Big Data and Data Mining Lesson Glossary

Welcome! This glossary contains many of the terms in this lesson. These terms are important for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

Term	Definition	Video where the term is introduced
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.	Data Scientists at New York University
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.	How Big Data is Driving Digital Transformation
Big Data Cluster	A distributed computing environment comprising thousands or tens of thousands of interconnected computers that collectively store and process large datasets.	What is Hadoop?
Broad Network Access	The ability to access cloud resources via standard mechanisms and platforms such as mobile devices, laptops, and workstations over networks.	Introduction to Cloud
Chief Data Officer (CDO)	An emerging role responsible for overseeing data-related initiatives, governance, and strategies, ensuring that data plays a central role in digital transformation efforts.	How Big Data is Driving Digital Transformation
Chief Information Officer (CIO)	An executive is responsible for managing an organization's information technology and computer systems, contributing to technology-related aspects of digital transformation.	How Big Data is Driving Digital Transformation
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.	Introduction to Cloud
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.	Introduction to Cloud
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.	Introduction to Cloud
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.	What is Hadoop?
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.	Cloud for Data Science
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.	What is Hadoop?
Data Science	An interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including programming, statistics, and analytical tools.	Data Scientists at New York University
Deep Learning	A subset of machine learning that involves artificial neural networks inspired by the human brain, capable of learning	Data Scientists at New York University

Term	Definition	Video where the term is introduced
	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.	How Big Data is Driving Digital Transformation
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.	How Big Data is Driving Digital Transformation
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.	What is Hadoop?
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.	Data Scientists at New York University
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.	What is Hadoop?
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.	Introduction to Cloud
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.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
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.	What is Hadoop?
Measured Service	A characteristic where users are billed for cloud resources based on their actual usage, with resource utilization transparently monitored, measured, and reported.	Introduction to Cloud
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.	Introduction to Cloud
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.	Introduction to Cloud
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.	What is Hadoop?
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.	What is Hadoop?
Resource Pooling	A cloud characteristic where computing resources are shared and dynamically assigned to multiple consumers, promoting economies of scale and cost-efficiency.	Introduction to Cloud
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.	Cloud for Data Science
Spilling to Disk	A technique used in memory-constrained situations where data is temporarily written to disk storage when memory	Big Data Processing Tools: Hadoop, HDFS,

Term	Definition	Video where the term is introduced
	resources are exhausted, ensuring uninterrupted processing.	Hive, and Spark
STEM Classes	Science, Technology, Engineering, and Mathematics (STEM) courses typically taught in high schools prepare students for technical careers, including data science.	Data Scientists at New York University
Variety	The diversity of data types, including structured and unstructured data from various sources such as text, images, video, and more, posing data management challenges.	Foundations of Big Data
Velocity	The speed at which data accumulates and is generated, often in real-time or near-real-time, drives the need for rapid data processing and analytics.	Foundations of Big Data
Veracity	The quality and accuracy of data, ensuring that it conforms to facts and is consistent, complete, and free from ambiguity, impacts data reliability and trustworthiness.	Foundations of Big Data
Video Tracking System	A system used to capture and analyze video data from games, enabling in-depth analysis of player movements and game dynamics, contributing to data-driven decision-making in sports.	How Big Data is Driving Digital Transformation
Volume	The scale of data generated and stored is driven by increased data sources, higher-resolution sensors, and scalable infrastructure.	Foundations of Big Data
V's of Big Data	A set of characteristics common across Big Data definitions, including Velocity, Volume, Variety, Veracity, and Value, highlighting the rapid generation, scale, diversity, quality, and value of data.	Foundations of Big Data