Introduction to Big Data with Spark and Hadoop

Module 1 Glossary: What Is Big Data?

Welcome! This alphabetized glossary contains many of the terms in this course. This comprehensive glossary also includes additional industry-recognized terms not used in course videos. These terms are essential for you to recognize when working in the industry, participating in user groups, and in other professional certificate programs.

Estimated reading time: 12 minutes

Term	Definition
Apache Spark	An open-source, in-memory application framework used for distributed data processing and iterative analysis of large data sets.
Apache HBase	A robust NoSQL datastore that efficiently manages storage and computation resources independently of the Hadoop ecosystem.
Business intelligence (BI)	Encompasses various tools and methodologies designed to convert data into actionable insights efficiently.
Big data	Data sets whose volume, velocity, or variety exceeds the capacity of conventional relational databases to effectively manage, capture, and process with minimal latency. Key characteristics of big data include substantial volume, high velocity, and diverse variety.
Big data analytics	Uses advanced analytic techniques against large, diverse big data sets that include structured, semi-structured, and unstructured data from different sources and sizes, from terabytes to zettabytes. It helps companies gain insights from the data collected by IoT devices.
Big data programming tools	Programming tools are the final component of big data commercial tools. These programming tools perform large-scale analytical tasks and operationalize big data. They also provide all necessary functions for data collection, cleaning, exploration, modeling, and visualization. Some popular tools you can use for programming include R, Python, SQL, Scala, and Julia.
Committer	Most open-source projects have formal processes for contributing code and include various levels of influence and obligation to the project: Committer, contributor, user, and user group. Typically, committers can modify the code directly.
Cloud computing	Allows customers to access infrastructure and applications over the internet without needing on-premises installation and maintenance. By leveraging cloud computing, companies can utilize server capacity on-demand and rapidly scale up to handle the extensive computational requirements of processing large data sets and executing complex mathematical models.
Cloud providers	Offer essential infrastructure and support, providing shared computing resources encompassing computing power, storage, networking, and analytical software. These providers also offer software as a service model featuring specific solutions, enabling enterprises to gather, process, and visualize data efficiently. Prominent examples of cloud service providers include AWS, IBM, GCP, and Oracle.
Extract, transform, and load (ETL)	A systematic approach that involves extracting data from various sources, transforming it to meet specific requirements, and loading it into a data warehouse or another centralized

Term	Definition
process	data repository.
Hadoop	An open-source software framework that provides dependable distributed processing for large data sets through the utilization of simplified programming models.
Hadoop Distributed File System (HDFS)	A file system distributed on multiple file servers, allowing programmers to access or store files from any network or computer. It is the storage layer of Hadoop. It works by splitting the files into blocks, creating replicas of the blocks, and storing them on different machines. It is built to access streaming data seamlessly. It uses a command-line interface to interact with Hadoop.
Hive	A data warehouse infrastructure employed for data querying and analysis, featuring an SQL-like interface. It facilitates report generation and utilizes a declarative programming language, enabling users to specify the data they want to retrieve.
Internet of Things (IoT)	A system of physical objects connected through the internet. A thing or device can include a smart device in our homes or a personal communication device such as a smartphone or computer. These collect and transfer massive amounts of data over the internet without manual intervention by using embedded technologies.
Machine data	Refers to information generated by various sources, including the Internet of Things (IoT) sensors embedded in industrial equipment, as well as weblogs that capture user behavior and interactions.
Map	MapReduce converts a set of data into another set of data, and the elements are fragmented into tuples (key or value pairs).
MapReduce	A program model and processing technique used in distributed computing based on Java. It splits the data into smaller units and processes big data. It is the first method used to query data stored in HDFS. It allows massive scalability across hundreds or thousands of servers in a Hadoop cluster.
NoSQL databases	NoSQL databases are built from the ground up to store and process vast amounts of data at scale and support a growing number of modern businesses. NoSQL databases store data in documents rather than relational tables. Types of NoSQL databases include pure document databases, key-value stores, wide-column databases, and graph databases such as MongoDB, CouchDB, Cassandra, and Redis.
Open-source software	Not only is the runnable version of the code free, but the source code is also completely open, meaning that every line of code is available for people to view, use, and reuse as needed.
Price analytics	Helps understand market segmentation, identify the best price points for a product line, and perform margin analysis for maximum profitability.
Relational databases	Data is structured in the form of tables, with rows and columns, collectively forming a relational database. These tables are interconnected using primary and foreign keys to establish relationships across the data set.
Sentiment analysis	Utilizes social media conversations to gain insights into consumer opinions about a product. It is used to develop effective marketing strategies and establish customer connections based on their sentiments and preferences.
Social data	Comes from the likes, tweets and retweets, comments, video uploads, and general media that are uploaded and shared via the world's favorite social media platforms. Machine-generated data and business-generated data are data that organizations generate within their own operations.
Transactional data	Generated from all the daily transactions that take place both online and offline, such as invoices, payment orders, storage records, and delivery receipts.
Velocity	The speed at which data arrives. Velocity is one of the four main components used to describe the dimensions of big data.

Term	Definition
Volume	The increase in the amount of data stored over time. Volume is one of the four main components used to describe the dimensions of big data.
Variety	The diversity of data or the various data forms that need to be stored. Variety is one of the four main components used to describe the dimensions of big data.
Veracity	The certainty of data, as with a large amount of data available, makes it difficult to determine if the data collected is accurate. Veracity is one of the four main components used to describe the dimensions of big data.
Yet Another Resource Negotiator (YARN)	Serves as the resource manager bundled with Hadoop and is typically the default resource manager for numerous big data applications, such as HIVE and Spark. While it remains a robust resource manager, it's important to note that more contemporary container-based resource managers, such as Kubernetes, are gradually emerging as the new standard practices in the field.

Author(s)

- Rashi Kapoor

Changelog

Date	Version	Changed by	Change Description
2023-09-11	0.2	Kunal Merchant	Basic QC edit
2023-09-05	0.1	Sameeksha Saxena	Initial version created