Professional Certificate in Big Data and Cloud Computing

Foundations of Big Data

Foundations of Big Data Glossary

A

1. Aggregation: A process where raw data is combined to form a summary for analysis. Aggregation helps in reducing data size and complexity for faster processing. For example, calculating the average sales per month from daily sales data.

2. Apache Hadoop: An open-source software framework that is used for distributed storage and processing of large data sets using a cluster of computers. Hadoop consists of the Hadoop Distributed File System (HDFS) and MapReduce.

3. Artificial Intelligence (AI): The simulation of human intelligence processes by machines, especially computer systems. AI includes tasks such as learning, reasoning, problem-solving, perception, and language understanding.

B

1. Big Data: Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. Big data is characterized by the 3Vs: Volume, Velocity, and Variety.

2. Business Intelligence (BI): The use of data analysis tools and techniques to extract actionable insights from raw data to support decision-making in organizations. BI helps in understanding trends, identifying opportunities, and making informed decisions.

C

1. Cloud Computing: The delivery of computing services—including servers, storage, databases, networking, software, analytics, and intelligence—over the Internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale.

2. Cluster: A group of interconnected computers that work together as a single system to process large volumes of data. Clusters are commonly used in big data environments for distributed computing and storage.

3. CSV (Comma-Separated Values): A file format used to store tabular data in plain text, where each line corresponds to a row in the table, and columns are separated by commas. CSV files are commonly used for data exchange between applications.

D

1. Data Mining: The process of discovering patterns, trends, and insights from large data sets using statistical techniques, machine learning, and artificial intelligence. Data mining helps in uncovering valuable information hidden in data.

2. Data Quality: The measure of the accuracy, completeness, consistency, and reliability of data. High data quality is essential for making informed decisions and ensuring the effectiveness of data analysis.

3. Data Visualization: The representation of data in graphical or visual formats to make complex data sets more understandable and accessible. Data visualization helps in identifying trends, patterns, and outliers in data.

E

1. ETL (Extract, Transform, Load): A process used to collect data from various sources, transform it into a consistent format, and load it into a data warehouse or database for analysis. ETL is essential for data integration and data migration.

2. Exabyte: A unit of information equal to one quintillion bytes, or 2^60 bytes. Exabytes are used to measure extremely large data sets, especially in the context of big data analytics.

F

1. Feature Engineering: The process of selecting, creating, and transforming features (variables) in data sets to improve the performance of machine learning models. Feature engineering plays a crucial role in building accurate predictive models.

2. Framework: A software platform or structure that provides a foundation for developing applications, systems, or solutions. Frameworks help in simplifying development, enforcing best practices, and promoting reusability.

3. Frequency Distribution: A summary of the frequency of occurrences of values in a data set. Frequency distributions are used in descriptive statistics to understand the distribution of data and identify patterns or outliers.

G

1. GPU (Graphics Processing Unit): A specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. GPUs are commonly used in parallel processing for big data analytics.

2. Graph Database: A database that uses graph structures for semantic queries with nodes, edges, and properties to represent and store data. Graph databases are suitable for analyzing and querying interconnected data sets.

H

1. Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data and is designed to handle large data sets across multiple machines. HDFS is the primary storage system used by Apache Hadoop.

2. High-Performance Computing (HPC): The use of supercomputers and parallel processing techniques to solve complex computational problems efficiently. HPC is essential for processing large-scale simulations and big data analytics.

I

1. Indexing: The process of creating data structures to improve the speed of data retrieval and query performance. Indexing is used in databases and search engines to quickly locate specific data based on predefined criteria.

2. Internet of Things (IoT): The network of physical devices, vehicles, home appliances, and other items embedded with sensors, software, and connectivity to exchange data over the Internet. IoT generates vast amounts of data for analysis.

J

1. JSON (JavaScript Object Notation): A lightweight data-interchange format that is easy for humans to read and write and easy for machines to parse and generate. JSON is commonly used for transmitting data between a server and web application.

K

1. K-means Clustering: A popular clustering algorithm used in machine learning to partition data into K clusters based on similarity. K-means clustering is an unsupervised learning technique that helps in grouping similar data points.

2. K-nearest Neighbors (KNN): A classification algorithm that assigns a class label to an input data point based on the majority class of its K-nearest neighbors. KNN is a simple yet effective algorithm for pattern recognition and classification.

L

1. Logistic Regression: A statistical model used for binary classification tasks where the output variable takes only two possible values (e.g., 0 or 1). Logistic regression estimates the probability of a binary outcome based on input features.

2. Machine Learning: A branch of artificial intelligence that enables systems to learn from data, identify patterns, and make decisions without being explicitly programmed. Machine learning algorithms improve their performance over time.

M

1. MapReduce: A programming model used for processing and generating large data sets with a parallel, distributed algorithm on a cluster. MapReduce consists of two main phases: Map (data processing) and Reduce (aggregation).

2. Metadata: Data that describes other data, providing information about the content, structure, format, and context of data sets. Metadata helps in organizing, managing, and understanding data assets.

N

1. Normalization: The process of organizing data in a database efficiently to reduce redundancy and dependency. Normalization helps in eliminating data anomalies and improving data integrity in relational databases.

2. NoSQL (Not Only SQL): A term used to describe non-relational databases that provide flexible data models for handling unstructured and semi-structured data. NoSQL databases are designed for scalability, high availability, and performance.

O

1. Overfitting: A machine learning phenomenon where a model learns the noise or random fluctuations in the training data rather than the underlying patterns. Overfitting leads to poor generalization and decreased performance on unseen data.

2. Open Source: Software that is freely available for use, modification, and distribution by anyone. Open-source software promotes collaboration, transparency, and innovation in the development community.

P

1. Predictive Analytics: The use of statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. Predictive analytics helps in making informed decisions and anticipating trends.

2. Privacy: The protection of personal and sensitive information from unauthorized access, use, or disclosure. Privacy is a critical consideration in big data analytics to ensure legal compliance and maintain trust with users.

Q

1. Query: A request for information from a database or data set using a specific set of criteria or conditions. Queries are used to retrieve, filter, and manipulate data to extract meaningful insights.

2. Quantitative Data: Numerical data that can be measured and expressed in numerical form. Quantitative data is used in statistical analysis to perform calculations, comparisons, and predictions.

R

1. Regression Analysis: A statistical technique used to model the relationship between a dependent variable

and one or more independent variables. Regression analysis helps in predicting continuous outcomes based on input features.

2. Relational Database: A type of database that stores and organizes data in tables with predefined relationships between them. Relational databases use structured query language (SQL) for data manipulation and retrieval.

S

1. Scalability: The ability of a system to handle a growing amount of work or its potential to accommodate growth. Scalability is a crucial factor in big data environments to ensure performance, availability, and efficiency.

2. Semi-Structured Data: Data that does not fit into a structured format like a relational database but includes tags, markers, or keys to separate elements. Semi-structured data is common in web pages, emails, and JSON files.

T

1. Text Mining: The process of extracting meaningful information and patterns from unstructured text data using natural language processing (NLP) and machine learning techniques. Text mining helps in analyzing large volumes of text for insights.

2. Time Series Analysis: A statistical technique used to analyze and forecast time-dependent data points collected at regular intervals. Time series analysis helps in understanding patterns, trends, and seasonality in temporal data.

U

1. Unstructured Data: Data that does not have a predefined format or organization and cannot be easily stored in a traditional database. Unstructured data includes text, images, videos, and social media content.

2. Usability: The ease of use and effectiveness of a system, application, or product for end-users to achieve specific goals efficiently. Usability is essential in big data solutions to ensure user adoption and satisfaction.

V

1. Validation: The process of checking and confirming the accuracy, completeness, and reliability of data to ensure its quality and integrity. Validation is crucial in data analysis to prevent errors and ensure the credibility of results.

2. Visualization: The graphical representation of data and information to communicate insights, trends, and patterns effectively. Data visualization tools help in interpreting complex data and making data-driven decisions.

W

1. Web Scraping: The process of extracting data from websites by using automated tools or bots to collect information for analysis. Web scraping is commonly used in data mining, market research, and competitive analysis.

2. Workflow: The sequence of tasks, processes, and operations involved in the management and analysis of data. Workflows help in organizing and automating data processing tasks to improve efficiency and productivity.

X

1. X-axis: The horizontal line on a graph that represents the independent variable or data points in a data set. The x-axis is used to plot and visualize quantitative data along a common scale.

2. XLSX (Excel Spreadsheet): A file format used to store spreadsheet data in Microsoft Excel, containing rows and columns of cells with text, numbers, and formulas. XLSX files are commonly used for data analysis and reporting.

Y

1. Y-axis: The vertical line on a graph that represents the dependent variable or output values in a data set. The y-axis is used to plot and visualize the response or outcome based on the independent variable.

2. Yottabyte: A unit of information equal to one septillion bytes, or 2^80 bytes. Yottabytes are used to measure extremely large data sets, especially in the context of big data analytics.

Z

1. Zero-Day Attack: A cyber attack that exploits a software vulnerability on the same day it becomes known to the public, before a fix or patch is available. Zero-day attacks pose a significant threat to data security and privacy.

2. Zettabyte: A unit of information equal to one sextillion bytes, or 2^70 bytes. Zettabytes are used to measure large data volumes, especially in the context of global data storage and transmission.

This glossary provides a comprehensive overview of key terms and concepts related to the foundations of big data in the context of the Professional Certificate in Big Data and Cloud Computing. Understanding these terms is essential for professionals working in data analytics, machine learning, and cloud computing to effectively leverage big data for decision-making and innovation.