Google Cloud

# Course 3 glossary

## Terms and definitions from Course 3

### A

**Algorithm:** A process or set of rules followed for a specific task

**Automated transformation:** The use of processing and scripting tools with less or no programming, compared to manual transformation

### B

**Batch ingestion:** A method of collecting data over time, and then processing it in groups, also called batches

**Batch loading:** A method where data is moved to destination storage in groups called batches at a predetermined schedule

### D

**Data aggregation:** The process of gathering data and expressing it in a summary form

**Data cleaning:** The process of fixing or removing data quality issues

**Data conversion:** The process of changing the format of data to improve compatibility, readability, or make data more secure

**Data derivation:** The process of combining and processing existing data using an algorithm to create new data

**Data enrichment:** The process of adding additional information to data

**Data integrity:** The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

**Data mapping:** The process of matching fields from one data source to another

**Data mapping rules:** A set of instructions that define how the fields will be matched

**Data pipeline:** A series of processes that transports data from different sources to their final destination for storage and analysis

**Data profiling:** The process of exploring data to identify quality issues

**Data standardization:** The process of ensuring that all the data in a dataset is in a common format

**Data transformation:** The process of taking raw data and converting it into a usable format

**Data validation:** The process of checking and rechecking the quality of data so that it's complete, accurate, secure, and consistent

**Deduplication:** The process of eliminating a dataset's redundant data

**Duplicate data:** A record that repeats the information–in whole or in part–of another record

## E

**Exact duplicate:** A record where all the data repeats another record

**Extract-load-transform (ELT):** A type of data pipeline that enables data to be gathered from data lakes, loaded into a unified destination system, and transformed into a useful format

**Extract-transform-load (ETL):** A type of data pipeline that enables data to be gathered from source systems, converted into a useful format, and brought into a data warehouse, or other unified destination system

**Extract:** The stage of retrieving data from one or more sources in a data pipeline

## I

**Incremental loading:** A method where only the data that has changed since the last load is moved to destination storage

## L

**Load:** The stage of inserting data into the target database, data store, data warehouse, or data lake in a data pipeline

## M

**Manual transformation:** The use of coding languages to affect data transformation, without the aid of software programs

**Missing data:** When the rows you expect to be returned do not appear in the joined table

## N

**NULL value:** No available value for a field

## O

**Outer join:** A join that returns both matched and unmatched rows from one or both tables

## P

**Pandas:** An open source data analysis and manipulation tool that can be used with Python

**Partial duplicate:** A record where only part of the data repeats another record

**Python:** A high-level, general-purpose programming language

## R

**R:** A programming language used for statistical computing

## S

**Streaming ingestion:** A method of collecting and processing data as soon as it becomes available

**Streaming loading:** A method where the data is moved to destination storage in a continuous stream, as soon as it becomes available

**Structured Query Language (SQL):** Used to get data out of relational databases

## T

**Time sensitive data:** Data that must be acted on within a specific time frame, or it loses value

**Transform:** The stage of taking data, cleaning it, and putting it into a standard format in a data pipeline

## V

**Validation rules:** A set of instructions that specify the standards that must be met for data to be valid, and how to handle data errors