**Assignment -1**

**Santosh Kumar Velgapuri**

**BIG DATA MARKDOWN TOPICS**

**- Big Data with example and types**

Big Data refers to enormous, complicated datasets that are challenging to process and analyze using conventional techniques. It is a vast collection of both structured and unstructured data that has been gathered from numerous websites, social media platforms, and business applications. Big data can be applied in many different contexts, such as assessing customer interactions, projecting market changes, and predicting consumer behaviour and trends.

Types of big data:

Batch processing:

Data processing in batches is a technique applied in Big Data applications. It entails gathering a large amount of data, processing it in phases, and then archiving the results. Here the incoming large data is analyzed using batch processing instead of separating the data into smaller sections. Data processing in batches is effective because it relieves the system from having to read and write data all the time. Instead, the data can be handled in a single step, making it more time and resource efficient.

Streaming Data:

Streaming data is a type of data which is produced continually by numerous sources, including sensors, Internet of Things (IoT) devices, and online applications. This kind of large data is produced in real-time and must be evaluated fast in order to yield insights and prompt action. Various sources, including social media, financial transactions, and IoT, can produce streaming data. These data sets are often enormous, therefore processing and analyzing those calls for specific tools and methods. Example as discussed in class for twitter the hashtags and hot topics of Real-time trend and anomaly detection and the creation of predictive models are also possible with streaming data. Furthermore, it can be used to create data pipelines and data-driven applications.

Graph Data:

Big data that is organized into nodes and edges is known as graph data. In order to uncover patterns, correlations, and other insights within the data, graphs can be used to show relationships between things like people, places, or objects. Each node in a graph contains information about itself, such as its age, location, or social connections, while each edge represents a connection between two nodes.  Businesses may get more information from their data and make better decisions by utilizing graph data.

Spatio-temporal:

Spatio-temporal big data refers to big data that emphasizes the spatial and temporal dimensions of data. It stands out due to its size, speed, and need for tremendous precision and detail in order for it to be useful. Forecasting the weather and tracking the spread of disease are two examples of applications using spatial-temporal data. In addition, it's used to map cities, monitor environmental changes, and enhance public transportation by examining traffic patterns. example tracking the location of NASA satellite for every instance is also a spatio temporal data.  Organizations can utilize spatiotemporal data to analyze the distribution of news and other media across diverse locations and develop effective marketing strategies.

**- 6 ‘V’s of Big Data (define each)**

Volume- The Volume in Big data which deals with the amount of data produced from many sources, dealing with storage using various tools and frameworks that are required to process and analyse such data.

Velocity - The velocity of big data deals with the speed at which the data is created, captured and streamed like real time streaming data. The data which is generated is at a very high speed like the social media data and capturing it.

Veracity - The veracity of big data deals with analysis performed on the data and increasing the quality of data by cleaning and removing the noise of data. It mainly refers to the quality on how accurate is the data and the degree to which the data can be trusted for usage.

Variety - The vareity of data deals with the forms of data. Structured data, semi structured data and unstructured data are various formats present in big data. They include data of sensors, images, videos, text and audios.

Variability - The variability of big data deals with the ways in which the data can be used and formatted. As the data collected may be inconsistent, it can damage process to handle and manage like the example of sensor data whose data quality might not be constant. The data changes during the processing and life cycle so it is inconsistent which leads to dealing with other ways to process.

Value - The value of big data deals with the usefulness of data for the intended purpose. The more value that data has the more business value it gets and more information can be extracted from the data. Volume, velocity, veracity, variety, and variability are components that adds the value to data.

**- Phases of Big Data analysis (discuss each)**

Big Data has 5 major phases:

Phase 1: Data Acquisition and Recording

Phase 2: Information Extraction and Cleaning

Phase 3: Data Integration, Aggregation, and Representation

Phase 4: Query Processing, Data Modeling, and Analysis

Phase 5: Interpretation

There are 5 main phases of big data analysis, they are

**Data Acquisition and Recording:**

Data acquisition and recording are the processes of collecting and storing data from various sources. Data can be collected from a variety of sources such as sensors, Internet of Things (IoT) devices, and other online applications and also the manual entry. Once it is acquired, the data is then stored in a database for later use. Data acquisition and recording are important for a variety of applications, such as analyzing trends, making decisions, tracking performance and more. Most of the data collected is of no use. We need to know our goals such that we can compress the data using filtration of used information and also to make sure we don’t discard the useful information. Few of the sources generate the Meta data automatically.

**Information extraction and Cleaning:**

After acquisition of data and making filtration placing the goals in mind, now the useful information is extracted from the raw data and organizing the required information in useful structured annotation. The extraction can be done manually and automated format. Automation is done using advanced algorithms, tools and using natural language processing (Discussed in class).

Data Cleaning is a process of making the data to near accuracy by filling missing values, clearing the noise and completeness of extracted data. At this step we need to make sure that the data after cleaning is of no errors. Data cleaning also involves and ensures that the data is up-to-date in order to avoid errors. Else the results with an inaccurate date always leads to flaws and wrong results.

**Data Integration, Aggregation and Representation:**

Data Integration:

Data integration creates a single fact base for analytics by combining multiple types and formats of data from any source across an organization into a data lake or data warehouse. Making decisions based on this single data collection enables firms to improve customer experience, departmental coordination, and decision-making.

Data Aggregation:

Data aggregation refers to any procedure that gathers data and presents it in a highly efficient manner. When data is aggregated, totals or summary statistics are used in place of the often assembled atomic data rows. Instead of using collections of observable aggregates, summary statistics derived from those data are used. Aggregate data is typically found in a data warehouse since it can quickly answer analytical queries and greatly reduce the time needed to query large data sets.

Data aggregation is frequently used to produce relevant summary data for business analysis and to provide statistical analysis for groups of people. Data aggregators are software tools that are frequently used for large-scale aggregation. Tools for gathering, analyzing, and presenting data

Data Representation:

Data representation is the process of organizing and structuring data in a way that makes it easier to understand, analyze, and interpret. It involves transforming raw data into an organized form that makes it easier to analyze, visualize, and extract meaningful insights from the data. Data representation can take many forms, such as tables, graphs, charts, diagrams, and maps.

Data can be represented in many ways such as events, objects, and ideas. Data can be anything from a name to a number to the hues in an image to the notes in a song

**Query Processing, Data Modelling And Analysis:**

Query processing:

**Query processing** is a process of getting relative or useful information from the dataset by using interpretation queries. It allows the analyst to optimise his time an efforts in getting relative information from large set of data.

data modelling:

Building a model to represent data and its relationships is called "**data modelling**.". using the modelling we can create a data structure that can be used to create the meaningful representation of the data along various fields and various data sets like E-R representation, Relational database model and semantic data.

After all this we need to analyse the data using various tools to get the conclusion, we need to go through various processes of inspecting, transformation tracking, insight knowledge and modelling of data. This not only involves SQL querying but much more to analyse the data. Some analysis are not easy in SQL like unstructured querying so we might use various database querying in this process.

Interpretation:

This is the last phase of data analysis where any individual analyzes the data. Here, we can develop insightful hypotheses and reach conclusions. At this stage, data patterns are identified, conclusions are drawn, and suggestions are made forth. This stage allows for decision-making, strategy development, and error minimization, all of which contribute to increased corporate efficiency.

**- Challenges in Big Data analysis (discuss each)**

Heterogeneity:

Big data is distinguished by its heterogeneity and incompleteness because of the enormous volume of data it contains that was gathered from numerous sources. The variety and incompleteness of the data create numerous challenges for data analysis and interpretation, including issues with data integration, quality, privacy, representation, and correctness.

Scaling:

Scaling is one of the major challenge in the big data. Companies must upgrade and adapt their systems in order to properly handle and analyze vast volumes of data. Big data's difficulty is further increased by the data's complexity. It is challenging to deal with and evaluate the data since it is frequently unstructured and changing. In order to properly handle the complexity and scope of big data, companies must adopt strategies.

Timeliness:

A unique difficulty in massive data is timeliness. Due to the abundance of data, it may be challenging for businesses to swiftly assess it and take the appropriate action. Also, if data is not updated often, it may quickly become outdated. To solve this issue, businesses must put in place efficient data management systems that transport data quickly across sources. Businesses should be able to respond quickly on any new data insights they find in order to stay ahead and compete in the market.

Privacy:

Big data poses a number of privacy challenges, one of which is figuring out how to maintain data security while also allowing for public access. It gets more challenging to maintain data security and privacy as the amount of data being gathered, saved, and analyzed keeps rising. Big data systems frequently involve huge, distributed databases that store delicate personal data; these databases need to be protected from unauthorized access and use.

Human Collaboration:

Across industries, organizations are utilizing big data to address complex problems, but the success of the data-driven solutions they generate depends on how well people and machines collaborate during the development process. It is critical for people to grasp the data and insights that machines may supply in order to use them effectively. Machines have little autonomy in making decisions, thus humans must provide the necessary context and ethical considerations on their behalf.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_X\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_