Glossary

This book includes many technical terms that can be confusing to even the most seasoned technologist. This glossary is a list of terms that may be unfamiliar to you.

abstraction — minimizing the complexity of something by hiding the details and just providing the relevant information. It's about providing a high-level specification rather than going into a lot of detail about how something works. In the cloud, for instance, in an IaaS delivery model, the infrastructure is abstracted from the user.

advanced analytics — algorithms for complex analyses of either structured or unstructured data, which includes sophisticated statistical models, machine learning, neural networks, text analytics, and other advanced data mining techniques. Advanced analytics does not include database query, reporting, and OLAP cubes.

algorithm — a step-by-step description of a specific process, procedure, or method.

Apache Spark — an open-source parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.

Apache Software Foundation — a nonprofit, community-led organization responsible for coordinating the development and distribution of more than 150 open source software projects.

API (application programming interface) — a collection of routines, protocols, and tools that define the interface to a software component, allowing external components access to its functionality without requiring them to know internal implementation details.

Big Data — a relative term referring to data that is difficult to process with conventional technology due to extreme values in one or more of three attributes: volume (how much data must be processed), variety (the complexity of the data to be processed) and velocity (the speed at which data is produced or at which it arrives for processing). As data management technologies improve, the threshold for what is considered big data rises. For example, a terabyte of slow-moving simple data was once considered big data, but today that is easily managed. In the future, a yottabyte data set may be manipulated on desktop, but for now it would be considered big data as it requires extraordinary measures to process.

business rules — constraints or actions that refer to the actual commercial world but may need to be encapsulated in service management or business applications.

business service — an individual function or activity that is directly useful to the business.

cache — an efficient memory management approach to ensure that future requests for previously used data can be achieved faster. Cache may be implemented in hardware as a separate high-speed memory component or in software (e.g., in a web browser's cache). In either case, the cache stores the most frequently used data and is the first place searched by an application.

cloud computing — a computing model that makes IT resources such as servers, middleware, and applications available as services to business organizations in a self-service manner.

columnar or column-oriented database — a database that stores data across columns rather than rows. This is in contrast to a relational database that stores data in rows.

construction grammar — an approach to linguistic modeling that uses the "construction" (a pairing of structure and meaning) as the basic unit of language. In NLP, construction grammars are used to search for a semantically defined deep structure.

corpus — a machine-readable representation of the complete record of a particular individual or topic.

data at rest — data at rest is placed in storage rather than used in real time.

data cleansing — software used to identify potential data-quality problems. If a customer is listed multiple times in a customer database because of variations in the spelling of her name, the data-cleansing software makes corrections to help standardize the data.

data federation — data access to a variety of data stores, using consistent rules and definitions that enable all the data stores to be treated as a single resource.

data in motion — data that moves across a network or in-memory for processing in real time.

data mining — the process of exploring and analyzing large amounts of data to find patterns.

data profiling — a technique or process that helps you understand the content, structure, and relationships of your data. This process also helps you validate your data against technical and business rules.

data quality — characteristics of data such as consistency, accuracy, reliability, completeness, timeliness, reasonableness, and validity. Data-quality software ensures that data elements are represented in a consistent way across different data stores or systems, making the data more trustworthy across the enterprise.

data transformation — a process by which the format of data is changed so that it can be used by different applications.

data warehouse — a large data store containing the organization's historical data, which is used primarily for data analysis and data mining. It is the data system of record.

database — a computer system intended to store large amounts of information reliably and in an organized fashion. Most databases provide users convenient access to the data, along with helpful search capabilities.

Database Management System (DBMS) — software that controls the storage, access, deletion, security, and integrity of primarily structured data within a database.

disambiguation — a technique within NLP for resolving ambiguity in language.

distributed computing — the capability to process and manage processing of algorithms across many different nodes in a computing environment.

distributed filesystem — a distributed filesystem is needed to manage the decomposition of structured and unstructured data streams.

elasticity — the ability to expand or shrink a computing resource in real time, based on scaling a single integrated environment to support a business.

ETL (Extract, Transform, Load) — tools for locating and accessing data from a data store (data extraction), changing the structure or format of the data so that it can be used by the business application (data transformation), and sending the data to the business application (data load).

federation — the combination of disparate things so that they can act as one—as in federated states, data, or identity management—and to make sure that all the right rules apply.

framework — a support structure for developing and managing software.

graph databases — makes use of graph structures with nodes and edges to manage and represent data. Unlike a relational database, a graph database does not rely on joins to connect data sources.

governance — the process of ensuring compliance with corporate or governmental rules, regulations, and policies. Governance is often associated with risk management and security activities across computing environments.

Hadoop — an Apache-managed software framework derived from MapReduce. Big Table Hadoop enables applications based on MapReduce to run on large clusters of commodity hardware. Hadoop is designed to parallelize data processing across computing nodes to speed up computations and hide latency. The two major components of Hadoop are a massively scalable distributed file system that can support petabytes of data and a massively scalable MapReduce engine that computes results in batch.

Hadoop Distributed File System (HDFS) — HDFS is a versatile, resilient, clustered approach to managing files in a Big Data environment. HDFS is not the final destination for files. Rather it is a data "service" that offers a unique set of capabilities needed when data volumes and velocity are high.

Hidden Markov Models (HMMs) — statistical models used to interpret "noisy" sequences of words or phrases based on probabilistic states.

hybrid cloud — a computing model that includes the use of public and private cloud services that are intended to work together.

information integration — a process using software to link data sources in various departments or regions of the organization with an overall goal of creating more reliable, consistent, and trusted information.

infrastructure — can be either hardware or software elements that are necessary for the operation of anything, such as a country or an IT department. The physical infrastructure that people rely on includes roads, electrical wiring, and water systems. In IT, infrastructure includes basic computer hardware, networks, operating systems, and other software that applications run on top of.

Infrastructure as a Service (IaaS) — infrastructure, including a management interface and associated software, provided to companies from the cloud as a service.

in-memory database — a database structure in which information is managed and processed in memory rather than on disk.

latency — the amount of time lag that enables a service to execute in an environment. Some applications require less latency and need to respond in near real time, whereas other applications are less time-sensitive.

lexical analysis — a technique used within the context of language processing that connects each word with its corresponding dictionary meaning.

machine learning — a discipline grounded in computer science, statistics, and psychology that includes algorithms that learn or improve their performance based on exposure to patterns in data, rather than by explicit programming.

markup language — a way of encoding information that uses plain text containing special tags often delimited by angle brackets (< and >). Specific markup languages are often created based on XML to standardize the interchange of information between different computer systems and services.

MapReduce — designed by Google as a way of efficiently executing a set of functions against a large amount of data in batch mode. The "map" component distributes the programming problem or tasks across a large number of systems and handles the placement of the tasks in a way that balances the load and manages recovery from failures. When the distributed computation is completed, another function called "reduce" aggregates all the elements back together to provide a result.

metadata — the definitions, mappings, and other characteristics used to describe how to find, access, and use the company's data and software components.

metadata repository — a container of consistent definitions of business data and rules for mapping data to its actual physical locations in the system.

morphology — the structure of a word. Morphology gives the stem of a word and its additional elements of meaning.

multitenancy — refers to the situation in which a single instance of an application runs on a SaaS vendor's servers but serves multiple client organizations (tenants), keeping all their data separate. In a multitenant architecture, a software application partitions its data and configuration so that each customer has a customized virtual application instance.

neural networks — neural network algorithms are designed to emulate human/animal brains. The network consists of input nodes, hidden layers, and output nodes. Each of the units is assigned a weight. Using an iterative approach, the algorithm continuously adjusts the weights until it reaches a specific stopping point.

neuromorphic — refers to a hardware or software architecture designed with elements or components that simulate neural activities.

neurosynaptic — refers to a hardware or software architecture designed with elements or components that simulate the activities of neurons and synapses. (it is a more restrictive term than neuromorphic.)

NoSQL (Not only SQL) — NoSQL is a set of technologies that created a broad array of database management systems that are distinct from a relational database systems. One major difference is that SQL is not used as the primary query language. These database management systems are also designed for distributed data stores.

ontology — a representation of a specific domain that includes relationships between their elements, and often containing rules and relationships between categories and criteria for inclusion within a category.

phonology — the study of the physical sounds of a language and how those sounds are uttered in a particular language.

Platform as a Service (PaaS) — a cloud service that abstracts the computing services, including the operating software and the development, deployment, and management life cycle. It sits on top of Infrastructure as a Service (IaaS).

pragmatics — the aspect of linguistics that tackles one of the fundamental requirements for cognitive computing: the capability to understand the context of how words are used.

process — a high level end-to-end structure useful for decision making and normalizing how things get done in a company or organization.

predictive analytics — a statistical or data mining solution consisting of algorithms and techniques that can be used on both structured and unstructured data (together or individually) to determine future outcomes. It can be deployed for prediction, optimization, forecasting, simulation, and many other uses.

private cloud — unlike a public cloud, which is generally available to the general public, a private cloud is a set of computing resources within the corporation that serves only the corporation, but which is set up to be managed as a set of self-service options.

provisioning — makes resources available to users and software. A provisioning system makes applications available to users and makes server resources available to applications.

public cloud — a resource that is available to any consumer either on a fee per transaction service or as a free service.

quantum computing — an approach to computation based on properties of quantum mechanics, specifically those dealing with elementary units that may exist in multiple states simultaneously (in contrast with binary computers, whose basic elements always resolve to a 1 or 0).

real time — real time processing is used when a computer system accepts and updates data at the same time, feeding back immediate results that influence the data source.

registry — a single source for all the metadata needed to gain access to a web service or software component.

reinforcement learning — a special case of supervised learning in which the cognitive computing system receives feedback on its performance to guide it to a goal or good outcome.

repository — a database for software and components, with an emphasis on revision control and configuration management (where they keep the good stuff, in other words).

Relational Database Management System (RDBMS) — a database management system that organizes data in defined tables.

REST (Representational State Transfer) — REST is designed specifically for the Internet and is the most commonly used mechanism for connecting one web resource (a server) to another web resource (a client). A RESTful API provides a standardized way to create a temporary relationship (also called "loose coupling") between and among web resources.

scoring — the process of assigning a confidence level for a hypothesis.

semantics — in computer programming, what the data means as opposed to formatting rules (syntax).

semi-structured data — semi-structured data has some structures that are often manifested in images and data from sensors.

service — purposeful activity carried out for the benefit of a known target. Services often consist of a group of component services, some of which may also have component services. Services always transform something and complete by delivering an output.

service catalog — a directory of IT services provided across the enterprise, including information such as service description, access rights, and ownership.

SLA (service-level agreement) — an SLA is a document that captures the understanding between a service user and a service provider as to quality and timeliness. It may be legally binding under certain circumstances.

service management — the ability to monitor and optimize a service to ensure that it meets the critical outcomes that the customer values and the stakeholders want to provide.

silo — in IT, a silo is an application, data, or service with a single narrow focus, such as human resources management or inventory control, with no intention or preparation for use by others.

Software as a Service (SaaS) — software as a Service is the delivery of computer applications over the Internet on a per user per month charge basis.

Software Defined Environment (SDE) — an abstraction layer that unifies the components of virtualization in IaaS so that the components can be managed in a unified fashion.

spatial database — a spatial database that is optimized for data related to where an object is in a given space.

SQL (Structured Query Language) — SQL is the most popular computer language for accessing and manipulating databases.

SSL (Secure Sockets Layer) — SSL is a popular method for making secure connections over the Internet, first introduced by Netscape.

streaming data — an analytic computing platform that is focused on speed. Data is continuously analyzed and transformed in memory before it is stored on a disk. This platform allows for the analyzing of large volumes of data in real time.

structured data — data that has a defined length and format. Examples of structured data include numbers, dates, and groups of words and numbers called strings (for example, for a customer's name, address, and so on).

supervised learning — refers to an approach that teaches the system to detect or match patterns in data based on examples it encounters during training with sample data.

Support Vector Machine (SVM) — a machine learning algorithm that works with labeled training data and outputs results to an optimal hyperplane. A *hyperplane* is a subspace of the dimension minus one (that is, a line in a plane).

syntactical analysis — helps the system understand the meaning in context with how the term is used in a sentence.

taxonomy — provides context within the ontology. Taxonomies are used to capture hierarchical relationships between elements of interest. For example, a taxonomy for the U.S. Generally Accepted Accounting Principles

(GAAP) represents the accounting standards in a hierarchical structure that captures the relationships between them.

text analytics — the process of analyzing unstructured text, extracting relevant information, and transforming it into structured information that can be leveraged in various ways.

unstructured data — information that does not follow a specified data format. Unstructured data can be text, video, images, and such.

unsupervised learning — refers to a machine learning approach that uses inferential statistical modeling algorithms to *discover* rather than *detect* patterns or similarities in data. An unsupervised learning system can identify new patterns, instead of trying to match a set of patterns it encountered during training.

Watson — watson is a cognitive system developed by IBM that combines capabilities in NLP, machine learning, and analytics.

XML — the eXtensible Markup Language is a language designed to enable the creation of documents readable by humans and computers. It is formally defined as an open standard by a set of rules under the auspices of the World Wide Web Consortium, an international standards organization.