Glossary terms from module 2

Terms and definitions from Course 3 Module 2

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

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 enrichment: The process of adding additional information to data

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

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

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

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

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

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

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