Introduction to Big Data with Spark and Hadoop

Module 2 Glossary: Introduction to Hadoop Ecosystem

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 to recognize when working in the industry, participating in user groups, and other professional certificate programs.

Estimated reading time: 12 minutes

Term	Definition
Anomaly detection	A process in machine learning that identifies data points, events, and observations that deviate from a data set's normal behavior. Detecting anomalies from time series data is a pain point that is critical to address for industrial applications.
Apache	This open-source HTTP server implements current HTTP standards to be highly secure, easily configurable, and highly extendible. The Apache Software License by the Apache Software Foundation builds and distributes it.
Apache Cassandra	It is a scalable, NoSQL database specifically designed not to have a single point of failure.
Apache Nutch	An extensible and scalable web crawler software product to aggregate data from the web.
Apache ZooKeeper	A centralized service for maintaining configuration information to maintain healthy links between nodes. It provides synchronization across distributed applications. It also tracks server failure and network partitions by triggering an error message and then repairing the failed nodes.
Big data	Data sets whose type or size supersedes the ability of traditional relational databases to manage, capture, and process the data with low latency. Big data characteristics include high volume, velocity, and variety.
Big data analytics	Uses advanced analytic techniques against large, diverse big data sets, including structured, semi-structured, and unstructured data, from varied sources and sizes, from terabytes to zettabytes.

Term	Definition
Block	Minimum amount of data written or read, and also offers fault tolerance. The default block size can be 64 or 128 MB, depending on the user's system configuration. Each file stored need not take up the storage of the preconfigured block size.
Clusters	These servers are managed and participate in workload management. They allow enterprise applications to supersede the throughput achieved with a single application server.
Command-line interface (CLI)	Used to enter commands that enable users to manage the system.
Commodity hardware	Consists of low-cost workstations or desktop computers that are IBM-compatible and run multiple operating systems such as Microsoft Windows, Linux, and DOS without additional adaptations or software.
Data ingestion	The first stage of big data processing. It is a process of importing and loading data into IBMÂ® WatsonX.data. You can use the Ingestion jobs tab from the Data manager page to load data securely and easily into WatsonX.data console.
Data sets	Created by extracting data from packages or data modules. They gather a customized collection of items that you use frequently. As users update their data set, dashboards and stories are also updated.
Data warehouse	Stores historical data from many different sources so users can analyze and extract insights from it.
Distributed computing	A system or machine with multiple components on different machines. Each component has its own job, but the components communicate with each other to run as one system for the end user.
Driver	Receives query statements submitted through the command line and sends the query to the compiler after initiating a session.
Executor	Executes tasks after the optimizer has split the tasks.
Extended Hadoop Ecosystem	Consists of libraries or software packages commonly used with or installed on top of the Hadoop core.
Fault tolerance	A system is fault-tolerant if it can continue performing despite parts failing. Fault tolerance helps to make your remote-boot infrastructure more robust. In the case of OS deployment servers, the whole system is fault-tolerant if the servers back up each other.
File system	An all-comprehensive directory structure with a root (/) directory and other directories and files under a logical volume. The complete

Term	Definition
	information about the file system centralized in the /etc/filesystems file.
Flume	A distributed service that collects, aggregates, and transfers big data to the storage system. Offers a simple yet flexible architecture that streams data flows and uses an extensible data model, allowing online analytic applications.
Hadoop	An open-source software framework offering reliable distributed processing of large data sets using simplified programming models.
Hadoop Common	Fundamental part of the Apache Hadoop framework. It refers to a collection of primary utilities and libraries that support other Hadoop modules.
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 can access streaming data seamlessly. It uses a command-line interface to interact with Hadoop.
Hadoop Ecosystem	It splits big data analytics processing tasks into smaller tasks. The small tasks are performed in conjunction using an algorithm (MapReduce) and then distributed across a Hadoop cluster (nodes that perform parallel computations on big data sets).
Hadoop Ecosystem stages	The four main stages are: Ingest, store, process, analyze, and access.
HBase	A column-oriented, non-relational database system that runs on top of the Hadoop Distributed File System (HDFS). It provides real-time wrangling access to the Hadoop file system. It uses hash tables to store data in indexes and allow for random data access, making lookups faster.
High-throughput	Throughput quantifies the data processed in a timeframe. The target system needs robust throughput for heavy workloads with substantial data changes from the source database to prevent latency spikes. Performance objectives are frequently outlined with throughput targets. High throughput is achieved when most messages are delivered successfully, whereas low successful delivery rates indicate poor throughput and network performance.
Hive	It is a data warehouse infrastructure used in data query and analysis with an SQL-like interface. It helps in generating and creating reports. It is a declarative programming language allowing users to express which data they wish to receive.

Term	Definition
Hive client	Hive provides different communication drivers depending on the application type. For example, Java-based applications use JDBC drivers, and other applications use ODBC drivers. These drivers communicate with the servers.
Hive server	Used to execute queries and enable multiple clients to submit requests. It can support JDBC and ODBC clients.
Hive services	Client interactions and query operations are done through the Hive services. The command-line interface acts as an interface for the Hive service. The driver takes in query statements, monitors each session's progress and life cycle, and stores metadata generated from the query statements.
Hive Web Interface	A web-based user interface that interacts with Hive through a web browser. It offers a graphical user interface (GUI) to browse tables, execute Hive queries, and manage Hive resources.
HMaster	The master server that monitors the region server instances. It assigns regions to region servers and distributes services to different region servers. It also manages any changes to the schema and metadata operations.
Hue	An acronym for Hadoop user experience. It allows you to upload, browse, and query data. Users can run Pig jobs and workflow in Hue. It also provides an SQL editor for several query languages, like Hive and MySQL.
Impala	A scalable system that allows nontechnical users to search for and access the data in Hadoop.
InputSplits	Created by the logical division of data. They serve as an input to a single Mapper job.
JDBC client	Component in the Hive client allows Java-based applications to connect to Hive.
Low latency data access	A type of data access allowing minimal delays, not noticeable to humans, between an input processed and corresponding output offering real-time characteristics. It is crucial for internet connections using trading, online gaming, and Voice over IP.
Map	Job in MapReduce converts a set of data into another set of data. The elements fragment into tuples (key/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

Term	Definition
	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.
Meta store	Stores the metadata, the data, and information about each table, such as the location and schema. In turn, the meta store, file system, and job client communicate with Hive storage and computing to perform the following: Metadata information from tables store in some databases and query results, and data loaded from the tables store in a Hadoop cluster on HDFS.
Node	A single independent system for storing and processing big data. HDFS follows the primary/secondary concept.
ODBC (Open Database Connectivity) Client	Component in the Hive client, which allows applications based on the ODBC protocol to connect to Hive.
Optimizer	Performs transformations on the execution and splits the tasks to help speed up and improve efficiency.
Parallel computing	Workload for each job is distributed across several processors on one or more computers, called compute nodes.
Parser	A program that interprets the physical bit stream of an incoming message and creates an internal logical representation of the message in a tree structure. The parser also regenerates a bit stream from the internal message tree representation for an outgoing message.
Partitioning	This implies dividing the table into parts depending on the values of a specific column, such as date or city.
Pig Hadoop component	Famous for its multi-query approach, it analyzes large amounts of data. It is a procedural data flow and programming language that follows an order and set of commands.
Primary node	Also known as the name node, it regulates client file access and maintains, manages, and assigns tasks to the secondary node. The architecture is such that per cluster, there is one name node and multiple data nodes, the secondary nodes.
Rack	The collection of about forty to fifty data nodes using the same network switch.
Rack awareness	When performing operations such as read and write, the name node maximizes performance by choosing the data nodes closest to themselves. Developers can select data nodes on the same rack or nearby racks. It

Term	Definition
	reduces network traffic and improve cluster performance. The name node keeps the rack ID information to achieve rack awareness.
Read	In this operation, the client will request the primary node to acquire the location of the data nodes containing blocks. The client will read files closest to the data nodes.
Reduce	Job in MapReduce that uses output from a map as an input and combines data tuples into small sets of tuples.
Region	The basic building element and most negligible unit of the HBase cluster, consisting of column families. It contains multiple stores, one for each column family, and has two components: HFile and MemStore.
Region servers	These servers receive read and write requests from the client. They assign the request to a region where the column family resides. They serve and manage regions present in a distributed cluster. The region servers can communicate directly with the client to facilitate requests.
Relational database	Data is organized into rows and columns collectively, forming a table. The data is structured across tables, joined by a primary or a foreign key.
Relational Database Management System (RDBMS)	Traditional RDBMS maintains a database and uses the structured query language, SQL. It is suited for real-time data analysis, like data from sensors. It allows for as many read-and-write operations as a user may require. It can handle up to terabytes of data. It enforces that the schema must verify loading data before it can proceed. It may not always have built-in support for data partitioning.
Replication	The process of creating a copy of the data block. It is performed by rack awareness as well. It is done by ensuring data node replicas are in different racks. So, if a rack is down, users can obtain the data from another rack.
Replication factor	Defined as the number of times you make a copy of the data block. Users can set the number of copies they want, depending on their configuration.
Schema	It is a collection of named objects. It provides a way to group those objects logically. A schema is also a name qualifier; it provides a way to use the same natural name for several objects and prevent ambiguous references.
Secondary node	This node is also known as a data node. There can be hundreds of data nodes in the HDFS that manage the storage system. They perform read and write requests at the instructions of the name node. They also create, replicate, and delete file blocks based on instructions from the name node.

Term	Definition
Semi-structured data	Semi-structured data (JSON, CSV, XML) is the "bridge" between structured and unstructured data. It does not have a predefined data model and is more complex than structured data, yet easier to store than unstructured data.
Shuffle	Phase in which interim map output from mappers transfers to reducers. Every reducer fetches interim results for all values associated with the same key from multiple nodes. This is a network-intensive operation within the Hadoop cluster nodes.
Sqoop	An open-source product designed to transfer bulk data between relational database systems and Hadoop. It looks at the relational database and summarizes the schema. It generates MapReduce code to import and export data. It helps develop other MapReduce applications that use the records stored in HDFS.
Streaming	Implies HDFS provides a constant bitrate when transferring data rather than having the data transferred in waves.
Structured data	Structured data, typically categorized as quantitative data, is highly organized and easily decipherable by machine learning algorithms. Developed by IBM in 1974, structured query language (SQL) is the programming language used to manage structured data.
Unstructured data	Information lacking a predefined data model or not fitting into relational tables.
Write	In this operation, the Name node ensures that the file does not exist. If the file exists, the client gets an IO Exception message. If the file does not exist, the client is given access to start writing files.
Yet Another Resource Negotiator (YARN)	Prepares Hadoop for batch, stream, interactive, and graph processing.

Author(s)

- Niha Ayaz Sultan

Changelog

Date	Version	Changed by	Change Description
2023-09-06	0.2	Mary Stenberg	QA Pass with edits
2023-09-05	0.1	Sameeksha Saxena	Initial version created