Unit I: Introduction to Data Science and Big Data (6 Hrs)

1. Introduction to Data Science and Big Data

- **Data Science** is an interdisciplinary field that uses scientific methods, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- **Big Data** refers to datasets that are so large or complex that traditional data processing applications are inadequate.Studocu+1College Circulars+1

2. Defining Data Science and Big Data

- **Data Science** combines aspects of statistics, computer science, and domain expertise to analyze and interpret complex data.
- **Big Data** is characterized by the three Vs: Volume, Velocity, and Variety, which describe the massive amount of data, the speed at which it is generated, and the different types of data, respectively.

3. Big Data Examples

- **Social Media Data**: Platforms like Facebook and Twitter generate vast amounts of user data daily.
- **Sensor Data**: IoT devices collect data from various sensors in real-time.
- **Transactional Data**: E-commerce websites record user transactions, clicks, and preferences.Studocu

4. Data Explosion: Volume, Variety, Velocity, and Veracity

- **Volume**: The amount of data generated every second is enormous.
- **Variety**: Data comes in various formats—structured, semi-structured, and unstructured.
- **Velocity**: The speed at which new data is generated and needs to be processed.
- **Veracity**: The uncertainty of data, which can lead to inconsistencies and inaccuracies.

5. Big Data Infrastructure and Challenges

- **Infrastructure**: Involves distributed storage systems, parallel processing, and cloud computing.
- **Challenges**: Include data security, privacy concerns, and the need for scalable storage solutions.

6. Big Data Processing Architectures

- **Data Warehouse**: Central repository for storing large volumes of structured data.
- **Re-Engineering the Data Warehouse**: Modifying existing data warehouses to handle big data.
- **Shared Everything Architecture**: All nodes share the same storage and memory.
- **Shared Nothing Architecture**: Each node is independent and self-sufficient, which enhances scalability.Studocu

7. Big Data Learning Approaches

- **Supervised Learning**: Models are trained on labeled data.

- **Unsupervised Learning**: Models identify patterns in unlabeled data.
- **Reinforcement Learning**: Models learn by interacting with the environment and receiving feedback.AIT Pune+2dypatiltcs.com+2AIT Pune+2

8. Data Science – The Big Picture

- **Artificial Intelligence (AI)**: The simulation of human intelligence processes by machines.
- **Statistical Learning**: Focuses on understanding data and making predictions.
- **Machine Learning (ML)**: A subset of AI that enables systems to learn from data.
- **Data Mining**: The process of discovering patterns in large datasets.
- **Big Data Analytics**: The complex process of examining big data to uncover information.AIT Pune+3dypatiltcs.com+3kjei.edu.in+3

Unit II: Mathematical Foundation of Big Data (7 Hrs)

1. Probability

- **Random Variables**: Variables that take on different values based on the outcome of a random phenomenon.
- **Joint Probability**: The probability of two events occurring simultaneously.
- **Conditional Probability**: The probability of an event occurring given that another event has occurred.
- **Markov Chains**: Models that predict a system's future state based solely on its current state.

2. Tail Bounds

- Inequalities that provide bounds on the probability that a random variable deviates from some value (e.g., Markov's inequality, Chebyshev's inequality).

3. Markov Chains and Random Walks

- **Markov Chains**: Used to model systems that transition from one state to another on a state space.
- **Random Walks**: A mathematical formalization of a path consisting of a succession of random steps.

4. Pair-wise Independence and Universal Hashing

- **Pair-wise Independence**: Two events are pair-wise independent if the occurrence of one does not affect the probability of the other.
- **Universal Hashing**: A method to select a hash function at random from a family of hash functions with certain mathematical properties.

5. Approximate Counting and Median

- **Approximate Counting**: Algorithms that estimate the number of distinct elements in a dataset.

- **Approximate Median**: Algorithms that estimate the median value in a dataset without sorting the entire dataset.

6. Data Streaming Models and Statistical Methods

- **Flajolet-Martin Algorithm**: Estimates the number of distinct elements in a stream.
- **Distance Sampling and Random Projections**: Techniques for dimensionality reduction and similarity estimation.
- **Bloom Filters**: Space-efficient probabilistic data structures used to test whether an element is a member of a set.

7. Statistical Measures

- **Mode**: The value that appears most frequently in a dataset.
- **Variance**: Measures how far a set of numbers are spread out from their average value.
- **Standard Deviation**: The square root of the variance, representing data dispersion.
- **Correlation Analysis**: Determines the relationship between two variables.
- **Analysis of Variance (ANOVA)**: Analyzes the differences among group means in a sample.

Unit III: Big Data Processing (6 Hrs)

1. Big Data Analytics Ecosystem and Technologies

- **Ecosystem**: Includes tools like Hadoop, Spark, Hive, Pig, and NoSQL databases.
- **Technologies**: Focus on distributed computing and storage solutions.

2. Google File System (GFS)

- A scalable distributed file system developed by Google for large distributed data-intensive applications.

3. Hadoop Architecture

- **HDFS (Hadoop Distributed File System)**: Stores large files across multiple machines.
- **MapReduce**: A programming model for processing large data sets with a distributed algorithm.

4. Hadoop Storage: HDFS

- **NameNode**: Manages the metadata and namespace for HDFS.
- **DataNode**: Stores the actual data.
- **Secondary NameNode**: Performs housekeeping functions for the NameNode.

5. Common Hadoop Shell Commands

- Commands like `hdfs dfs -ls`, `hdfs dfs -put`, `hdfs dfs -get` are used to interact with HDFS.

6. Anatomy of File Write and Read

- **Write**: Client requests to write a file, which is divided into blocks and stored in DataNodes.
- **Read**: Client requests to read a file, and the NameNode provides the locations of the blocks.

7. MapReduce Paradigm

- **Map Function**: Processes input data and produces intermediate key-value pairs.
- **Reduce Function**: Aggregates the intermediate data to produce the final output.

8. Job and Task Trackers

- **Job Tracker**: Manages the distribution of tasks to specific nodes.
- **Task Tracker**: Runs on DataNodes and executes the tasks assigned by the Job Tracker.

9. Cluster Setup – SSH & Hadoop Configuration

- **SSH**: Used for secure communication between nodes.
- **Configuration**: Involves setting up configuration files like `core-site.xml`, `hdfs-site.xml`, and `mapred-site.xml`.

10. Introduction to NoSQL

- Non-relational databases designed for large-scale data storage and for massively-parallel, high-performance data processing across a large number of commodity servers.

11. Textual ETL Processing

- **ETL (Extract, Transform, Load)**: Processes that involve extracting data from various sources, transforming it into a suitable format, and loading it into a database or data warehouse.

Unit IV: Big Data Analytics (6 Hrs)

1. Big Data Analytics - Architecture and Life Cycle

- **Architecture**: Involves data sources, data storage, data processing, and data visualization layers. It ensures efficient handling and analysis of big data.
- **Life Cycle**: Comprises phases like data discovery, data preparation, model planning, model building, and deployment. This iterative process ensures continuous improvement and adaptation to new data. GeeksforGeeks

2. Types of Analysis

- **Descriptive Analysis**: Summarizes historical data to understand changes over time.
- **Diagnostic Analysis**: Examines data to understand causes of events and behaviors.
- **Predictive Analysis**: Uses statistical models to forecast future outcomes.

- **Prescriptive Analysis**: Suggests actions based on predictive insights to achieve desired outcomes. WikipediaIBM - United States

3. Analytical Approaches

- **Mathematical Manipulations**: Involves statistical computations and algorithms to derive insights.
- **Data Mining**: Extracts patterns and knowledge from large datasets.
- **Machine Learning**: Employs algorithms that improve automatically through experience.

4. Data Ingestion from Different Sources

- **CSV, JSON, HTML, Excel**: Common file formats for structured and semi-structured data.
- **MongoDB, MySQL, SQLite**: Databases used for storing and retrieving large volumes of data efficiently.advantagecourseware.s3.amazonaws.com

5. Data Cleaning

- **Handling Missing Values**: Techniques like imputation or deletion to manage incomplete data.
- **Data Imputation**: Replacing missing data with substituted values based on various strategies.

6. Data Transformation and Standardization

- **Transformation**: Converting data into a suitable format for analysis.
- **Standardization**: Scaling data to have a mean of zero and a standard deviation of one.

7. Handling Categorical Data

- **Two Categories**: Binary encoding (e.g., 0 and 1).
- **Multiple Categories**: One-hot encoding or label encoding to convert categories into numerical values.

8. Statistical and Graphical Analysis Methods

- **Statistical Methods**: Include measures like mean, median, mode, variance, and standard deviation.
- **Graphical Methods**: Utilize charts like histograms, box plots, and scatter plots to visualize data distributions and relationships.

9. Hive Data Analytics

- **Apache Hive**: A data warehouse software that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.

Unit V: Big Data Visualization (6 Hrs)

1. Introduction to Data Visualization

- The graphical representation of information and data to facilitate understanding of complex datasets.

2. Challenges to Big Data Visualization

- **Volume**: Handling massive datasets.
- **Variety**: Dealing with diverse data types and formats.
- **Velocity**: Managing the speed at which new data is generated and needs to be visualized.

3. Conventional Data Visualization Tools

- Tools like tables, histograms, scatter plots, line charts, bar charts, pie charts, area charts, flow charts, bubble charts, Venn diagrams, data flow diagrams, and entity relationship diagrams are commonly used for visualizing data. Prof. Bhavana Khivsara

4. Techniques for Visual Data Representations

- **Parallel Coordinates**: Used for plotting multivariate numerical data.
- **Treemaps**: Display hierarchical data as a set of nested rectangles.
- **Cone Trees and Semantic Networks**: Visualize hierarchical information and relationships between concepts.GeeksforGeeks+9Prof. Bhavana Khivsara+9PICT+9

5. Types of Data Visualization

- **Time-Series**: Line charts showing data points at successive time intervals.
- **Statistical**: Box plots, histograms, and scatter plots to show data distributions.
- **Hierarchical**: Tree diagrams and dendrograms to represent data with parent-child relationships.
- **Network**: Node-link diagrams to show relationships between entities.

6. Visualizing Big Data

- Involves using advanced tools and techniques to handle and represent large and complex datasets effectively.

7. Tools Used in Data Visualization

- **Proprietary Tools**: Tableau, Power BI.
- **Open-Source Tools**: D3.js, Google Chart API, Candela.sampig.github.io+4IJCert+4Wikipedia+4

8. Case Study: Analysis of a Business Problem of Zomato Using Visualization

- Utilizing data visualization techniques to analyze Zomato's business data, such as customer preferences, peak ordering times, and popular cuisines, to derive actionable insights.

9. Analytical Techniques Used in Big Data Visualization

- **Interactive Dashboards**: Allow users to interact with data visualizations for deeper insights.
- **Geospatial Mapping**: Visualizing data on maps to identify regional patterns.
- **Real-Time Visualization**: Displaying data as it is generated for immediate analysis.ResearchGate+3Wikipedia+3Prof. Bhavana Khivsara+3

10. Data Visualization Using Tableau

- **Tableau**: A powerful data visualization tool that allows for the creation of interactive and shareable dashboards. It supports various data sources and provides real-time data analytics.

Unit VI: Big Data Technologies Application and Impact (05 Hrs)

1. Social Media Analytics

Social media analytics involves collecting and analyzing data from social platforms like Facebook, Twitter, and Instagram to inform business decisions. This includes understanding user behavior, sentiment analysis, and measuring campaign effectiveness. By leveraging big data, companies can tailor their marketing strategies to target specific audiences more effectively.

2. Text Mining

Text mining, or text data mining, is the process of transforming unstructured text into structured data to identify meaningful patterns and insights. It enables organizations to analyze vast collections of textual materials, such as customer reviews or social media posts, to capture key concepts, trends, and hidden relationships. GWU Library Guides+2IBM - United States+2GeeksforGeeks+2

3. Mobile Analytics

Mobile analytics refers to the collection and analysis of data about user behavior on mobile devices and applications. It provides insights into app performance, user engagement, and helps identify areas for improvement to enhance user experience and retention. HeapGlassbox

4. Data Analytics Life Cycle of Case Studies

The data analytics life cycle encompasses stages from data collection to actionable insights. Case studies, such as those from companies like Netflix or Uber, demonstrate how data is utilized to optimize operations, personalize user experiences, and drive innovation.

5. Organizational Impact

Implementing big data analytics can significantly impact organizations by enhancing decision-making, improving operational efficiency, and fostering innovation. It enables businesses to respond swiftly to market changes and customer needs.

6. Understanding Decision Theory

Decision theory in the context of big data involves using data-driven approaches to make informed decisions. It combines statistical analysis and predictive modeling to evaluate potential outcomes and optimize decision-making processes. IBM - United States

7. Creating Big Data Strategy

Developing a big data strategy involves defining business objectives, identifying relevant data sources, and establishing processes for data management and analysis. A well-crafted strategy ensures that data initiatives align with organizational goals and deliver measurable value. Informa TechTarget

8. Big Data Value Creation Drivers

Value creation from big data arises from capabilities such as data democratization, contextualization, experimentation, and execution. These drivers enable organizations to harness data effectively, leading to improved decision-making and competitive advantage. IEDP

9. Michael Porter's Value Creation Models

Michael Porter's Value Chain model analyzes a company's activities to identify areas where value can be added. Integrating big data into this model allows organizations to optimize operations, enhance customer experiences, and create new revenue streams.

10. Big Data User Experience Ramifications

Big data analytics can significantly influence user experience by enabling personalized content, predictive recommendations, and responsive interfaces. However, it also raises concerns about privacy and data security, necessitating ethical considerations in data usage.

11. Identifying Big Data Use Cases

Big data use cases span various industries, including healthcare, finance, retail, and transportation. Examples include predictive maintenance in manufacturing, fraud detection in banking, and personalized marketing in retail.

12. Big Data Analytics Challenges and Research Directions

Challenges in big data analytics include data privacy concerns, integration of diverse data sources, and the need for skilled professionals. Future research directions focus on developing advanced analytics tools, real-time data processing, and addressing ethical implications of data usage.