enables analysis of large volumes of data <--

Disadvantages:

Difficult to process and analyze due to lack of structure

Requires specialized tools and techniques for analysis

May have inconsistencies and errors

Requires significant storage capacity

Examples:

Social media data (tweets, posts, comments)

Sensor data (temperature, pressure, humidity)

Emails and other forms of text data.

A very important term to be remembered in DA -

***Customer pattern, behaviour , insights and preference.***

\*\*this might be helpful in many places.

1. Classification of Digital data -



1. Challenges of Traditional Systems -

A - Traditional systems cannot perform efficiently on unstructured data

B - The data sets are usually of high volume, velocity and variety which need high computing resources to be processed. So traditional systems like databases, servers, networks are not sufficient enough to handle this large data assets.

1. **Storage capacity** : Traditional data systems are not designed to handle large volumes of data, and as a result, they lack the storage capacity required to manage big data. Big data requires massive storage capacity, which can be challenging for traditional systems to provide.
2. **Processing power**: Traditional data systems also lack the processing power needed to manage big data. Big data requires complex algorithms and data processing techniques that can be time-consuming and resource-intensive. Traditional systems are often not capable of handling these requirements.
3. **Data integration**: Big data often comes from multiple sources, including structured and unstructured data, and can be in different formats and structures. Traditional systems may not be able to integrate these disparate data sources effectively, leading to data silos and reduced insights.
4. **Real-time analysis**: Big data analytics often requires real-time or near real-time analysis, which can be a challenge for traditional data systems. These systems may not be able to process data quickly enough to provide real-time insights, limiting their usefulness for time-sensitive applications.
5. Data security: Traditional data systems may not be equipped to handle the security challenges posed by big data. Big data often includes sensitive information, such as personal data and financial information, which requires robust security measures to protect against breaches and unauthorized access.
6. Evolution of analystics scalability

**Analytics scalability in Big Data refers to the ability to efficiently process and analyze large volumes of data as the size of the dataset increases. It involves the *ability to scale data processing and analytics resources*, including hardware, software, and personnel, to match the growing volume of data.**

In traditional data systems, scalability can be limited by the capacity of the hardware and software used for data processing and analytics. However, Big Data systems are designed to be highly scalable, allowing organizations to process and analyze massive amounts of data.

The technologies are:

- MPP (massively parallel processing)

- Cloud computing (Appendix)

- Grid computing

- MapReduce (Hadoop)

1. Distributed VS Parallel Computing.
2. Explain Traditional Analytics Architecture and Modern in-Database Analytics architecture?

A - **Traditional Analytics Architecture**:

The traditional analytics architecture is typically characterized by a two-tiered approach where data is extracted from source systems and stored in a separate data warehouse or data mart. The data is then processed using analytics tools, which are typically installed on separate servers or workstations, and the results are displayed in reports or dashboards. The key components of traditional analytics architecture are:

1. Extract, Transform, and Load (ETL) processes: ETL processes are used to extract data from source systems, transform it into a format suitable for analysis, and load it into a data warehouse or data mart.
2. Data warehouse or data mart: Data warehouses or data marts are used to store large volumes of structured data, which can be analyzed using analytics tools.
3. Analytics tools: Analytics tools are used to process data and generate insights. These tools are typically installed on separate servers or workstations.
4. Reports and dashboards: Reports and dashboards are used to display the results of analytics processes.

**Modern in-Database Analytics Architecture**:

In-database analytics architecture is a modern approach to analytics that leverages the power of relational database systems to perform analytics tasks. In this architecture, analytics functions are executed inside the database, eliminating the need to extract and transfer data to a separate analytics system. The key components of in-database analytics architecture are:

1. Relational database management system (RDBMS): RDBMS is used to store and manage data.
2. In-database analytics functions: In-database analytics functions are built into the RDBMS and are used to perform analytics tasks directly in the database.
3. Analytics tools: Analytics tools are used to interact with the in-database analytics functions and display the results of analytics processes.
4. Data visualization tools: Data visualization tools are used to create reports and dashboards to display the results of analytics processes.

Advantages of Modern In-Database Analytics Architecture over Traditional Analytics Architecture:

Improved Performance: In-database analytics architecture can deliver better performance by performing analytics tasks directly in the database. This eliminates the need to transfer data to a separate analytics system, which can be time-consuming and resource-intensive.

Scalability: In-database analytics architecture can be more scalable than traditional analytics architecture. As the amount of data grows, in-database analytics functions can be distributed across multiple database servers, providing better performance and scalability.

Reduced Complexity: In-database analytics architecture can reduce the complexity of the analytics process by eliminating the need for separate ETL processes, data warehouses, or data marts.

Real-time analytics: In-database analytics architecture can enable real-time analytics by performing analytics tasks directly in the database as data is being generated.

1. Differences between Parallel Computing and Distributed Computing

**Parallel computing**

In parallel computing multiple processors perform multiple

tasks assigned to them simultaneously. Memory in parallel

systems may be shared or distributed.

In parallel computing, multiple operations are executed parallelly using multiple processors in a single computer located a single location. They have shared and distributed memory.

**Distributed computing**

It consists of multiple autonomous computers which appear

like a single system. So there is no shared memory and

computes communicate (message passing) with each other

using network. In distributed computing, multiple operations are running on multiple processors on multiple autonomous computers on multiple location. They communicate with each other over the internet. They only have distributed memory

**Difference -**

1. In parallel computing many operations are performed

simultaneously. In distributed computing system components are located at different locations.

2. In parallel computing single computer is used while

distributed computing uses multiple computers.

3. In parallel computing multiple processors perform

multiple operations and in distributed computing multiple

computers perform multiple operations.

4. Parallel computing may use shared or distributed memory but distributed computing has to use only distributed memory.

5. Processors communicate with each other through bus in

parallel computing. Computers communicate with each other

through message passing in distributed computing.

6. Parallel computing improves the system performance but

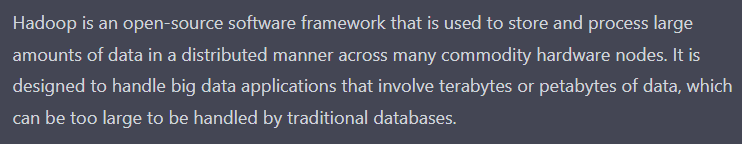
distributed computing improves system scalability, fault

tolerance and resource sharing capabilities.

1. Massively Parallel Processing System -

It is a storage structure designed to handle the coordinated processing of program operations by multipleprocessors.Thiscoordinated processing can work on different parts of a program, witheachprocessor using its own operating systemand memory. ThisallowsMPPdatabases to handle massive amounts of data and provide muchfasteranalyticsbased on large datasets.

Hadoop -



Why Hadoop

1. Low cost
2. Storage Flexibility
3. Scalability
4. Computing power

Pig -