Glossary: Gathering and Wrangling Data

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, and participating in other certificate programs.

Estimated reading time: 7 minutes

Term	Definition
Amazon Web Services (AWS)	A cloud-computing platform that offers economical online app development tools and services.
Application programming interface (API)	APIs are predefined rules that allow software applications to communicate with each other, process data transfers, and enable the sharing of application data and functionality between companies and external parties.
Customer relationship management (CRM) software	Helps companies measure and control their lead generation and sales pipelines.
Data analysis	Process that involves cleaning, transforming, and modeling data to uncover useful information to aid business decision-making.
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 cleaning	An essential process of preparing and validating data before performing analysis. It is an integral part of building a data model.
Data collection	The collection and evaluation of information from multiple sources to find answers, assess outcomes, and forecast trends.
Data encryption	The process of translating data from a readable to an unreadable format, also known as ciphertext, to protect sensitive information during transmission or storage.
Data governance policy	A document that outlines an organization's data management policies, roles, responsibilities, data quality, security, and access.
Data integration	The combination of technical and business processes that 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, semi-structured, and unstructured data. It can store data in its native format and process a variety of data while ignoring size limits.
Data marts	Partitions to manage one specific business function, department, or subject area in data warehouses, also called subsets. Data marts make detailed data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.
Data modeling	The process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures.
Data profiling	The process used to analyze data in a data warehouse for structure, content, relationships, and derivation rules. Profiling identifies anomalies, assesses data quality, and discovers, registers, and assesses metadata.
Data refinery	A tool that helps to explore data residing in a spectrum of data sources.
Data reporting	The gathering of raw data from various sources and transforming it into meaningful, easy-to-understand information to gain valuable insights into business performance.
Data repository	Also termed a data archive or library, it refers to a data set identified to be mined for reporting and analysis.
Data source	The physical or digital location where data is held in a data table, object, or other storage format.
Data storage	The process of saving digital information using computer devices facilitates the efficient completion of various digital tasks.
Data visualization	A graphical representation of information and data. It helps data visualization to understand trends, outliers, and patterns in data.
Data warehouse	A data warehouse pulls together data from many different sources into a single data repository for sophisticated analytics and decision support.
Data wrangling	An iterative process that involves data exploration, transformation, and validation, making it available for a credible and meaningful analysis.
Descriptive analytics	The process of utilizing statistical techniques to explain or summarize a specific data set. Descriptive analysis is also called descriptive statistics.
Diagnostic analytics	A type of analytics that helps identify the reason an event occurred. It lets you discover hidden correlations and connections between variables, determine causal relationships, detect anomalies, and isolate patterns.
Extract, load, and transport (ETL) process	A process that extracts, loads, and transforms data from multiple sources to a data warehouse or other unified data repository.
Global Positioning Systems (GPS)	A radio navigation system that accurately determines location, time, and velocity regardless of weather conditions.
Google DataPrep	An intelligent cloud data service allows visually exploring, cleaning, and preparing both structured and unstructured data for analysis.
Java	A programming language is 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.
Jupiter Notebook	An open-source web application widely used for data cleaning and transformation, statistical modeling, and data visualization.
Key performance indicator (KPI)	The performance measure for a specific objective that provides targets, milestones, and insights to help teams and individuals make better decisions.
Loatame	A data management platform.
NoSQL database	A non-tabular database that stores data with different data storage tables than relational tables.
Numoy or numerical Python	A package of Python.
Online transaction processing (OLTP)	A computerized system that allows real-time data processing and immediate response to users' queries.
OpenRefine	An open-source tool that allows you to import and export data in various formats, such as TSV, CSV, XLS, XML, and JSON. Using OpenRefine, you can clean data, transform it from one format to another, and extend data with web services and external data.
Pandas	A Python library used to work with data sets.
Predictive analytics	A type of analytics used to predict future outcomes. It relies on historical data and employs various techniques such as statistical modeling, data mining, and machine learning for predictions.
Prescriptive analytics	A type of data analytics that recommends the optimal course of action to achieve a specific goal, drawing from inputs from descriptive, diagnostic, and predictive analytics processes.
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 work with written code.
Relational database	Structured data with a well-defined schema.
Semi-structured data	Data that has some organizational properties but not a rigid schema, such as data from emails, XML, zipped files, binary executables, and TCP/IP protocols.
Snowflake schema	A multi-dimensional data model, an expansion of a star schema. It breaks down dimension tables into subdimensions.
Spreadsheets	Applications such as Microsoft Excel and Google Sheets that have a host of features and built-in formulae that can help you identify issues and clean and transform data.
Structured query language (SQL)	A computer language used to interact with a relational database.
Tableau	A tool to analyze and report large amounts of data through visual representation.
Talend	ETL tool for data integration that provides software solutions for data preparation, quality, integration, application integration, management, and big data. Talend is a separate product for all these solutions.
Trifacta wrangler	An interactive cloud-based service for cleaning and transforming data. It takes messy, real-world data and cleans and rearranges it into data tables, which can then be exported to Excel, Tableau, and R.
Watson Studio refinery	Available via IBM Watson Studio, it allows for discovering, cleaning, and transforming data with built-in operations. It transforms large amounts of raw data into consumable, quality information that's ready for analytics.

Author(s)

- Bhavika Chhatbar

Skills Network