Glossary: The Data Ecosystem

Welcome! This alphabetized glossary contains many terms used in this course. Understanding these terms is essential when working in the industry, participating in user groups, or participating in other certificate programs.

Estimated reading time: 8 minutes

Term	Definition
Apache Airflow	An open-source workflow management platform for data engineering pipelines.
Apache Beam	An open-source, unified programming model for batch and streaming data processing pipelines.
Apache HBase	A non-relational database that runs on Hadoop, providing real-time access to large data sets.
Apache Kafka	An open-source software platform used to handle real-time data feeds.
Apache Storm	A framework for distributed stream processing computation primarily written in the Clojure programming language.
Apache Spark Streaming	An extension of the core Spark API that allows for fault-tolerant stream processing of live data streams with high throughput and scalability.
Atomicity, consistency, isolation, and durability (ACID) compliance	A group of characteristics that ensure dependable and uniform processing of transactions in a database system.
BeautifulSoup	A Python library to get data out of HTML, XML, and other markup languages.
Big data stores	A larger, more complex data set, especially from new data sources.
Big data	A dynamic, large, and disparate volume of data being created by people, tools, and machines.
Cloudant	A fully managed, distributed database optimized for heavy workloads and fast-growing web and mobile apps.
Comma-separated values (CSV)	A text-formatted file uses commas to separate the values.
Conceptual data model	The model, created by business stakeholders and data architects, defines the system's scope, concepts, and rules.
CouchDB	An open-source NoSQL document database that collects and stores data in JSON-based document formats. Unlike relational databases, CouchDB uses a schema-free data model, which simplifies record management across various computing devices, mobile phones, and web browsers.
Customer relationship management (CRM) software	Software that helps companies measure and control their lead generation and sales pipelines.
Data abstraction	The process of simplifying a set of data to represent the whole.
Data analyst	A data professional who first gathers and understands the data, then analyzes and interprets it before visualizing it and, finally, weaving it into a story.
Data analytics	Focuses on extracting valuable information from data using various tools, techniques, processes, and algorithms. It includes data analysis and the interpretation of the results, keeping in mind specific business objectives.
Data fabric	An architecture that facilitates the end-to-end integration of various data pipelines and cloud environments through intelligent and automated systems.
Data integration	The combination of technical and business processes that are used to combine data from disparate sources into meaningful and valuable information.
Data lakes	A centralized repository designed to store, process, and secure large amounts of structured, semistructured, and unstructured data. It can store data in its native format and process any variety, ignoring size limits.
Data lookup	A way to fill in information based on rules.
Data marts	Data warehouses are segmented into smaller subsets, known as data marts. These data marts are designed to manage specific business functions, departments, or subject areas. By doing so, data marts make it easier for a defined group of users to access specific data, enabling them to quickly find crucial insights without wasting time searching through an entire data warehouse.
Data modeling	Creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures.D
Data repository	Data sets isolated to be mined for reporting and analysis. It is also known as a data archive or library.
Data science	Process that focuses on understanding the data. This involves data analysis, beginning with data loading, exploring, and cleaning. It creatively explores data, coming up with new solutions and inventions.
Data source	The physical or digital location where the data is held in a data table, object, or other storage format.
Data streams	The process of transmitting continuous data and feeding it into stream processing software to derive valuable insights.
Data visualization	The graphical representation of information and data. It helps data visualization to understand trends, outliers, and patterns in data.
Data warehouses	A storage architecture that pulls data from many sources into a single data repository for sophisticated analytics and decision support.
Database as a service	A cloud-computing service that allows users to access and use a cloud database system without purchasing and setting up their own hardware, installing their own database software, or managing the database themselves.
Database Management System (DBMS)	Software to store and retrieve users' data by considering the security of their information.
Denodo	A unified virtual data layer that allows enterprise users to access data across formats, protocols, and locations using techniques like search.
DocumentDB	A NoSQL database service that supports document data structures with some MongoDB 3.6 and 4.0 compatibility.
DynamoDB	A type of database developed by Amazon Web Services (AWS).
Enterprise resource planning (ERP) systems	A type of software system that enables businesses to automate and efficiently manage their key business processes to gain optimal performance.
Entity-relationship model (E-R model)	A high-level data model is created to define the data elements and their relationships for a specific system. It develops a conceptual design for the database and presents a simple and easy-to-design data view.
Extract, load, transform (ETL) process	A process that extracts, loads, and transforms data from multiple sources to a data warehouse or other unified data repository.
Flat files	Collection of data that is stored specifically in a two-dimensional database. It usually contains a series of records (or lines), where each record is usually a sequence of fields.
Global Positioning Systems (GPS)	A radio navigation system that accurately determines location, time, and velocity regardless of weather conditions.
Hadoop Distributed File System (HDFS)	A storage system for big data that runs on multiple commodity hardware devices connected through a network. HDFS provides scalable and reliable big data storage by partitioning files over multiple nodes.
Hierarchical model	A data model in which the data are organized into a tree-like structure.
Hive	A data warehouse for data query and analysis built on top of Hadoop.
Hadoop	A collection of tools that provides distributed storage and processing of big data.
Java	A programming language known for its platform independence, which allows Java programs to run on different operating systems without modification.
JavaScript object notation (JSON)	An open standard file format that uses readable text to store and transmit data objects consisting of attributes.
Linux	An open-source operating system developed from Unix.
Logical data model	Provides detailed descriptions of data elements and is utilized to create visual representations of data entities, attributes, keys, and relationships.
MongoDB	An open-source, nonrelational database management system (DBMS) that uses flexible documents instead of tables and rows to process and store various forms of data.
MySQL	An open-source relational database management system (RDBMS).
Network model	A database model conceived as a flexible way of representing objects and their relationships.
NoSQL database	A non-tabular database that stores data with different data storage tables than relational tables.
Online analytical processing (OLAP)	Software that is used to conduct multidimensional analysis on large volumes of data from a data warehouse, data mart, or other centralized data store.
Online Transaction Processing (OLTP)	A computerized system that allows real-time data processing and immediate response to users' queries.
Oracle Cloud	A cloud platform that offers complete cloud application suites across software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS).
Oracle database	A multi-model database management system generally used for online transaction processing (OLTP), data warehousing, and both workloads.
Pandas	A Python library used to work with data sets.
Physical data model	A database-specific model that represents relational data objects (for example, tables, columns, primary and foreign keys) and their relationships.
Platform as a service	A cloud computing model that provides customers a complete cloud platform, hardware, software, and infrastructure, for developing, running, and managing applications without the cost, complexity, and inflexibility that often come with building and maintaining that platform on-premises.
PostgreSQL	An open-source database that has a strong reputation for its reliability, flexibility, and support of open technical standards.
PowerShell	A cross-platform command-line shell and scripting language designed for automating tasks and managing configurations.
Python	An agile, dynamically typed, expressive, open-source programming language that supports multiple programming philosophies, including procedural, object-oriented, and functional. Python is a popular high-level programming language that is easily extensible through the use of third-party packages and often allows powerful functions to be written with a few lines of code.
Radio Frequency Identification (RFID) tags	A method for tracking goods through their tags.
Relational Database Service (RDS)	Organizes data into rows and columns, which collectively form a table. Data is typically structured across multiple tables, which can be joined together via a primary key or a foreign key.
Relational model	An approach to managing data using a structure and language consistent with first-order predicate logic.
Scala	A programming language designed for concise, elegant, and type-safe expression of programming patterns. This language seamlessly integrates object-oriented and functional features.
Scrapy	A free and open-source web-crawling framework written in Python and developed in Cambuslang.
Selenium	A testing platform for an open-source web user interface.
Spark	A distributed data analytics framework designed to perform complex data analytics in real-time.
SQL	Computer language used to interact with a relational database.
SQL database	A collection of highly structured tables where each row represents a data entity and every column represents a specific information field.
Statistical Analysis System (SAS)	A programming language that provides all the tools necessary to read, write, and create system files, SAS databases, and reports.
Structured data	The data that conforms to a defined structure follows a consistent order and is easily accessible to people or computer programs.
Tab-separated values (TSV)	A text-based file format that stores data.
Talend Open Studio	A free, open-source ETL tool for data integration and big data.
Unix	A group of multitasking, multiuser computer operating system.
Unstructured data	Typically categorized as qualitative data that cannot be processed and analyzed via conventional data tools and methods.
Velocity	A tool to provide insights to the business about how well software delivery is working and where to focus new processes, resources, or more automation.
Veracity	The term "Veracity" was coined by IBM to describe the challenges of managing data from disparate sources, which can be inconsistent and unreliable.
Web scraping	A technique used to collect online content and data generally gets saved in a local file so as to manipulate and analyze as needed.

Author(s)

- Bhavika Chhatbar