What is Data Science

Big Data and Data Mining

Term	Definition
Analytics	The process of examining data to draw conclusions and make informed decisions is a fundamental aspect of data science, involving statistical analysis and data-driven insights.
Big Data	Vast amounts of structured, semi-structured, and unstructured data are characterized by its volume, velocity, variety, and value, which, when analyzed, can provide competitive advantages and drive digital transformations.
Big Data Cluster	A distributed computing environment comprising thousands or tens of thousands of interconnected computers that collectively store and process large datasets.
Broad Network Access	The ability to access cloud resources via standard mechanisms and platforms such as mobile devices, laptops, and workstations over networks.
Cloud Computing	The delivery of on-demand computing resources, including networks, servers, storage, applications, services, and data centers, over the Internet on a pay-for-use basis.
Cloud Deployment Models	Categories that indicate where cloud infrastructure resides, who manages it, and how cloud resources and services are made available

	to users, including public, private, and hybrid models.
Cloud Service Models	Models based on the layers of a computing stack, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), represent different cloud computing offerings.
Commodity Hardware	Standard, off-the-shelf hardware components are used in a big data cluster, offering cost-effective solutions for storage and processing without relying on specialized hardware.
Data Algorithms	Computational procedures and mathematical models used to process and analyze data made accessible in the cloud for data scientists to deploy on large datasets efficiently.
Data Replication	A strategy in which data is duplicated across multiple nodes in a cluster to ensure data durability and availability, reducing the risk of data loss due to hardware failures.
Data Science	An interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including programming, statistics, and analytical tools.
Deep Learning	A subset of machine learning that involves artificial neural networks inspired by the human brain, capable of learning and making complex decisions from data on their own.
Digital Change	The integration of digital technology into business processes and operations leads to improvements and innovations in how organizations operate and deliver value to customers.
Digital Transformation	A strategic and cultural organizational change driven by data science, especially Big Data, to integrate digital technology across all areas of the organization, resulting in fundamental operational and value delivery changes.
Distributed Data	The practice of dividing data into smaller chunks and distributing them across multiple computers within a cluster enables parallel processing for data analysis.
Hadoop	A distributed storage and processing framework used for handling and analyzing large datasets, particularly well-suited for big data analytics and data science applications.
Hadoop Distributed File System (HDFS)	A storage system within the Hadoop framework that partitions and distributes files across multiple nodes, facilitating parallel data access and fault tolerance.

Infrastructure as a Service (IaaS)	A cloud service model that provides access to computing infrastructure, including servers, storage, and networking, without the need for users to manage or operate them.
Java-Based Framework	Hadoop is implemented in Java, an open-source, high-level programming language, providing the foundation for building distributed storage and processing solutions.
Map Process	The initial step in Hadoop’s MapReduce programming model, where data is processed in parallel on individual cluster nodes, often used for data transformation tasks.
Measured Service	A characteristic where users are billed for cloud resources based on their actual usage, with resource utilization transparently monitored, measured, and reported.
On-Demand Self-Service	The capability for users to access and provision cloud resources such as processing power, storage, and networking using simple interfaces without human interaction with service providers.
Rapid Elasticity	The ability to quickly scale cloud resources up or down based on demand, allowing users to access more resources when needed and release them when not in use.
Reduce Process	The second step in Hadoop's MapReduce model is where results from the mapping process are aggregated and processed further to produce the final output, typically used for analysis.
Replication	The act of creating copies of data pieces within a big data cluster enhances fault tolerance and ensures data availability in case of hardware or node failures.
Resource Pooling	A cloud characteristic where computing resources are shared and dynamically assigned to multiple consumers, promoting economies of scale and cost-efficiency.
Skills Network Labs (SN Labs)	Learning resources provided by IBM, including tools like Jupyter Notebooks and Spark clusters, are available to learners for cloud data science projects and skill development.

Deep Learning and Machine Learning (Data Sciences)

Term	Definition
Artificial Neural Networks	Collections of small computing units (neurons) that process data and learn to make decisions over time.
Bayesian Analysis	A statistical technique that uses Bayes' theorem to update probabilities based on new evidence.
Business Insights	Accurate insights and reports generated by generative AI can be updated as data evolves, enhancing decision-making and uncovering hidden patterns.
Cluster Analysis	The process of grouping similar data points together based on certain features or attributes.
Coding Automation	Using generative AI to automatically generate and test software code for constructing analytical models, freeing data scientists to focus on higher-level tasks.
Data Mining	The process of automatically searching and analyzing data to discover patterns and insights that were previously unknown.
Decision Trees	A type of machine learning algorithm used for decision-making by creating a tree-like structure of decisions.
Deep Learning Models	Includes Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) that create new data instances by learning patterns from large datasets.
Five V's of Big Data	Characteristics used to describe big data: Velocity, volume, variety, veracity, and value.
Generative AI	A subset of AI that focuses on creating new data, such as images, music, text, or code, rather than just analyzing existing data.
Market Basket Analysis	Analyzing which goods tend to be bought together is often used for marketing insights.
Naive Bayes	A simple probabilistic classification algorithm based on Bayes' theorem.
Natural Language Processing (NLP)	A field of AI that enables machines to understand, generate, and interact with human language, revolutionizing content creation and chatbots.

Precision vs. Recall	Metrics are used to evaluate the performance of classification models.
Predictive Analytics	Using machine learning techniques to predict future outcomes or events.
Synthetic Data	Artificially generated data with properties similar to real data, used by data scientists to augment their datasets and improve model training.