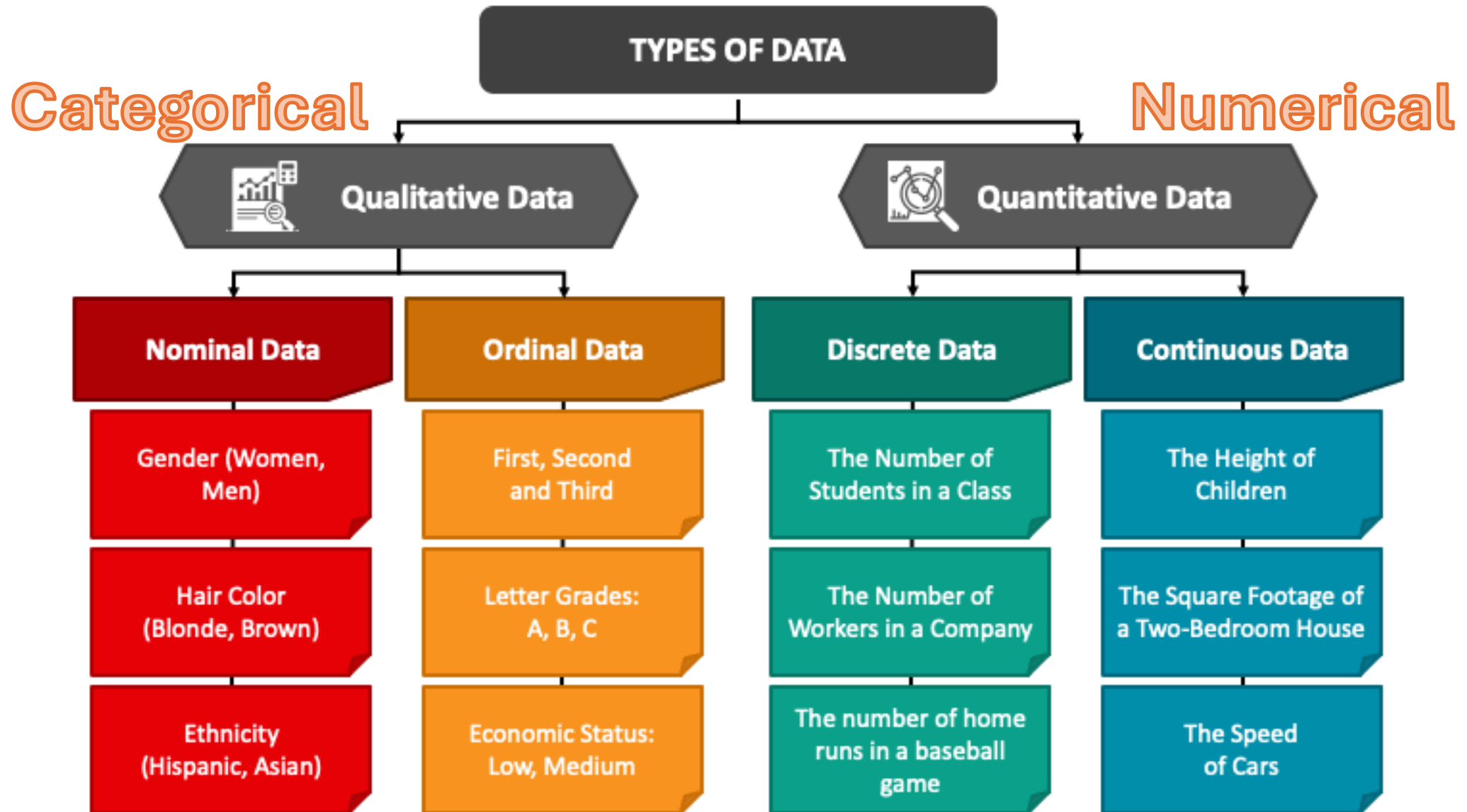




Data and Data Sources



Categorical Data

- The objects being studied are grouped into categories based on some **qualitative** trait.
- The resulting data are merely labels or categories.
- E.g. Hair color (blonde, brown, red, black, etc.), Opinion of students about riots (ticked off, neutral, happy), Smoking status (smoker, non-smoker), etc.
- **Nominal:** A type of categorical data in which objects fall into **unordered** categories.
 - E.g. Hair color (blonde, brown, red, black, etc.), Smoking status (smoker, non-smoker), etc.
- **Ordinal:** A type of categorical data in which **order** is important.
 - E.g. Class (fresh, sophomore, junior, senior, super senior), Opinion of students about riots (ticked off, neutral, happy), etc.
- **Binary:** A type of categorical data in which there are only two categories. Can be nominal or ordinal.
 - E.g. Attendance (present, absent), Smoking status (smoker, non-smoker), etc.

Numerical

- The objects being studied are “measured” based on some **quantitative** trait.
- The resulting data are set of numbers.
- E.g., Cholesterol level, Height, Age, SAT score, Number of students late for class, Time to complete a homework assignment, etc.
- **Discrete:** Only certain values are possible (there are gaps between the possible values).
 - E.g., SAT scores, Number of students late for class, etc.
- **Continuous:** Theoretically, any value within an interval is possible with a fine enough measuring device.
 - E.g., Cholesterol level, Height, Time to complete a homework assignment, etc.

Types of Data

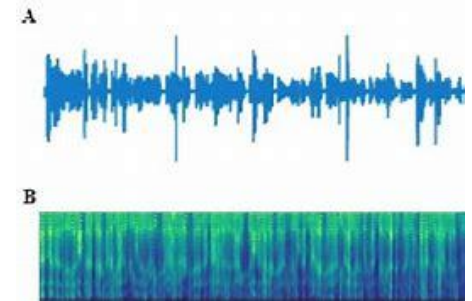
- The type(s) of data collected in a study determine the type of statistical analysis used. For instance:
- Categorical data are commonly summarized using “percentages” (or “proportions”).
 - 11% of students have a tattoo
 - 2%, 33%, 39%, and 26% of the students in class are, respectively, freshmen, sophomores, juniors, and seniors
- Numerical data are typically summarized using “averages” (or “means”).
 - Average number of siblings Fall 1998 Stat 250 students have is 1.9.
 - Average weight of male Fall 1998 Stat 250 students is 173 pounds.
 - Average weight of female Fall 1998 Stat 250 students is 138 pounds.

Types of Data

Based on Structure

- **Structured Data** – Organized, stored in relational databases (e.g., SQL tables).
 - Example: Customer records, transaction logs.
- **Unstructured Data** – No predefined format, difficult to analyze (e.g., text, images, videos).
 - Example: Social media posts, raw audio files.
- **Semi-Structured Data** – Some structure but not strictly formatted (e.g., JSON, XML).
 - Example: Emails, NoSQL databases.

	A	B	C	D	E
1	Cookie Sales by Region				
2	SalesRep	Region	# Orders	Total Sales	
3	Bill	West	217	\$41,107	
4	Frank	West	268	\$72,707	
5	Harry	North	224	\$41,676	
6	Janet	North	286	\$87,858	
7	Joe	South	226	\$45,606	
8	Martha	East	228	\$49,017	
9	Mary	West	234	\$57,967	
10	Ralph	East	267	\$70,702	
11	Sam	East	279	\$77,738	
12	Tom	South	261	\$69,496	
13					
14					
15					



```
{
  "widget": {
    "debug": "on",
    "window": {
      "title": "Sample Konfabulator Widget",
      "name": "main_window",
      "width": 500,
      "height": 500
    },
    "image": {
      "src": "Images/Sun.png",
      "name": "sun1",
      "hOffset": 250,
      "vOffset": 250,
      "alignment": "center"
    },
    "text": {
      "data": "Click Here",
      "size": 36,
      "style": "bold",
      "name": "text1",
      "hOffset": 250,
      "vOffset": 100,
      "alignment": "center",
      "onMouseUp": "sun1.opacity = (sun1.opacity / 100) * 90;"
    }
  }
}
```

Types of Data

Text Data

- **Definition:** Data in the form of natural language text.
- **Examples:** Social media posts, customer reviews, research papers.
- **Use Cases:** Sentiment analysis, language translation, text classification.

Image Data

- **Definition:** Data in the form of images.
- **Examples:** Photographs, medical scans, satellite images.
- **Use Cases:** Image classification, object detection, image segmentation.

Audio Data

- **Definition:** Data in the form of sound recordings.
- **Examples:** Speech recordings, music, environmental sounds.
- **Use Cases:** Speech recognition, audio classification, music generation.

Video Data

- **Definition:** Data in the form of moving images.
- **Examples:** Surveillance footage, video clips, movies.
- **Use Cases:** Action recognition, video summarization, video segmentation.

Types of Datasets

Datasets can be categorized based on their characteristics, structure, and the type of problem they are used to solve. Different types of datasets serve various purposes and are used in different applications. Understanding the nature of the data and choosing the right type of dataset for a specific problem is crucial for developing effective and accurate prediction models.

Here are the primary types of datasets:

1. Structured vs. Unstructured Dataset

Structured Data

- Definition:** Data that is organized in a predefined manner, often in tabular format with rows and columns.
- Examples:** Spreadsheets, SQL databases.
- Use Cases:** Financial records, customer databases, sensor data.

Unstructured Data

- Definition:** Data that does not have a predefined format or organization.
- Examples:** Text documents, images, audio files, videos.
- Use Cases:** Natural language processing (NLP), image recognition, speech-to-text conversion.

Types of Datasets

2. Labeled vs. Unlabeled Datasets

Labeled Data

- Definition:** Data that has been tagged with one or more labels, providing explicit information about the target variable.
- Examples:** Annotated images (with objects labeled), spam vs. non-spam emails.
- Use Cases:** Supervised learning tasks such as classification and regression.

Unlabeled Data

- Definition:** Data without any labels or target variables.
- Examples:** Raw text, unlabeled images, customer behavior data.
- Use Cases:** Unsupervised learning tasks such as clustering, anomaly detection.

3. Time Series Datasets

- Definition:** Data points collected or recorded at specific time intervals.
- Examples:** Stock prices, weather data, sensor readings.
- Use Cases:** Forecasting, anomaly detection, trend analysis.

Types of Datasets

4. Training, Validation, and Test Sets

Training Set

- Definition:** The portion of the dataset used to train the machine learning model.
- Purpose:** To allow the model to learn patterns and relationships in the data.

Validation Set

- Definition:** A subset of the dataset used to tune model parameters and make decisions about model architecture.
- Purpose:** To provide an unbiased evaluation of a model fit on the training dataset while tuning hyperparameters.

Test Set

- Definition:** The portion of the dataset used to evaluate the final model performance.
- Purpose:** To provide an unbiased assessment of the model's performance on unseen data.

Popular sources for datasets

- <https://www.kaggle.com/>
- [Hugging Face – The AI community building the future.](#)
- [Home - UCI Machine Learning Repository](#)
- [Dataset Storage and Dataset Search Platform | IEEE DataPort](#)

Data Quality and Issues

Data quality is crucial for accurate analysis and decision-making. . Poor data quality leads to incorrect insights and business risks.

Key Dimensions of Data Quality

Accuracy – Data should be correct and free from errors.

Completeness – No missing values or gaps in data.

Consistency – Uniform format and values across different sources.

Timeliness – Data should be up to date and relevant.

Validity – Data should conform to predefined formats and rules.

Uniqueness – No duplicate or redundant records.

Data Quality and Issues

Common Data Issues

- **Missing Data** – Incomplete records causing bias in analysis.
- **Duplicate Data** – Multiple records for the same entity.
- **Inconsistencies** – Data mismatch across different systems.
- **Incorrect Data** – Human or system errors in data entry.
- **Data Drift** – Changes in data patterns over time affecting model accuracy.
- **Bias in Data** – Unrepresentative data leading to skewed results.
- **Scalability**: Handling large volumes of data requires scalable storage and processing solutions.

Data Quality and Issues

Improving Data Quality

- **Data Cleaning** – Removing errors, duplicates, and inconsistencies.
- **Data Standardization** – Ensuring a uniform format across datasets.
- **Validation Techniques** – Implementing rule-based validation and automated error detection.
- **Data Governance** – Setting policies for data integrity, security, and compliance.
- **Continuous Monitoring** – Regular data audits and updates to maintain quality.

Association Analysis and Prediction Analysis

Feature	Association Analysis	Prediction Analysis
Goal	To discover relationships and patterns between variables.	To build models that forecast future outcomes or unknown values.
Focus	Identifying "what goes with what."	Predicting "what will happen."
Purpose	Understanding the relationships and dependencies within data.	Maximizing the accuracy of predictions on new data.
Output	Rules, correlations, or patterns describing relationships.	A predictive model that produces forecasts.
Evaluation	Statistical significance of relationships (e.g., support, confidence, lift).	Accuracy metrics (e.g., precision, recall, accuracy, RMSE).
Interpretability	Often high; relationships are typically easier to understand.	Varies; complex models may be "black boxes" with low interpretability.
Example	Market basket analysis (finding which items are frequently bought together).	Predicting customer churn or forecasting sales.
Key Question	"What variables are related?"	"What outcome is most likely?"
Typical Methods	Association rule mining, correlation analysis.	Regression analysis, classification algorithms, time series analysis.



Data and Data Sources

Market Basket analysis



- It identifies associations between products in transactions.
- Uses Association Rule Mining to generate rules like "If a customer buys X, they are likely to buy Y."
- Commonly applied in retail, e-commerce, and recommendation systems.

Association Analysis

Association Analysis is a data mining technique used to **discover relationships or patterns** between items in large datasets. It is widely used in **market basket analysis, recommendation systems, fraud detection, and web usage mining**.

Objective:

To find frequent item-sets and association rules that describe how items are related within a dataset.

Association Rules

Association rules are statements in the form of:

If $X \Rightarrow Y$, Which means, **if item X appears, item Y is also likely to appear.**

Association Analysis

Example:

- **{Bread, Butter} → {Milk}** (People who buy bread and butter often buy milk)
- **{Laptop} → {Mouse}** (People who buy a laptop are likely to buy a mouse)

Key metrics used to evaluate association rules:

1. Support

Measures how frequently an itemset appears in the dataset.

$$\text{Support}(X) = \frac{\text{Frequency of } X \text{ in dataset}}{\text{Total transactions}}$$

Association Analysis

Example: If Milk appears in 30 out of 100 transactions, then:

$$\text{Support}(\text{Milk}) = \frac{30}{100} = 30\%$$

2. Confidence

Measures how often **Y** appears when **X** is present.

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

Example: If Bread appears in 50 transactions, and in 40 of them, Milk is also bought:

$$\text{Confidence}(\text{Bread} \Rightarrow \text{Milk}) = \frac{40}{50} = 80\%$$

Association Analysis

3. Lift

Measures how much **stronger** the association is compared to a random occurrence.

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)}$$

If $\text{Lift} > 1$: X and Y are positively correlated (buying one increases the likelihood of buying the other).

If $\text{Lift} < 1$: X and Y are negatively correlated (buying one reduces the likelihood of buying the other).

Problems

A retailer wants to analyze buying patterns based on 500 transactions in a week:

- {Laptop} appears in 100 transactions.
- {Laptop, Mouse} together appear in 60 transactions.
- {Mouse} appears in 150 transactions.

Questions:

1. What is the confidence of the rule {Laptop} \rightarrow {Mouse}?
2. What is the confidence of the rule {Mouse} \rightarrow {Laptop}?

Problems: Supermarket Transactions

Transaction Dataset

Transaction ID	Items Purchased
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Bread, Butter, Eggs
T5	Bread, Butter, Eggs

Step 1: Compute Support

- $\text{Support}(\text{Milk})$
- $\text{Support}(\text{Bread})$
- $\text{Support}(\text{Butter})$
- $\text{Support}(\{\text{Milk}, \text{Bread}\})$
- $\text{Support}(\{\text{Bread}, \text{Butter}\})$

Step 2: Compute Confidence

- $\text{Confidence}(\text{Milk} \rightarrow \text{Bread})$
- $\text{Confidence}(\text{Bread} \rightarrow \text{Butter})$

Step 3: Compute Lift

- $\text{Lift}(\text{Milk} \rightarrow \text{Bread})$
- $\text{Lift}(\text{Bread} \rightarrow \text{Butter})$

Problems

Transaction Data

Transaction ID	Items Purchased
T1	Apple, Banana, Milk
T2	Apple, Banana
T3	Apple, Banana, Milk
T4	Banana, Milk, Bread
T5	Apple, Bread
T6	Banana, Bread
T7	Apple, Banana, Bread

Lift(Apple → Banana) 0.87

Lift(Banana → Bread) 0.60

Applications of Market Basket analysis

Retail:

- Optimize product placement (**e.g., placing Milk near Bread**).
- Identify frequently bought-together items for promotions.

E-commerce & Recommendations:

- Suggest items frequently bought together (**Amazon's "Customers who bought this also bought..."**).
- Improve personalized recommendations.

Healthcare: Analyze patient symptoms and medications that are frequently prescribed together.

Finance: Detect fraud by identifying unusual spending patterns.

Practice

Q. Using the following transactional dataset of customer purchases. Find:

- i. Frequent Itemset/s
- ii. Association rules
- iii. Support, confidence and lift of the rules

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Practice

Frequent Itemsets (Let's use a minimum support of 2):

•Individual Items:

- Bread: $5/8 = 0.625$ (Support = 0.625)
- Milk: $6/8 = 0.75$ (Support = 0.75)
- Eggs: $3/8 = 0.375$ (Support = 0.375)
- Butter: $4/8 = 0.5$ (Support = 0.5)
- Cheese: $1/8 = 0.125$ (Support = 0.125)

•Pairs:

- {Bread, Milk}: $3/8 = 0.375$ (Support = 0.375)
- {Bread, Butter}: $3/8 = 0.375$ (Support = 0.375)
- {Milk, Butter}: $3/8 = 0.375$ (Support = 0.375)
- {Milk, Eggs}: $2/8 = 0.25$ (Support = 0.25)
- {Bread, Eggs}: $2/8 = 0.25$ (Support = 0.25)

•Triplets:

- {Bread, Milk, Butter}: $2/8 = 0.25$ (Support = 0.25)

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Practice

Support, Confidence, and Lift:

- Support:** The proportion of transactions that contain the itemset.
- Confidence:** The probability that a transaction containing A also contains B ($A \rightarrow B$).
- Lift:** The ratio of the observed support to the support if A and B were independent. A lift greater than 1 suggests a positive association.

Association Rules (Using the frequent itemsets):

•{Bread, Milk} \rightarrow {Butter}:

- Support = $2/8 = 0.25$
- Confidence = $2/3 = 0.666$
- Lift = $(2/8) / ((3/8) * (4/8)) = 1.33$

•{Bread, Butter} \rightarrow {Milk}:

- Support = $2/8 = 0.25$
- Confidence = $2/3 = 0.666$
- Lift = $(2/8) / ((3/8) * (6/8)) = 0.888$

•{Milk, Butter} \rightarrow {Bread}:

- Support = $2/8 = 0.25$
- Confidence = $2/3 = 0.666$
- Lift = $(2/8) / ((3/8) * (5/8)) = 1.066$

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Practice

Transaction ID	Items Purchased
1	Bread, Milk, Eggs
2	Bread, Butter
3	Milk, Butter
4	Bread, Milk, Butter, Cheese
5	Eggs, Milk
6	Bread, Eggs
7	Milk
8	Bread, Butter, Milk

Rule	Support	Confidence	Lift
{Bread} -> {Milk}	0.375	0.6	0.8
{Bread} -> {Butter}	0.375	0.6	1.2
{Milk} -> {Bread}	0.375	0.5	0.8
{Milk} -> {Butter}	0.375	0.5	1
{Butter} -> {Bread}	0.375	0.75	1.5
{Butter} -> {Milk}	0.375	0.75	1.25
{Bread, Milk} -> {Butter}	0.25	0.666	1.33
{Bread, Butter} -> {Milk}	0.25	0.666	0.888
{Milk, Butter} -> {Bread}	0.25	0.666	1.066



Data and Data Sources

Data Catalogue

A data catalog is essentially an organized inventory of an organization's data assets. It uses metadata (data about data) to make it easier for users to find, understand, and use data.

Features:

- **Metadata Management:** Storing information about data sources, schemas, data lineage, and other relevant details.
- **Data Discovery:** Enabling users to search and find relevant data assets.
- **Data Lineage:** Tracking the origin and flow of data through various systems.
- **Data Governance:** Enforcing policies and controls related to data usage and access.

[what-is-a-data-catalog/](#)

Data Catalogue and Data Pipelines

A data catalog provides the "what" and "where" of data, while data pipelines handle the "how" of moving and transforming it.

- A data catalog enhances data pipelines by providing visibility into the data's origin, transformations, and quality.
- Data pipelines populate the data catalog with metadata as data moves through the stages.
- Data catalogs help data engineers and analysts understand the impact of changes to data pipelines.

Data Pipelines

A **data pipeline** is a set of processes that automate the movement, transformation, and processing of data from source to destination.

It ensures data is collected, cleaned, enriched, and stored efficiently for analysis or machine learning.

- [guide-to-data-pipelines/](#)

Key Components of a Data Pipeline

A data pipeline is a series of processes that move and transform data from one or more sources to a destination. Common stages include:

Extraction/Collection: Retrieving data from various sources, such as databases, APIs, files, or streaming platforms.

Ingestion: Bringing the extracted data into a staging area. Batch data ingestion and streaming data ingestion.

Storage: Storing data in a data warehouse, data lake, or a database. In case of ETL this step comes after data preparation.

Key Components of a Data Pipeline

Data Preparation/Transformation:

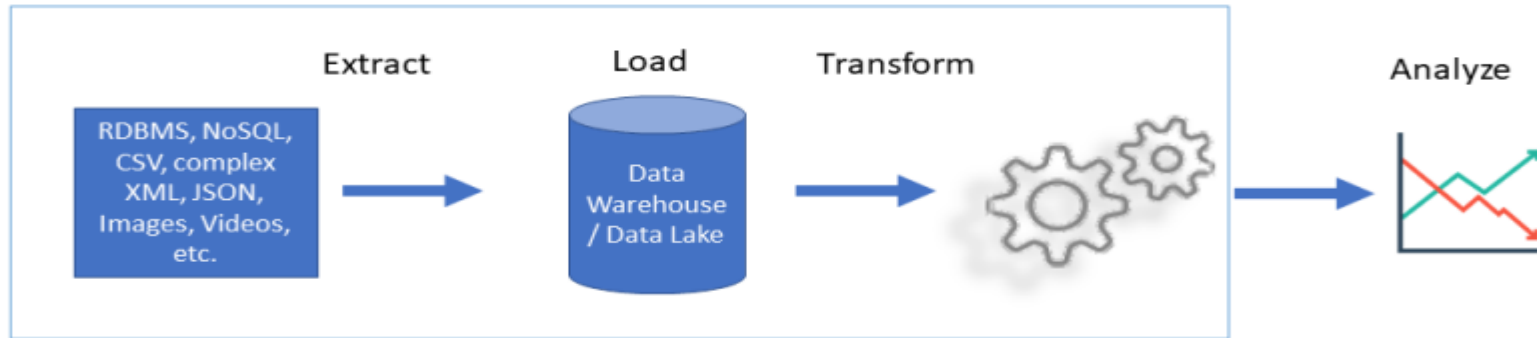
- **Cleaning:** Identifying and correcting errors, inconsistencies, and missing values in the data.
- **Wrangling:** Transforming and structuring the data to make it suitable for analysis or other purposes.
This may involve: Filtering, Aggregating, Joining, Formatting.
- **Exploration and data analysis:** This is where data analysts begin to look at the data, to find patterns, and to understand the data that has been gathered. Querying, reporting, or using data for machine learning.

Versioning & Monitoring : Maintaining a history of data changes, allowing for tracking and rollback if necessary. This is very important for data governance, and for reproducibility of analysis. Managing dependencies, scheduling tasks, and ensuring system reliability.

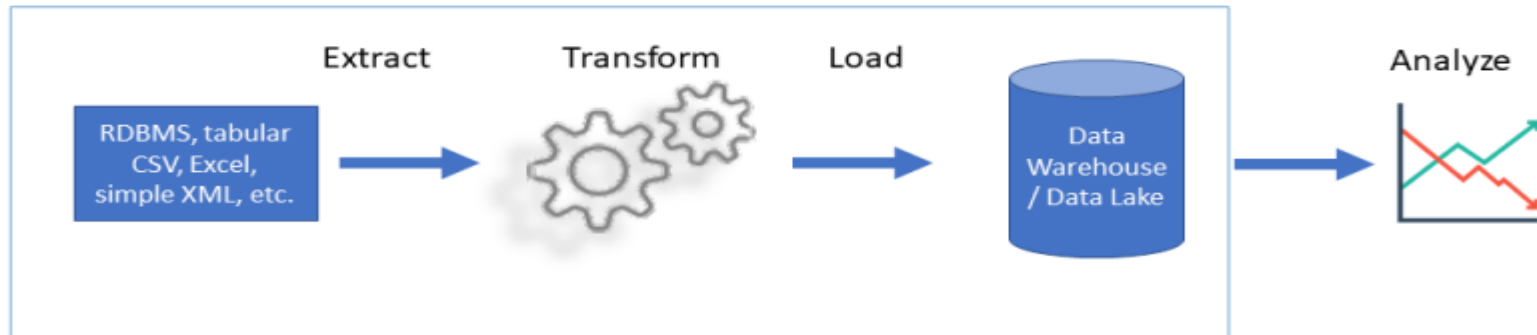
Stages of a Data Pipeline

ELT vs ETL

ELT



ETL



Common Data Pipeline Patterns

Batch Processing Pipeline

- Processes data in chunks at scheduled intervals.
- Suitable for large-scale ETL workloads.
- **Example:** Nightly aggregation of customer transactions for financial reporting.

Technology Stack:

- ◆ Apache Spark, Apache Hadoop, Airflow, AWS Glue

Streaming Data Pipeline

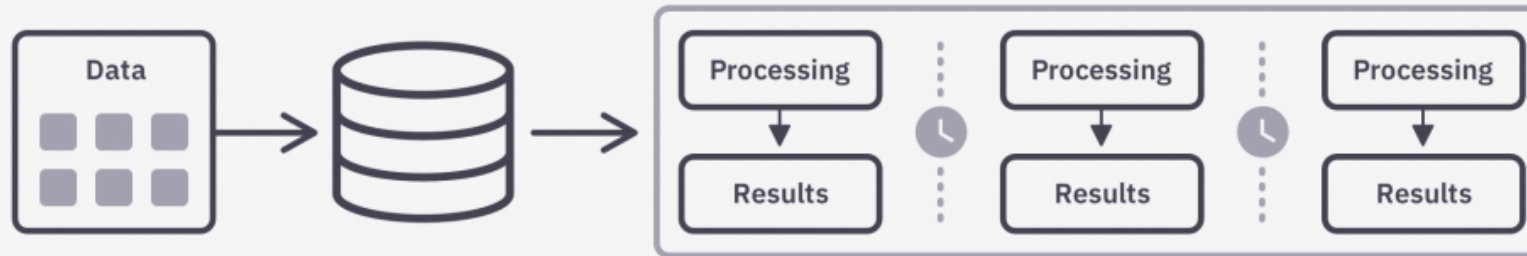
- Processes data in real-time or near real-time.
- Suitable for applications like fraud detection, live analytics, and IoT.
- **Example:** Monitoring website clicks or detecting fraudulent credit card transactions.

Technology Stack:

- ◆ Apache Kafka, Apache Flink, Spark Streaming, AWS Kinesis

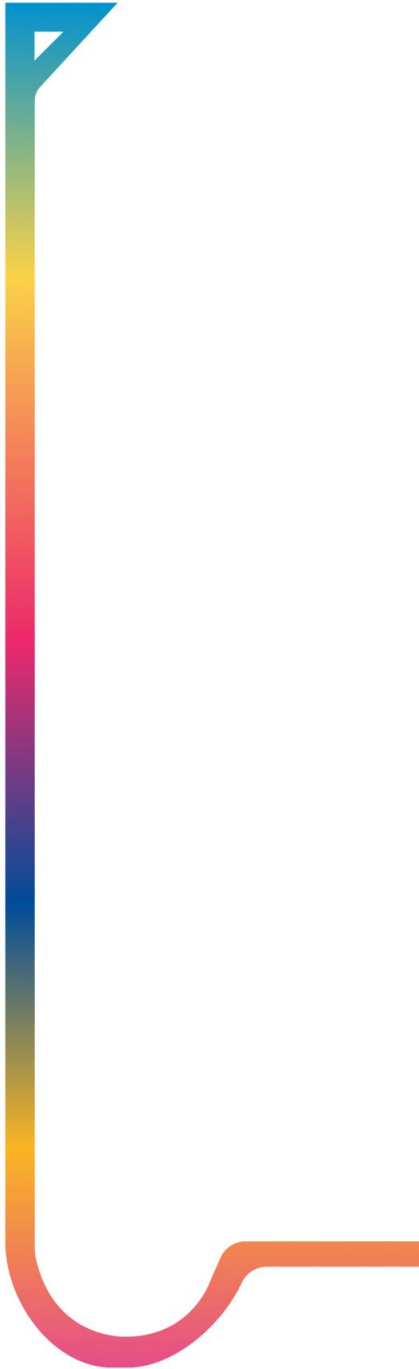
Common Data Pipeline Patterns

Batch Processing



Data Stream Processing





Feature	Batch Processing	Stream Processing
Data Processing	Processes a large volume of data at once.	Processes data as it arrives, record by record.
Latency	High latency, as processing happens after data collection.	Low latency, providing near real-time insights.
Throughput	Can handle vast amounts of data at once.	Optimized for real-time but might handle less data volume at a given time.
Use Case	Ideal for historical analysis or large-scale data transformations.	Best for real-time analytics, monitoring, and alerts.
Complexity	Relatively simpler to implement with predefined datasets.	More complex, requires handling continuous streams.
Data Scope	Operates on a finite set of data.	Operates on potentially infinite streams of data.
Error Handling	Errors can be identified and corrected before execution.	Requires real-time handling of errors and failures.
Resource Usage	Resource-intensive during processing, idle otherwise.	Continuous use of resources.
Cost	Cost-effective for large volumes of data.	More expensive due to continuous processing.

Common Data Pipeline Patterns

Lambda Architecture (Hybrid Batch + Stream)

- Combines batch and real-time processing.
- **Example:** A weather app that uses real-time sensor data for short-term forecasts and batch data for long-term trends.

Layers:

- **Batch Layer:** Stores historical data.
- **Speed Layer:** Processes real-time data.
- **Serving Layer:** Merges both for a unified view.

Technology Stack:

- ◆ Apache Kafka, Apache Spark, HDFS, NoSQL Databases

Common Data Pipeline Patterns

Data Lake + Data Warehouse Hybrid

- Stores **raw data** in a **data lake** (e.g., AWS S3, Azure Data Lake).
- Transforms and moves structured data into a **data warehouse** (e.g., Snowflake, Redshift).

Example: An e-commerce company storing all transactions in a data lake but using a warehouse for analytics.

Technology Stack:

- ◆ AWS S3, Azure Data Lake, Snowflake, BigQuery

Best Practices for Data Pipelines

- ✓ **Use a Scalable Architecture** – Design for growing data volume.
- ✓ **Ensure Data Quality** – Use validation and anomaly detection.
- ✓ **Automate Orchestration** – Schedule and monitor pipelines with Apache Airflow.
- ✓ **Optimize Performance** – Use caching, indexing, and parallel processing.
- ✓ **Implement Security & Governance** – Encrypt data, use access controls, and comply with GDPR.

Data Transformation

Need of data transformation:

- To ensure data is consistent and compatible across different systems.
- To improve data quality by cleaning and standardizing it.
- To make data more suitable for specific analytical or modeling needs.

Data transformation includes:

1. Data cleaning:

- Removing or correcting errors, inconsistencies, and duplicates.
- Handling missing values (e.g., imputation).

2. Standardization:

- Formatting data consistently (e.g., date formats, units of measure).
- Normalizing or scaling numerical data.

Data Transformation

3. Structuring:

- Changing the data's organization (e.g., pivoting, aggregating).
- Converting data types (e.g., string to integer).

4. **Enrichment:** Adding new data or deriving new values from existing data. Joining data from multiple sources.

5. **Filtering:** Removing unwanted data.

6. **Aggregation:** Summarizing data.

Feature Management

Feature Selection

Selecting the most relevant features from a dataset while eliminating redundant or irrelevant ones.

- **Methods:**

- **Filter Methods** (e.g., Correlation, Mutual Information)
- **Wrapper Methods** (e.g., Recursive Feature Elimination)
- **Embedded Methods** (e.g., Lasso Regression)

Feature Engineering

Creating new features from raw data to enhance model learning.

- **Common Techniques:**

- **Polynomial Features** (e.g., x^2, x^3, x^2x^3)
- **Domain-Specific Transformations** (e.g., Date-Time Feature Extraction)
- **Aggregations and Grouping** (e.g., Mean Purchase Amount per User)

Feature Management

Feature Transformation

Modifying features to meet the assumptions of machine learning algorithms.

•Techniques:

- **Scaling** (Standardization, Min-Max Normalization)
- **Encoding** (One-Hot Encoding, Label Encoding)
- **Log Transformations** (for skewed data)

Feature Store & Feature Versioning

Managing and reusing features efficiently in ML pipelines.

- **Feature Store Tools:** Tecton, Feast, AWS SageMaker Feature Store
- **Feature Versioning:** Tracking feature changes across different ML models.