

FREE RESOURCE · REFERENCE · CROSS-CUTTING

Data Glossary

Plain-language definitions of 122 data and analytics terms, with the distinctions practitioners most often get wrong.

At a glance	
What it is	A reference glossary of 122 data, analytics and AI terms in plain English, each tagged by discipline and cross-referenced, plus a section on terms that are commonly confused.
Who it is for	All practitioners and the business colleagues they work with: analysts, engineers, stewards, product owners, managers and anyone new to the field.
How it is used	Look a term up from the A-Z index (every letter and "See also" reference is a link), or use it to align a team on shared vocabulary at the start of a project.
What you provide	Nothing - this is a reference. Where your organisation has its own governed business glossary, your agreed definitions take precedence for business terms.
What the outputs mean	Definitions describe widely used professional meaning. Some terms (for example "data product") are used differently by different vendors and authors; the definition given is the common core.

Version 1.0 · Licence class: Standard Use Licence v1.0 (see the full licence text supplied with this resource) · Publisher: dataintelligenceinsights.com

Conventions used in this resource

Every entry, letter and "See also" reference is a link: select it to jump. The PDF bookmarks panel also lists each section.

This resource is provided under the dataintelligenceinsights.com Standard Use Licence v1.0. © dataintelligenceinsights.com. All rights reserved.

How to use this glossary

Terms are listed alphabetically. Each entry shows the discipline it belongs to, a definition written to be understood without specialist background, and related terms. Legal concepts (such as personal data) are described at a general level; for legal interpretation in your jurisdiction, take professional advice.

Discipline tags: Governance · Architecture · Engineering · Quality · Analytics & BI · AI & ML · Privacy & security · Management

Jump to letter: [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [R](#) [S](#) [T](#) [V](#)

Commonly confused terms

Terms	The distinction
Data governance vs data management	Governance decides and assures (who decides, which rules, how compliance is checked); management executes (the day-to-day work of modelling, integrating, securing and maintaining data).
Data owner vs data steward vs data custodian	The owner is accountable and approves; the steward does the day-to-day definition and quality work; the custodian operates the technology that holds the data.
Business glossary vs data dictionary vs data catalogue	A glossary defines business meaning; a dictionary describes physical fields; a catalogue is the searchable inventory that can hold and link both.
Data warehouse vs data lake vs lakehouse	A warehouse holds modelled, cleansed data for analysis; a lake holds raw data in native formats; a lakehouse adds warehouse-style management and transactions to lake storage via open table formats.
Data mesh vs data fabric	Mesh is mainly an organisational model (domain ownership of data products); fabric is mainly a technology approach (metadata-driven integration across systems). They can coexist.
Metric vs KPI	Every KPI is a metric, but a KPI is tied to a strategic objective and a decision. A dashboard with fifty "KPIs" usually has a handful of KPIs and many metrics.
Master data vs reference data	Master data describes core business entities (customers, products); reference data is the controlled code lists used to classify other data (countries, currencies).
Anonymisation vs pseudonymisation	Anonymised data cannot reasonably be linked back to a person; pseudonymised data can be with additional information and remains personal data.
ETL vs ELT	The difference is where transformation happens: before loading, in a separate engine (ETL), or after loading, inside the target platform (ELT).
Accuracy vs validity	Valid data conforms to rules and formats (a well-formed date); accurate data matches reality (the correct date). Data can be valid and still wrong.

A-Z of terms

A

A/B test ANALYTICS & BI

A controlled experiment that randomly splits users between two variants to measure which performs better on a defined metric, with statistical testing to separate real effects from chance.

See also: [Statistical significance](#)

Access control PRIVACY & SECURITY

Rules and mechanisms that decide who can see or change which data, typically implemented through roles, groups and policies rather than individual grants.

See also: [Role-based access control \(RBAC\)](#), [Least privilege](#)

Ad hoc analysis ANALYTICS & BI

One-off analysis to answer a specific question, as opposed to recurring reporting.

See also: [Self-service analytics](#)

Aggregation ANALYTICS & BI

Combining detailed records into summary values (sum, count, average) at a coarser level of detail.

See also: [Grain](#)

Anonymisation PRIVACY & SECURITY

Irreversibly altering personal data so individuals can no longer be identified by any reasonably likely means. Truly anonymised data generally falls outside data protection law; pseudonymised data does not.

See also: [Pseudonymisation](#), [Personal data](#)

API (application programming interface) ARCHITECTURE

A defined interface through which software systems request and exchange data or functions.

See also: [Data pipeline](#)

B

Batch processing ENGINEERING

Processing data in scheduled groups (for example nightly) rather than continuously as it arrives.

See also: [Stream processing](#), [Micro-batch](#)

Bias (algorithmic) AI & ML

Systematic error in a model's outputs that disadvantages particular groups, often inherited from unrepresentative training data or flawed proxies.

See also: [Fairness](#), [Training data](#)

Business glossary GOVERNANCE

A governed list of business terms with agreed definitions, owners and relationships. It defines meaning; a data dictionary describes physical structure.

See also: [Data dictionary](#), [Data catalogue](#)

Business intelligence (BI) ANALYTICS & BI

The technologies and practices for collecting, integrating, analysing and presenting business data to support decisions, typically through reports and dashboards.

See also: [Dashboard](#), [Semantic layer](#)

C

Canonical model ARCHITECTURE

A shared, system-independent representation of key entities used to translate between systems without point-to-point mappings.

See also: [Data model](#)

Cardinality ARCHITECTURE

The number of distinct values in a column, or the numeric relationship between entities (one-to-one, one-to-many, many-to-many).

See also: [Data model](#)

Change data capture (CDC) ENGINEERING

A technique that identifies and streams inserts, updates and deletes from a source database so targets stay synchronised without full reloads.

See also: [Incremental load](#)

Chief data officer (CDO) MANAGEMENT

The executive accountable for an organisation's data strategy, governance and the value it gains from data.

See also: [Data governance](#), [Data strategy](#)

Cloud data warehouse ARCHITECTURE

A managed, cloud-hosted analytical database that separates storage from compute and scales on demand.

See also: [Data warehouse](#), [Lakehouse](#)

Conformed dimension ARCHITECTURE

A dimension (such as Date or Customer) defined once and shared across multiple fact tables so results can be compared consistently.

See also: [Dimensional model](#), [Dimension table](#)

Critical data element (CDE) GOVERNANCE

A data element whose error would cause material regulatory, financial, customer or operational harm, and which therefore receives enhanced governance and quality control.

See also: [Data quality rule](#), [Data owner](#)

D**Dashboard** ANALYTICS & BI

A visual display of the most important information needed to monitor or decide, arranged on a single screen.

See also: [Key performance indicator \(KPI\)](#), [Data visualisation](#)

Data catalogue GOVERNANCE

A searchable inventory of data assets with technical and business metadata, owners, classifications and often lineage, used to find and understand data.

See also: [Metadata](#), [Business glossary](#)

Data classification PRIVACY & SECURITY

Assigning data to sensitivity categories (for example Public, Internal, Confidential, Restricted) that determine how it must be handled.

See also: [Access control](#), [Personal data](#)

Data contract ENGINEERING

A formal agreement between a data producer and its consumers specifying schema, semantics, quality expectations and service levels for a dataset.

See also: [Data product](#), [Schema evolution](#)

Data custodian GOVERNANCE

The technical role responsible for operating the systems that store and move data, implementing controls defined by owners.

See also: [Data owner](#), [Data steward](#)

Data dictionary GOVERNANCE

A description of the physical data structures - tables, columns, types, allowed values - in a system or dataset.

See also: [Business glossary](#), [Metadata](#)

Data domain GOVERNANCE

A logical grouping of data around a business area (such as Customer, Product or Finance) used to assign ownership.

See also: [Data mesh](#), [Data owner](#)

Data engineering ENGINEERING

The discipline of building and operating the pipelines and platforms that collect, transform and deliver data reliably for analysis and applications.

See also: [Data pipeline](#), [DataOps](#)

Data fabric ARCHITECTURE

An architecture approach that uses metadata, integration and automation to provide unified access to data distributed across many systems.

See also: [Data virtualisation](#), [Data mesh](#)

Data governance GOVERNANCE

The exercise of decision rights, accountability and controls over data so it is managed as an asset - who decides what, by which rules, and how compliance is checked.

See also: [Data management](#), [Data stewardship](#)

Data lake ARCHITECTURE

A repository that stores large volumes of raw data in native formats, typically on low-cost object storage, for later processing.

See also: [Lakehouse](#), [Data swamp](#)

Data lineage GOVERNANCE

The record of where data comes from, how it is transformed and where it goes, from source to consumption.

See also: [Impact analysis](#), [Metadata](#)

Data literacy MANAGEMENT

The ability to read, interpret, question and communicate with data appropriately for one's role.

See also: [Self-service analytics](#)

Data management GOVERNANCE

The broader set of practices - architecture, modelling, integration, quality, security, metadata and more - that deliver and look after data across its lifecycle. Governance directs data management; it does not replace it.

See also: [Data governance](#)

Data mart ARCHITECTURE

A subject-specific subset of a data warehouse built for a particular team or business function.

See also: [Data warehouse](#)

Data mesh ARCHITECTURE

A socio-technical approach that decentralises analytical data ownership to business domains, which publish data as products on a self-service platform under federated governance.

See also: [Data product](#), [Data domain](#)

Data minimisation PRIVACY & SECURITY

The principle of collecting and keeping only the personal data that is adequate, relevant and necessary for a stated purpose.

See also: [Purpose limitation](#), [Retention](#)

Data model ARCHITECTURE

A structured representation of data entities, attributes and relationships, at conceptual, logical or physical level.

See also: [Dimensional model](#), [Normalisation](#)

Data observability ENGINEERING

Automated monitoring of data health - freshness, volume, schema, distribution and lineage - to detect and diagnose issues before consumers do.

See also: [Data quality](#), [DataOps](#)

Data owner GOVERNANCE

The senior business role accountable for a data domain or dataset: approving definitions, access and quality standards.

See also: [Data steward](#), [Data custodian](#)

Data pipeline ENGINEERING

An automated sequence of steps that moves and transforms data from sources to a destination.

See also: [ETL \(extract, transform, load\)](#), [ELT \(extract, load, transform\)](#)

Data platform ARCHITECTURE

The integrated set of technologies used to ingest, store, process, govern and serve data across an organisation.

See also: [Cloud data warehouse](#), [Lakehouse](#)

Data product ARCHITECTURE

A dataset or data service managed like a product: with an owner, defined consumers, documentation, quality guarantees and a lifecycle.

See also: [Data mesh](#), [Data contract](#)

Data profiling QUALITY

Analysing a dataset's content and structure - value distributions, nulls, patterns, duplicates - to understand its quality and characteristics.

See also: [Data quality](#)

Data quality QUALITY

The degree to which data is fit for its intended use, usually assessed across dimensions such as completeness, validity, accuracy, consistency, uniqueness and timeliness.

See also: [Data quality dimension](#), [Data quality rule](#)

Data quality dimension QUALITY

A measurable aspect of data quality. Common dimensions: completeness, validity, accuracy, consistency, uniqueness, timeliness.

See also: [Data quality](#)

Data quality rule QUALITY

A testable statement that data must satisfy, with a threshold, used to measure quality automatically.

See also: [Data quality](#), [Critical data element \(CDE\)](#)

Data residency PRIVACY & SECURITY

Requirements on the geographic location where data is stored or processed, arising from law, regulation or contract.

See also: [Data sovereignty](#)

Data sovereignty PRIVACY & SECURITY

The principle that data is subject to the laws of the country where it is collected or stored.

See also: [Data residency](#)

Data steward GOVERNANCE

The role responsible day-to-day for the definition, quality and correct use of data in a domain, on behalf of the data owner.

See also: [Data owner](#), [Data stewardship](#)

Data stewardship GOVERNANCE

The operational activities that put governance into practice: maintaining definitions, resolving issues, monitoring quality and supporting users.

See also: [Data steward](#)

Data strategy MANAGEMENT

A plan that links business objectives to the data capabilities, investments and behaviours needed to achieve them.

See also: [Chief data officer \(CDO\)](#)

Data swamp ARCHITECTURE

A data lake that has become unusable because data lacks metadata, ownership and quality control.

See also: [Data lake](#)

Data virtualisation ARCHITECTURE

Querying data across multiple sources through a logical layer without physically copying it.

See also: [Data fabric](#)

Data visualisation ANALYTICS & BI

Representing data graphically so patterns, comparisons and trends can be understood quickly.

See also: [Dashboard](#)

Data warehouse ARCHITECTURE

An integrated, historical store of cleansed data structured for analysis and reporting across subject areas.

See also: [Data mart](#), [Data lake](#)

DataOps ENGINEERING

Applying DevOps and lean practices - automation, testing, version control, monitoring - to data pipelines to improve speed and reliability.

See also: [Data engineering](#), [Data observability](#)

Descriptive analytics ANALYTICS & BI

Analysis that explains what has happened.

See also: [Diagnostic analytics](#), [Predictive analytics](#)

Diagnostic analytics ANALYTICS & BI

Analysis that explains why something happened.

See also: [Descriptive analytics](#)

Dimension table ARCHITECTURE

In a dimensional model, a table of descriptive attributes (who, what, where, when) used to filter and group facts.

See also: [Fact table](#), [Star schema](#)

Dimensional model ARCHITECTURE

A modelling approach that organises analytical data into facts (measurements) and dimensions (context), optimised for querying.

See also: [Star schema](#), [Snowflake schema](#)

Drift (model) AI & ML

Degradation of a model's performance over time because the data it sees (data drift) or the relationship it learned (concept drift) has changed.

See also: [Model monitoring](#)

E

ELT (extract, load, transform) ENGINEERING

Loading raw data into the target platform first and transforming it there, using the platform's compute.

See also: [ETL \(extract, transform, load\)](#)

Embedding AI & ML

A numerical vector representation of text, images or other data that captures meaning so similar items sit close together.

See also: [Vector database](#), [Retrieval-augmented generation \(RAG\)](#)

Entity resolution QUALITY

Determining which records across or within systems refer to the same real-world entity.

See also: [Master data management \(MDM\)](#)

ETL (extract, transform, load) ENGINEERING

Extracting data from sources, transforming it in a separate processing layer, then loading it into the target.

See also: [ELT \(extract, load, transform\)](#), [Data pipeline](#)

Explainability AI & ML

The extent to which the reasons behind a model's output can be understood by people.

See also: [Model governance](#)

F

Fact table ARCHITECTURE

In a dimensional model, a table of measurable events at a defined grain, with foreign keys to dimensions.

See also: [Dimension table](#), [Grain](#)

Fairness AI & ML

The property that an AI system's outcomes do not unjustifiably disadvantage individuals or groups; it must be defined and measured for each use case.

See also: [Bias \(algorithmic\)](#)

Feature (ML) AI & ML

An input variable used by a machine-learning model.

See also: [Feature store](#)

Feature store AI & ML

A system for defining, storing and serving machine-learning features consistently for training and prediction.

See also: [Feature \(ML\)](#)

Federated governance GOVERNANCE

A governance model in which a central function sets standards and policies while business domains own and execute governance for their data.

See also: [Data mesh](#), [Operating model](#)

Foundation model AI & ML

A large model trained on broad data that can be adapted to many tasks, such as a large language model.

See also: [Large language model \(LLM\)](#)

G

Golden record QUALITY

The single, best-available version of an entity's data, assembled from multiple sources through matching and survivorship rules.

See also: [Master data management \(MDM\)](#)

Grain ARCHITECTURE

The level of detail one row represents in a table, for example one row per order line per day.

See also: [Fact table](#)

H

Hallucination AI & ML

A generative AI output that is fluent but factually wrong or unsupported by its sources.

See also: [Retrieval-augmented generation \(RAG\)](#)

I

Idempotency ENGINEERING

The property that running a process more than once produces the same result as running it once - essential for safe pipeline re-runs.

See also: [Data pipeline](#)

Impact analysis GOVERNANCE

Using lineage to determine which downstream reports, processes and users are affected by a change or an issue.

See also: [Data lineage](#)

Incremental load ENGINEERING

Loading only new or changed data since the last run, rather than reloading everything.

See also: [Change data capture \(CDC\)](#)

Ingestion ENGINEERING

Bringing data from source systems into a data platform.

See also: [Data pipeline](#)

K**Key performance indicator (KPI)** ANALYTICS & BI

A measure, tied to a strategic objective, that shows whether performance is on track and prompts action when it is not.

See also: [Metric](#), [Dashboard](#)

L**Lakehouse** ARCHITECTURE

An architecture that combines low-cost data-lake storage in open table formats with warehouse-style management, transactions and performance.

See also: [Data lake](#), [Data warehouse](#), [Open table format](#)

Large language model (LLM) AI & ML

A foundation model trained on large volumes of text to understand and generate language.

See also: [Foundation model](#), [Hallucination](#)

Least privilege PRIVACY & SECURITY

Granting each user or system only the minimum access needed to perform its function.

See also: [Access control](#)

M**Machine learning (ML)** AI & ML

Methods that let systems learn patterns from data to make predictions or decisions without being explicitly programmed for each case.

See also: [Model governance](#), [Training data](#)

Master data GOVERNANCE

Core shared business entities - customers, products, suppliers, locations - used across many processes.

See also: [Reference data](#), [Master data management \(MDM\)](#)

Master data management (MDM) GOVERNANCE

The processes and technology that create and maintain a consistent, authoritative version of master data across systems.

See also: [Golden record](#), [Master data](#)

Metadata GOVERNANCE

Data that describes other data: technical (schemas, types), business (definitions, owners) and operational (run times, volumes, lineage).

See also: [Data catalogue](#)

Metric ANALYTICS & BI

Any quantified measure. Every KPI is a metric; most metrics are not KPIs.

See also: [Key performance indicator \(KPI\)](#)

Metric layer ANALYTICS & BI

A governed layer where business metrics are defined once and reused by every reporting and analytics tool.

See also: [Semantic layer](#)

Micro-batch ENGINEERING

Processing data in very small, frequent batches (seconds to minutes) to approximate real-time.

See also: [Batch processing](#), [Stream processing](#)

MLOps AI & ML

Practices for deploying, monitoring, versioning and governing machine-learning models reliably in production.

See also: [Model monitoring](#), [DataOps](#)

Model governance AI & ML

Policies and controls covering the approval, documentation, validation, monitoring and retirement of analytical and AI models.

See also: [MLOps](#), [Explainability](#)

Model monitoring AI & ML

Ongoing measurement of a deployed model's performance, data inputs and outputs to detect drift and failures.

See also: [Drift \(model\)](#)

N**Normalisation** ARCHITECTURE

Structuring a relational model to reduce redundancy by separating data into related tables.

See also: [Data model](#), [Dimensional model](#)

O**OLAP** ANALYTICS & BI

Online analytical processing: fast, multidimensional analysis of aggregated data, slicing and drilling across dimensions.

See also: [OLTP](#)

OLTP ARCHITECTURE

Online transaction processing: systems optimised for many small, concurrent read/write transactions, such as order entry.

See also: [OLAP](#)

Open table format ARCHITECTURE

A table specification (such as Apache Iceberg, Delta Lake or Apache Hudi) that adds transactions, schema evolution and time travel to files in object storage.

See also: [Lakehouse](#)

Operating model MANAGEMENT

How an organisation arranges structure, roles, decision rights, processes, funding and technology to deliver a capability.

See also: [Federated governance](#)

Orchestration ENGINEERING

Scheduling and coordinating dependent data tasks and pipelines, including retries and alerts.

See also: [Data pipeline](#)

P**Personal data** PRIVACY & SECURITY

Any information relating to an identified or identifiable living individual, directly or indirectly.

See also: [Special category data](#), [Anonymisation](#)

Predictive analytics ANALYTICS & BI

Using statistical and machine-learning models to estimate what is likely to happen.

See also: [Prescriptive analytics](#)

Prescriptive analytics ANALYTICS & BI

Analysis that recommends actions to achieve a desired outcome, often through optimisation or simulation.

See also: [Predictive analytics](#)

Prompt injection AI & ML

An attack in which crafted input causes a generative AI system to ignore its instructions or disclose or act on data inappropriately.

See also: [Large language model \(LLM\)](#)

Pseudonymisation PRIVACY & SECURITY

Replacing identifiers with pseudonyms so data cannot be attributed to a person without separately held additional information. Pseudonymised data is still personal data.

See also: [Anonymisation](#)

Purpose limitation PRIVACY & SECURITY

The principle that personal data is collected for specified, explicit purposes and not further used in incompatible ways.

See also: [Data minimisation](#)

R

Reference data GOVERNANCE

Controlled value sets used to classify other data, such as country codes, currencies or status codes.

See also: [Master data](#)

Retention PRIVACY & SECURITY

The period for which data must or may be kept, after which it is deleted or archived according to policy.

See also: [Data minimisation](#)

Retrieval-augmented generation (RAG) AI & ML

A pattern in which a generative model is given relevant documents retrieved from a knowledge source at query time, to ground its answers.

See also: [Embedding](#), [Hallucination](#)

Reverse ETL ENGINEERING

Moving modelled data from the analytical platform back into operational tools such as CRM or marketing systems.

See also: [ELT \(extract, load, transform\)](#)

Role-based access control (RBAC) PRIVACY & SECURITY

Granting permissions to roles rather than individuals, and assigning people to roles.

See also: [Access control](#)

S

Schema ARCHITECTURE

The defined structure of a dataset: its fields, types and constraints.

See also: [Schema evolution](#)

Schema evolution ENGINEERING

Changing a dataset's structure over time in a controlled way that does not break consumers.

See also: [Data contract](#)

Self-service analytics ANALYTICS & BI

Enabling business users to access governed data and build their own analysis without specialist help for every request.

See also: [Semantic layer](#), [Data literacy](#)

Semantic layer ANALYTICS & BI

A business-friendly abstraction over physical data that defines entities, relationships and metrics consistently for all tools.

See also: [Metric layer](#)

Service level agreement (SLA) MANAGEMENT

An agreed, measurable commitment about a service, such as data freshness by 06:00 on 99% of days.

See also: [Data contract](#)

Slowly changing dimension (SCD) ARCHITECTURE

Techniques for handling changes to dimension attributes over time, such as overwriting (Type 1) or keeping full history (Type 2).

See also: [Dimension table](#)

Snowflake schema ARCHITECTURE

A dimensional model in which dimensions are normalised into multiple related tables.

See also: [Star schema](#)

Special category data PRIVACY & SECURITY

Personal data needing extra protection under data protection law, such as health, biometric, racial or ethnic origin, religious belief or sexual orientation.

See also: [Personal data](#)

Star schema ARCHITECTURE

A dimensional model with a central fact table joined directly to denormalised dimension tables.

See also: [Dimensional model](#), [Snowflake schema](#)

Statistical significance ANALYTICS & BI

A judgement that an observed effect is unlikely to be due to chance alone, given a pre-set threshold.

See also: [A/B test](#)

Stream processing ENGINEERING

Processing data continuously, event by event, as it is generated.

See also: [Batch processing](#)

Synthetic data AI & ML

Artificially generated data that mimics the statistical properties of real data, used for testing or training where real data is restricted.

See also: [Training data](#)

T**Training data** AI & ML

The data used to fit a machine-learning model.

See also: [Bias \(algorithmic\)](#), [Feature \(ML\)](#)

V**Vector database** AI & ML

A database optimised to store embeddings and find the most similar vectors quickly.

See also: [Embedding](#)

Scope of these definitions

Definitions in this glossary are general professional descriptions written for shared understanding. They are not legal definitions and do not replace your organisation's governed business glossary or professional advice.