GENERATION DATA

USING DATA FOR PROFIT

GLOSSARY

This programme has been funded with support from the European Commission

Erasmus+

Welcome to Generation Data's Glossary – our alphabetical list of terms or words found in the specific subject of data, with brief explanations.

Algorithm
A mathematical formula or statistical process run by software to perform an analysis of data. It usually consists of multiple calculations steps and can be used to automatically process data or solve problems.

Anonymization
The severing of links between people in a database and their records to prevent the discovery of the source of the records.

Analytics
The process of collecting, processing and analyzing raw data from a data mass to generate insights that inform fact-based decision-making. In many cases, it involves software-based analysis using algorithms.

Artificial Intelligence
The process of developing intelligence machines and software that can perceive the environment and take the corresponding action as and when required and even learn from those actions.

ACID Test
Stands for atomicity, consistency, isolation, and durability (ACID) test of data. These four attributes are the benchmarks for ensuring the validity of a data transaction.

Amazon Web Services
A collection of cloud computing services offered by Amazon to help businesses carry out large scale computing operations (such as big data projects) without having to invest in their own server

farms and data storage warehouses. Storage space, processing power and software operations are rented rather than having to be bought and installed from scratch.

Batch Processing
A standard computing strategy that involves processing data in large sets. This practice becomes imperative for non-time sensitive work that operates on very large datasets. The process is scheduled and at a later time, the results are retrieved by the system.

Biometrics
Using analytics and technology in identifying people by one or many of their physical characteristics, such as fingerprint recognition, facial recognition, iris recognition, etc.

Call Detail Record (CDR) analysis
CDRs contain data that a telecommunications company collects about phone calls, such as time and length of call. This data can be used in any number of analytical applications.

Cloud Computing
A broad term that refers to any Internet-based application that runs on remote servers, rather than locally. Data stored "in the cloud" is typically accessible over the internet, wherever in the world the owner of that data might be.

Clustering analysis
The process of identifying objects that are similar to each other and cluster them in order to understand the differences as well as the similarities within the data.

Cold data storage
Storing old data that is hardly used on low-power servers. Retrieving the data will take longer.

Comparative analysis
It ensures a step-by-step procedure of comparisons and calculations to detect patterns within very large data sets.

Dashboard
A graphical representation of the analyses performed by algorithms.

Data access
The act or method of viewing or retrieving stored data.

Data aggregation
The act of collecting data from multiple sources for the purpose of reporting or analysis.

Data architecture and design
How enterprise data is structured. The actual structure or design varies depending on the eventual end result required. Data architecture has three stages: conceptual representation of business entities, the logical representation of the relationships among those entities, and the physical construction of the system to support the functionality.

Database
A digital collection of data and the structure around which the data is organized. The data is typically entered into and accessed via a database management system (DBMS).

Database administrator (DBA)
A person, often certified, who is responsible for supporting and maintaining the integrity of the structure and content of a database.

Database management system (DBMS)
Software that collects and provides access to data in a structured format.

Data center
A physical facility that houses a large number of servers and data storage devices. Data centers might belong to a single organization or sell their services to many organizations.

Data cleansing
The process of reviewing and revising data in order to delete duplicates, correct errors and provide consistency.

Data collection
Any process that captures any type of data.

Data custodian
A person responsible for the database structure and the technical environment, including the storage of data.

Data exhaust
The data that a person creates as a byproduct of a common activity–for example, a cell call log or web search history.

Data feed
A means for a person to receive a stream of data. Examples of data feed mechanisms include RSS or Twitter.

Data governance
A set of processes or rules that ensure the integrity of the data and that data management best practices are met.

Data integration
The process of combining data from different sources and presenting it in a single view.

Data integrity
The measure of trust an organization has in the accuracy, completeness, timeliness, and validity of the data.

Data Lake
A storage repository that can hold a huge amount of raw data in its original format. Data Lake uses a flat architecture to store data, unlike a hierarchical data warehouse, which stores data in files or folders.

Data migration
The process of moving data between different storage types or formats, or between different computer systems.

Data mining
The process of deriving patterns or knowledge from large data sets. The purpose is to refine data into a more comprehensible and cohesive set of information.

This programme has been funded with support from the European Commission. The author is solely responsible for this publication (communication) and the Commission accepts no responsibility for any use that may be made of the information contained therein.

Co-funded by the
Erasmus+ Programme
of the European Union

Data model, data modeling
A data model defines the structure of the data for the purpose of communicating between functional and technical people to show data needed for business processes, or for communicating a plan to develop how data is stored and accessed among application development team members.

Data point
An individual item on a graph or a chart.

Data profiling
The process of collecting statistics and information about data in an existing source.

Data quality
The measure of data to determine its worthiness for decision making, planning, or operations.

Data replication
The process of sharing information to ensure consistency between redundant sources.

Data repository
The location of permanently stored data.

Data scientist
An expert in extracting insights and value from data. Usually someone that has skills in analytics, computer science, mathematics, statistics, creativity, data visualisation and communication as well as business and strategy.

Data Science
A discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning, and database engineering to solve complex problems.

Data security
The practice of protecting data from destruction or unauthorized access.

Data set
A collection of data, typically in tabular form.

Data source
Any provider of data–for example, a database or a data stream.

Data steward
A person responsible for data stored in a data field.

Data structure
A specific way of storing and organizing data.

Data visualization
A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.

Data warehouse
A place to store data for the purpose of reporting and analysis.

This programme has been funded with support from the European Commission. The author is solely responsible for this publication (communication) and the Commission accepts no responsibility for any use that may be made of the information contained therein.

Co-funded by the
Erasmus+ Programme
of the European Union

Dark Data
All the data that is gathered and processed by enterprises not used for any meaningful purposes. It is called dark because it is unused and unexplored. This includes social network feeds, call center logs, meeting notes, etc.

Distributed File System
Data storage system designed to store large volumes of data across multiple storage devices (often cloud based commodity servers), to decrease the cost and complexity of storing large amounts of data.

Distributed processing
The execution of a process across multiple computers connected by a computer network.

Econometrics
The application of statistical and mathematical theories in economics for testing hypotheses and forecasting future trends. Econometrics makes use of economic models, tests them through statistical trials and then compare the results against real-life examples. It can be subdivided into two major categories: theoretical and applied.

Extract, transform, and load (ETL)
A process used in data warehousing to prepare data for use in reporting or analytics.
The acronym for extract, transform, and load. It refers to the process of ‘extracting’ raw data, ‘transforming’ by cleaning/enriching the data for ‘fit for use’ and ‘loading’ into the appropriate repository for the future use.

Exploratory analysis
Finding patterns within data without standard procedures or methods. It is a means of discovering the data and to find the data sets main characteristics.

Hadoop
Apache Hadoop is one of the most widely used software frameworks in big data. It is a collection of programs which allow storage, retrieval and analysis of very large data sets using distributed hardware (allowing the data to be spread across many smaller storage devices rather than one very large one).

Internet of Things
Ordinary devices that are connected to the internet at any time anywhere via sensors that collect, analyse and transmit data to increase their usefulness.

Juridical Data Compliance
In the legal terminology, the word “Compliance” refers to the act of adherence to the law of the land. In business terms, in case of any organization, compliance implies strict adherence to the laws, regulations, guidelines, and specification that are relevant to the life cycle of a business entity. Juridical data compliance is commonly used in the context of cloud-based solutions, where the data is stored in a different country or continent. Data storage in a server or data center located in a foreign country must abide by the data security laws of the nation.

Latency
Any delay in a response or delivery of data from one point to another.

Legacy system

Outdated computer systems, programming languages or application software that are used instead of available upgraded versions. Legacy systems are also associated with terminology or processes that are no longer applicable to current contexts or content, thus creating confusion.

Load balancing
The process of distributing workload across a computer network or computer cluster to optimize performance.

Location analytics
Location analytics brings mapping and map-driven analytics to enterprise business systems and data warehouses. It allows you to associate geospatial information with datasets.

Location data
Data that describes a geographic location.

Log file
A log file is defined as a file that maintains a registry of events, processes, messages, and communication between various communicating software applications and the operating system. Log files are present in the executable software, operating systems, and programs whereby all the messages and process details are recorded.

Machine-generated data
Any data that is automatically created from a computer process, application, or other non-human source.

Machine learning
The use of algorithms to allow a computer to analyze data for the purpose of "learning" what action to take when a specific pattern or event occurs.

MapReduce
The software procedure of breaking up an analysis into pieces that can be distributed across different computers in different locations. It first distributes the analysis (map) and then collects the results back into one report (reduce). Several companies including Google and Apache (as part of its Hadoop framework) provide MapReduce tools.

Mashup
The method of merging different datasets into a single application to improve output. For instance, combining job listings with demographic data.

Metadata
The data that serves to provide context or additional information about other data. For example, information about the title, subject, author, typeface, enhancements, and size of the data file of a document constitute metadata about that document. It may also describe the conditions under which the data stored in a database was acquired, its accuracy, date, time, a method of compilation and processing, etc.

Natural Language Processing
Software algorithms designed to allow computers to more accurately understand everyday human speech, allowing us to interact more naturally and efficiently with them.

NoSQL

Database management systems that do not (or not only) use relational tables generally used in traditional database systems. It refers to data storage and retrieval systems that are designed for handling large volumes of data but without tabular categorisation (or schemas).

Object Databases
They store data in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.

Operational Databases
Operational Databases carry out regular operations of an organization and are generally very important to a business. They generally use online transaction processing that allows them to enter, collect and retrieve specific information about the company

Parse
The division of data, such as a string, into smaller parts for analysis.

Predictive analytics
Using statistical functions on one or more datasets to predict trends or future events.

Public data
Public information or data sets that were created with public funding.

Query
A request for data or information from a database table or combination of tables. This data may be generated as results returned by Structured Query Language (SQL) or as pictorials, graphs, or complex results, e.g., trend analyses from data-mining tools. SQL is the most well-known and widely-used query language.

R
A popular open source software environment used for analytics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

Real-time data
Data that is created, processed, stored, analysed and visualized within milliseconds

Recommendation engine
An algorithm that analyzes a customer's purchases and actions on an e-commerce site and then uses that data to recommend complementary products.

Reference data
Data that describes an object and its properties. The object may be physical or virtual.

Root-cause analysis
The process of determining the main cause of an event or problem.

Scalability
The ability of a system or process to maintain acceptable performance levels as workload or scope increases.

Schema
The structure that defines the organization of data in a database system.

This programme has been funded with support from the European Commission. The author is solely responsible for this publication (communication) and the Commission accepts no responsibility for any use that may be made of the information contained therein.

Co-funded by the
Erasmus+ Programme
of the European Union

Search data
Aggregated data about search terms used over time.

Semi-structured data
Data that is not structured by a formal data model, but provides other means of describing the data and hierarchies.

Sentiment analysis
The application of statistical functions on comments people make on the web and through social networks to determine how they feel about a product or company.

Server
A physical or virtual computer that serves requests for a software application and delivers those requests over a network.

Software-As-A-Service (SAAS)
The growing tendency of software producers to provide their programs over the cloud – meaning users pay for the time they spend using it (or the amount of data they access) rather than buying software outright.

Spatial Analysis
It refers to analyzing spatial data such geographic data or topological data to identify and understand patterns and regularities within data distributed in geographic space.

Structured v Unstructured Data: Structured data is anything than can be put into a table and organized in such a way that it relates to other data in the same table. Unstructured data is everything that can't – email messages, social media posts and recorded human speech, for example.

Storage
Any means of storing data persistently.

Structured Query Language (SQL)
A programming language designed specifically to manage and retrieve data from a relational database system.

Tag
A tag is a piece of information that describes the data or content that it is assigned to. Tags are nonhierarchical keywords used for Internet bookmarks, digital images, videos, files and so on.

Taxonomy
Taxonomy refers to the classification of data according to a pre-determined system with the resulting catalog. It provides a conceptual framework for easy access and retrieval.

Text analytics
The application of statistical, linguistic, and machine learning techniques on text-based sources to derive meaning or insight.

Transactional data
Data that changes unpredictably. Examples include accounts payable and receivable data, or data about product shipments.

This programme has been funded with support from the European Commission. The author is solely responsible for this publication (communication) and the Commission accepts no responsibility for any use that may be made of the information contained therein.

Co-funded by the
Erasmus+ Programme
of the European Union

Volume
The amount of data, ranging from megabytes to brontobytes.

Visualization
A visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.

XML Databases
XML Databases allow data to be stored in XML format. XML databases are often linked to document-oriented databases. The data stored in an XML database can be queried, exported and serialized into any format needed.

This programme has been funded with support from the European Commission. The author is solely responsible for this publication (communication) and the Commission accepts no responsibility for any use that may be made of the information contained therein.

Co-funded by the
Erasmus+ Programme
of the European Union