Postgraduate Certificate in Data-Driven Science Journalism

Big Data Technologies

Big Data Technologies:

Big Data: Big data refers to large and complex data sets that are difficult to process using traditional data processing applications. These data sets are characterized by their volume, velocity, and variety, often exceeding the capabilities of conventional databases and software tools.

Data-driven: Data-driven refers to the practice of making decisions and guiding actions based on data analysis and interpretation rather than intuition or personal experience. Data-driven approaches rely on collecting and analyzing data to gain insights and inform decision-making processes.

Science Journalism: Science journalism is a specialized form of journalism that focuses on reporting scientific research, discoveries, and developments to the public. Science journalists often translate complex scientific concepts into accessible and engaging stories for a general audience.

Postgraduate Certificate: A postgraduate certificate is a qualification awarded upon completion of a specialized program of study at the postgraduate level. It is typically shorter in duration than a master's degree and provides focused training in a specific area of study.

Data Science: Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Data science combines elements of statistics, machine learning, data analysis, and programming to uncover patterns and trends in data.

Journalism: Journalism is the practice of gathering, assessing, creating, and presenting news and information to an audience. Journalists investigate and report on current events, trends, and issues to inform the public and promote transparency and accountability.

Technology: Technology refers to the tools, systems, and methods used to solve practical problems or achieve specific goals. In the context of big data technologies, technology encompasses software, hardware, networks, and algorithms designed to process, analyze, and visualize large volumes of data.

Data Analysis: Data analysis is the process of inspecting, cleansing, transforming, and modeling data to uncover useful information, inform conclusions, and support decision-making. Data analysis techniques include statistical analysis, data mining, machine learning, and visualization.

Data Processing: Data processing involves converting raw data into a usable format through various operations such as sorting, filtering, aggregating, and summarizing. Data processing is a key step in preparing data for analysis and interpretation.

Data Visualization: Data visualization is the graphical representation of data to communicate information

clearly and efficiently. Data visualizations can take the form of charts, graphs, maps, and dashboards, helping users understand complex data sets and identify patterns.

Machine Learning: Machine learning is a subset of artificial intelligence that enables systems to learn from data and improve performance on specific tasks without being explicitly programmed. Machine learning algorithms can recognize patterns, make predictions, and optimize outcomes based on training data.

Artificial Intelligence: Artificial intelligence (AI) refers to the simulation of human intelligence processes by machines, particularly computer systems. AI technologies enable machines to perform tasks that typically require human intelligence, such as speech recognition, decision-making, and problem-solving.

Algorithm: An algorithm is a set of instructions or rules designed to solve a specific problem or perform a particular task. In the context of big data technologies, algorithms are used to process and analyze large data sets efficiently and effectively.

Database: A database is an organized collection of data stored and accessed electronically. Databases are designed to manage, manipulate, and retrieve data for various applications, ranging from simple spreadsheets to complex relational databases.

Data Management: Data management involves the processes, policies, and practices used to acquire, store, analyze, and protect data throughout its lifecycle. Effective data management ensures data quality, integrity, security, and accessibility for users.

Data Quality: Data quality refers to the accuracy, completeness, consistency, and reliability of data for a specific purpose or application. High data quality is essential for making informed decisions, conducting analysis, and producing reliable results.

Data Integration: Data integration is the process of combining data from different sources or formats to create a unified view of information. Data integration tools and techniques help organizations consolidate and harmonize data for analysis and reporting.

Data Warehousing: Data warehousing is the process of storing and managing large volumes of structured data from multiple sources in a centralized repository. Data warehouses enable organizations to analyze historical data, extract insights, and support decision-making processes.

Cloud Computing: Cloud computing is a model for delivering computing services over the internet on a pay-as-you-go basis. Cloud computing provides on-demand access to a shared pool of resources, including servers, storage, and applications, without the need for on-premises infrastructure.

Hadoop: Hadoop is an open-source framework for distributed storage and processing of big data sets across clusters of commodity hardware. Hadoop includes the Hadoop Distributed File System (HDFS) for storage and MapReduce for processing large-scale data sets.

Spark: Spark is an open-source cluster computing framework that provides in-memory processing capabilities for big data analytics. Spark offers faster data processing speeds than Hadoop's MapReduce through its resilient distributed dataset (RDD) abstraction.

NoSQL: NoSQL, or "not only SQL," is a category of database systems that provide flexible and scalable storage solutions for unstructured and semi-structured data. NoSQL databases are designed to handle large volumes of data and support distributed computing environments.

Real-time Analytics: Real-time analytics is the process of analyzing data as it is generated or received to provide immediate insights and actionable information. Real-time analytics enables organizations to make timely decisions, detect anomalies, and respond to events in real time.

Data Mining: Data mining is the process of discovering patterns, trends, and insights from large data sets using statistical and machine learning techniques. Data mining helps uncover hidden relationships and valuable information for decision-making and prediction.

Predictive Analytics: Predictive analytics is the practice of using historical data, statistical algorithms, and machine learning techniques to forecast future outcomes and trends. Predictive analytics helps organizations anticipate risks, opportunities, and customer behavior.

Internet of Things (IoT): The Internet of Things (IoT) refers to a network of interconnected devices, sensors, and objects that exchange data and communicate with each other over the internet. IoT technologies enable real-time monitoring, automation, and data collection in various industries.

Data Privacy: Data privacy encompasses the principles and practices related to protecting individuals' personal information and ensuring the confidentiality and security of sensitive data. Data privacy regulations govern how organizations collect, store, and use personal data to safeguard individuals' privacy rights.

Data Security: Data security involves protecting data from unauthorized access, disclosure, alteration, or destruction. Data security measures include encryption, access controls, authentication, and auditing to ensure the confidentiality and integrity of data.

Data Governance: Data governance is the framework of policies, procedures, and processes that define how organizations manage and control their data assets. Data governance ensures data quality, compliance, and accountability across the organization.

Data Ethics: Data ethics refers to the moral principles and guidelines that govern the responsible use of data, including privacy, transparency, fairness, and accountability. Data ethics considerations are essential in data-driven decision-making to uphold ethical standards and avoid bias or discrimination.

Data Literacy: Data literacy is the ability to read, analyze, interpret, and communicate data effectively. Data-literate individuals can understand data visualizations, evaluate data sources, and draw meaningful insights from data to inform decision-making.

Data Journalism: Data journalism is a form of journalism that uses data analysis, visualization, and storytelling techniques to report news and inform the public. Data journalists combine investigative reporting with data-driven approaches to uncover trends, patterns, and insights in data.

Python: Python is a popular programming language widely used in data science, machine learning, and big

data analytics. Python offers a rich set of libraries and tools for data manipulation, visualization, and analysis, making it a versatile language for data-driven projects.

R: R is a programming language and environment specifically designed for statistical computing and data analysis. R provides a wide range of packages and functions for data manipulation, visualization, and modeling, making it a powerful tool for statistical programming and research.

SQL: SQL, or Structured Query Language, is a standard programming language used to manage and manipulate relational databases. SQL enables users to retrieve, update, and manipulate data stored in database tables through a set of declarative commands and queries.

API: An API, or Application Programming Interface, is a set of rules and protocols that allows different software applications to communicate with each other. APIs enable data sharing, integration, and automation between systems, services, and platforms.

ETL: ETL stands for Extract, Transform, Load, a process used to extract data from source systems, transform it into a consistent format, and load it into a target database or data warehouse. ETL tools automate data integration tasks and ensure data quality and consistency.

Dashboard: A dashboard is a visual display of key performance indicators, metrics, and data insights designed to provide an overview of an organization's performance. Dashboards help users monitor trends, track progress, and make data-driven decisions based on real-time information.

Data Storytelling: Data storytelling is the practice of using data analysis, visualization, and narrative techniques to communicate insights and findings to a general audience. Data storytellers combine data-driven evidence with compelling narratives to engage and inform readers.

Natural Language Processing (NLP): Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. NLP technologies power chatbots, language translation, sentiment analysis, and text mining applications.

Deep Learning: Deep learning is a subset of machine learning that uses artificial neural networks to model and interpret complex patterns in data. Deep learning algorithms, such as deep neural networks, can learn from large data sets to perform tasks like image recognition and speech synthesis.

Blockchain: Blockchain is a decentralized and distributed ledger technology that securely records and verifies transactions across a network of computers. Blockchain technology ensures data integrity, transparency, and immutability through cryptographic principles and consensus mechanisms.

Quantum Computing: Quantum computing is a cutting-edge computing technology that leverages the principles of quantum mechanics to perform complex calculations and solve problems at an exponential speed. Quantum computers use quantum bits, or qubits, to process and store information in quantum states.

Cybersecurity: Cybersecurity involves protecting computer systems, networks, and data from cyber threats, attacks, and unauthorized access. Cybersecurity measures include encryption, firewall protection, intrusion

detection, and security protocols to safeguard digital assets.

Internet Security: Internet security encompasses the practices and technologies used to protect data and information transmitted over the internet from cyber threats and vulnerabilities. Internet security measures include secure protocols, encryption, and authentication mechanisms to ensure data confidentiality and integrity.

Data Storage: Data storage refers to the process of storing, organizing, and managing data in digital form for future use. Data storage technologies include hard drives, solid-state drives, cloud storage, and tape storage solutions to preserve and access data efficiently.

Data Compression: Data compression is the process of reducing the size of data files to save storage space, decrease transmission times, and optimize data transfer. Data compression algorithms encode data in a more efficient format without losing essential information or quality.

Parallel Computing: Parallel computing is a computing technique that divides and distributes tasks across multiple processors or computing cores to increase processing speed and efficiency. Parallel computing enables large-scale data processing, simulation, and scientific computations to be performed in parallel.

Scalability: Scalability refers to the ability of a system, network, or application to handle increased workloads, data volumes, and users without compromising performance or reliability. Scalable technologies can adapt and grow to meet changing demands and support expanding data requirements.

High Availability: High availability is a measure of a system's reliability and uptime, ensuring that services and applications are continuously accessible and operational. High availability technologies use redundancy, failover mechanisms, and load balancing to minimize downtime and ensure continuous service delivery.

Data Migration: Data migration is the process of transferring data from one system, platform, or storage location to another while preserving data integrity and consistency. Data migration tools help organizations move data seamlessly between databases, applications, and cloud environments.

Containerization: Containerization is a method of packaging and deploying applications in lightweight, portable containers that encapsulate software, dependencies, and configuration settings. Containerization technologies like Docker enable developers to build, deploy, and manage applications consistently across different environments.

DevOps: DevOps is a software development methodology that combines software development (Dev) and IT operations (Ops) to automate and streamline the software delivery process. DevOps practices focus on collaboration, continuous integration, and continuous deployment to accelerate development cycles and improve software quality.

Agile Methodology: Agile methodology is an iterative and incremental approach to software development that emphasizes flexibility, collaboration, and customer feedback. Agile teams work in short development cycles called sprints to deliver working software and adapt to changing requirements quickly.

Continuous Integration: Continuous Integration (CI) is a software development practice that involves

automatically integrating code changes into a shared repository multiple times a day. CI tools automate the build, test, and deployment processes to ensure code quality and detect errors early in the development cycle.

Machine-to-Machine (M2M) Communication: Machine-to-Machine (M2M) communication refers to the exchange of data and information between interconnected devices, sensors, and machines without human intervention. M2M technologies enable automated data transfer, monitoring, and control in various industries.

Edge Computing: Edge computing is a distributed computing paradigm that processes data closer to the source or "edge" of the network, rather than in centralized data centers. Edge computing reduces latency, bandwidth usage, and processing time for real-time data analysis and decision-making.

Robotic Process Automation (RPA): Robotic Process Automation (RPA) uses software robots or "bots" to automate repetitive tasks, processes, and workflows in business operations. RPA technologies mimic human actions to perform rule-based activities, data entry, and data manipulation with high accuracy and efficiency.

Dark Data: Dark data refers to unstructured or unutilized data that organizations collect but do not analyze or leverage for insights or decision-making. Dark data includes text files, images, videos, and other data types that remain untapped for valuable information.

Data Monetization: Data monetization is the process of generating revenue from data assets through the sale, licensing, or exchange of data products and services. Data monetization strategies include data sharing, data analytics, and data-driven marketing to create value from data assets.

Data Anonymization: Data anonymization is the process of removing personally identifiable information (PII) from data sets to protect individuals' privacy and comply with data protection regulations. Data anonymization techniques include masking, tokenization, and generalization to de-identify sensitive data.

Data Labeling: Data labeling is the process of annotating or tagging data samples to create labeled datasets for machine learning and AI model training. Data labeling tasks involve classifying images, transcribing text, and identifying patterns to teach algorithms to recognize and predict patterns in data.

Data Fusion: Data fusion involves combining information from multiple sources or sensors to create a comprehensive and accurate representation of a phenomenon or event. Data fusion techniques integrate data streams, sensor readings, and observations to improve situational awareness and decision-making.

Geospatial Data: Geospatial data refers to information that is tied to a specific geographic location or spatial coordinates. Geospatial data includes maps, satellite imagery, GPS coordinates, and location-based data used in mapping, navigation, and spatial analysis applications.

Open Data: Open data refers to publicly available data that can be freely accessed, used, and shared by anyone for any purpose. Open data initiatives promote transparency, collaboration, and innovation by making government, scientific, and organizational data accessible to the public.

Data Catalog: A data catalog is a centralized repository or database that organizes and indexes metadata about data assets, datasets, and data sources within an organization. Data catalogs help users discover, understand, and access data for analysis, reporting, and decision-making.

Data Lake: A data lake is a centralized repository that stores structured, unstructured, and semi-structured data at scale for big data analytics and processing. Data lakes enable organizations to store raw data in its native format and perform diverse analytics and data science tasks.

Data Warehouse: A data warehouse is a centralized repository that stores structured, historical data from multiple sources for reporting, analysis, and business intelligence purposes. Data warehouses organize and integrate data to provide a unified view of information for decision-making.

Metadata: Metadata is data that describes other data, providing information about data attributes, characteristics, and relationships. Metadata helps users understand and manage data assets, including data lineage, data quality, and data governance information.

Data Schema: A data schema is a blueprint or design that defines the structure, organization, and relationships of data elements in a database or data model. Data schemas specify data types, constraints, and rules for organizing and storing data in a consistent format.

Data Mining: Data mining is the process of discovering patterns, trends, and insights from large data sets using statistical and machine learning techniques. Data mining helps uncover hidden relationships and valuable information for decision-making and prediction.

Data Warehouse: A data warehouse is a centralized repository that stores structured, historical data from multiple sources for reporting, analysis, and business intelligence purposes. Data warehouses organize and integrate data to provide a unified view of information for decision-making.

Metadata: Metadata is data that describes other data, providing information about data attributes, characteristics, and relationships. Metadata helps users understand and manage data assets, including data lineage, data quality, and data governance information.

Data Schema: A data schema is a blueprint or design that defines the structure, organization, and relationships of data elements in a database or data model. Data schemas specify data types, constraints, and rules for organizing and storing data in a consistent format.