Data Engineering Glossary

She Ships Data · Free starter kit

How to use this glossary

These are the standard terms you'll hear in data engineering job posts, interviews and team meetings, explained in plain language. Every definition uses official industry vocabulary, so you can search any term and find it used the same way everywhere.

The examples use Ripple, the startup from the show, to make each idea concrete. Terms are grouped by topic, roughly in the order you'll meet them.

Storing data

Term	What it means	Ripple example
Database	An organized system for storing and retrieving data, usually as tables.	The app's database that records every subscription.
Table	Data arranged in rows (records) and columns (fields).	An `orders` table with one row per order.
Schema	The structure of a table or database: column names, data types and how tables relate. Also a named group of tables.	`orders` has `order_id`, `customer_id`, `amount` and `ordered_at`.
OLTP	Online transaction processing: databases built for many small, fast reads and writes from an application.	The checkout system saving each new order.
OLAP	Online analytical processing: systems built for large queries that summarize lots of data.	Calculating revenue by month for the board.

Term	What it means	Ripple example
Data warehouse	A central database built for analytics, holding cleaned, structured data. Examples: Snowflake, BigQuery, Redshift.	Where Ripple's revenue report should live instead of a laptop.
Data lake	Cheap storage that holds raw files of any kind (CSV, JSON, images) at large scale, often in cloud object storage.	Raw exports from every tool, dumped into cloud storage.
Lakehouse	An architecture that adds warehouse features (tables, transactions, governance) on top of data lake storage. Examples: Databricks, Apache Iceberg tables.	One platform for both raw files and clean reporting tables.
Object storage	Cloud storage for files, such as Amazon S3, Azure Blob Storage or Google Cloud Storage.	The bucket where nightly exports land.
CSV	Comma-separated values: a plain-text file format for tables. Easy to read, slow to query and loose about data types.	The spreadsheets Ripple started with.
Parquet	A columnar file format that compresses well and is fast for analytics.	Clean order data saved for the warehouse.
Data type	The kind of value a column holds, such as integer, decimal, text, date or timestamp.	`amount` should be a decimal, not text.

Moving data

Term	What it means	Ripple example
Data pipeline	A series of automated steps that moves data from where it's created to where it's used.	Orders flow from the app to the warehouse to the revenue dashboard.

Term	What it means	Ripple example
Source system	Where data originates, such as an app database, a SaaS tool or an API.	The payments provider and the subscription app.
Ingestion	Bringing data from source systems into your platform.	Pulling yesterday's orders from the app every night.
ETL	Extract, transform, load: data is cleaned and reshaped before it's loaded into the warehouse.	A script cleans the CSVs, then loads them.
ELT	Extract, load, transform: raw data is loaded first, then transformed inside the warehouse. The common modern approach.	Load raw orders as-is, then clean them with SQL.
Batch processing	Processing data in chunks on a schedule, such as hourly or nightly.	A nightly job that refreshes revenue.
Stream processing	Processing data continuously as each event arrives, within seconds. Examples: Kafka, Kinesis.	Alerting the moment a payment fails.
Change data capture (CDC)	Capturing only the rows that were inserted, updated or deleted in a source database, instead of copying everything.	Syncing just today's changed subscriptions.
Full load vs incremental load	A full load copies all data every run; an incremental load copies only new or changed data.	Reloading every order ever vs only today's.
Orchestration	Scheduling and coordinating pipeline steps so they run in the right order and retry on failure. Examples: Airflow, Dagster, Prefect.	Run ingestion, then transforms, then refresh the dashboard.

Term	What it means	Ripple example
DAG	Directed acyclic graph: how orchestrators describe a pipeline, as tasks connected by dependencies with no loops.	"Load orders" must finish before "build revenue" starts.
API	Application programming interface: a way for programs to request data from another system.	Pulling payouts from the payments provider.

Modeling data

Term	What it means	Ripple example
Data modeling	Designing how data is organized into tables and how those tables relate.	Deciding how orders, customers and products connect.
Primary key	A column (or columns) whose value uniquely identifies each row.	`order_id` in the `orders` table.
Foreign key	A column that points to the primary key of another table.	`customer_id` in `orders` points to `customers`.
Grain	What one row in a table represents. Agreeing on it prevents double counting.	One row per order, not one row per item in the order.
Fact table	A table of measurable events, usually with numbers you add up.	`fact_orders`: amount, quantity, discount.
Dimension table	A table of descriptive attributes used to filter and group facts.	`dim_customer`: name, city, signup date.
Star schema	A model with one fact table in the center joined to several dimension tables.	`fact_orders` joined to customer, product and date dimensions.

Term	What it means	Ripple example
Normalization	Splitting data into related tables so each fact is stored once. Common in OLTP databases.	Storing a customer's address once, not on every order.
Denormalization	Combining tables to reduce joins and make queries faster. Common in analytics.	A wide orders table that already includes the customer's city.
Slowly changing dimension (SCD)	A method for tracking changes to dimension values over time. Type 1 overwrites; Type 2 keeps history with a new row.	Keeping a customer's old plan when they upgrade.
Medallion architecture	Organizing a lakehouse into layers: bronze (raw), silver (cleaned) and gold (business-ready).	Raw exports, then clean orders, then the revenue table.

Keeping data reliable

Term	What it means	Ripple example
Data quality	How accurate, complete, consistent and timely data is for its purpose.	Does revenue in the dashboard match the bank?
Data validation	Automated checks that data meets rules before it's used, such as not null, unique or within a range.	Fail the run if any order has a negative amount.
Duplicate records	The same real-world event stored more than once, which inflates counts and totals.	One order appearing twice in revenue.
Deduplication	Removing duplicate records, often by keeping one row per key.	Keep the latest row for each `order_id`.

Term	What it means	Ripple example
Idempotency	A pipeline is idempotent if running it again with the same input gives the same result, with no duplicates.	Rerunning last night's load doesn't double the orders.
Null	A missing or unknown value. Not the same as zero or an empty string.	An order with no `discount` recorded.
Referential integrity	Every foreign key points to a row that exists.	No orders for customers missing from `customers`.
Data freshness	How recent the data is compared with the real world.	The dashboard shows data as of 6 a.m. today.
Backfill	Running a pipeline for past dates to fill in or fix historical data.	Reprocessing all of last year after fixing a bug.
Data lineage	A map of where data came from and every step that changed it.	Tracing a dashboard number back to the raw export.
Observability	Monitoring pipelines and data so problems are caught before users notice.	An alert when today's row count drops by half.
SLA	Service-level agreement: a promise about data, such as when it will be ready.	Revenue is updated by 8 a.m. every weekday.

Tools and performance

Term	What it means	Ripple example
SQL	Structured Query Language: the standard language for querying and transforming tables.	`SELECT SUM(amount) FROM orders;`
Python	The most common general-purpose language in data engineering, used for scripts, APIs and pipeline code.	A script that downloads and cleans exports.

Term	What it means	Ripple example
DuckDB	A fast, free analytics database that runs on your laptop with no server to set up.	Querying the Ripple dataset locally.
Apache Spark	An engine for processing very large datasets in parallel across many machines. Available on Databricks, AWS, Azure and Google Cloud.	Not needed at Ripple's size yet, but common in job posts.
dbt	A tool for writing transformations as SQL files, with testing, documentation and dependencies built in.	Turning raw orders into `fact_orders`.
Apache Airflow	A widely used open-source orchestrator that runs pipelines as DAGs written in Python.	Scheduling the nightly revenue pipeline.
Apache Kafka	A platform for streaming events between systems in real time.	Streaming each new order as it happens.
Table format	A layer that turns files in a data lake into reliable tables with transactions and version history. Examples: Delta Lake, Apache Iceberg.	Making lake files behave like warehouse tables.
Partitioning	Splitting a table into chunks by a column, often date, so queries read only what they need.	Orders partitioned by `order_date`.
Index	A structure that helps a database find rows quickly without scanning the whole table.	Finding one customer's orders fast.
Version control (Git)	A system that tracks every change to your code so you can review, share and undo it. GitHub hosts Git repositories.	The repo holding your pipeline code.

Term	What it means	Ripple example
CI/CD	Continuous integration and delivery: automatically testing and deploying code changes.	Tests run on every pull request before the pipeline changes go live.