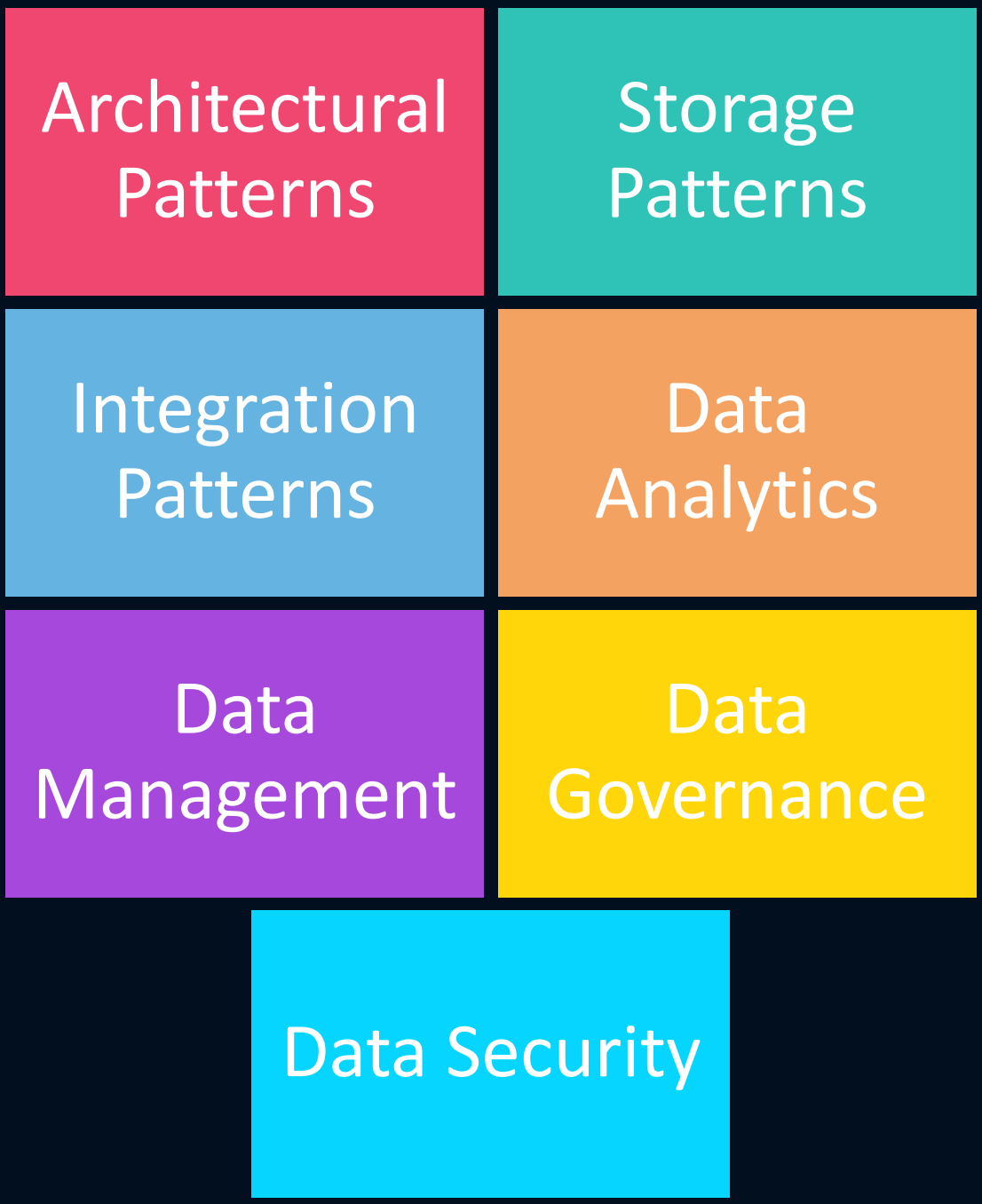
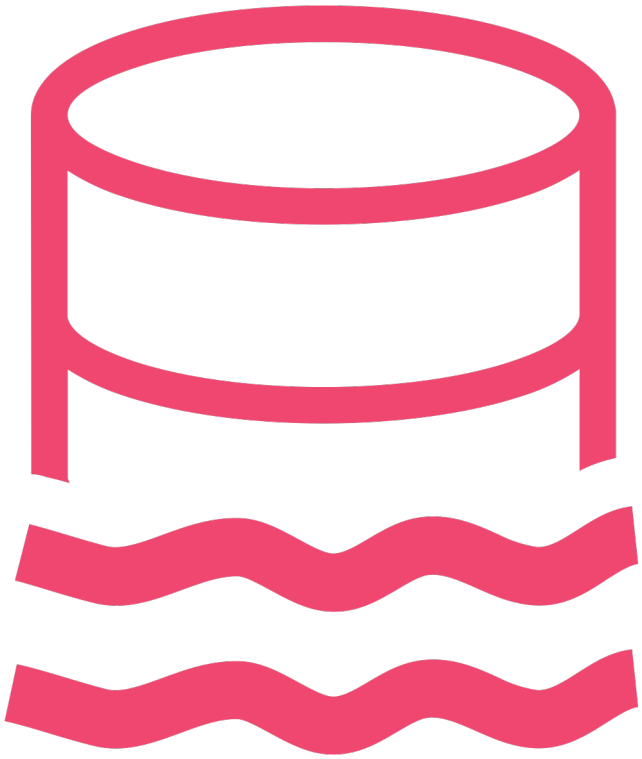
Data Patterns and Practices Library



Architectural Patterns

- Data Lake
- Data Warehouse
- Lambda Architecture
- Kappa Architecture
- Microservices Architecture
- Event-Driven Architecture
- Polyglot Persistence
- Data Mesh
- Data Vault
- Streaming-first Architecture

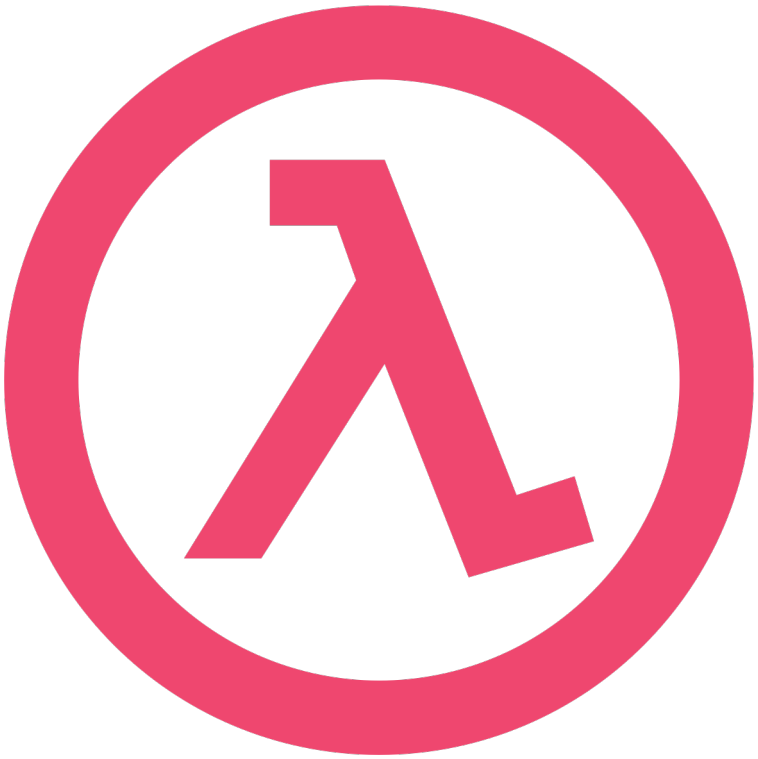


Data Lake

A centralised repository that allows storing structured and unstructured data at any scale. It enables raw data storage for various analytics purposes.

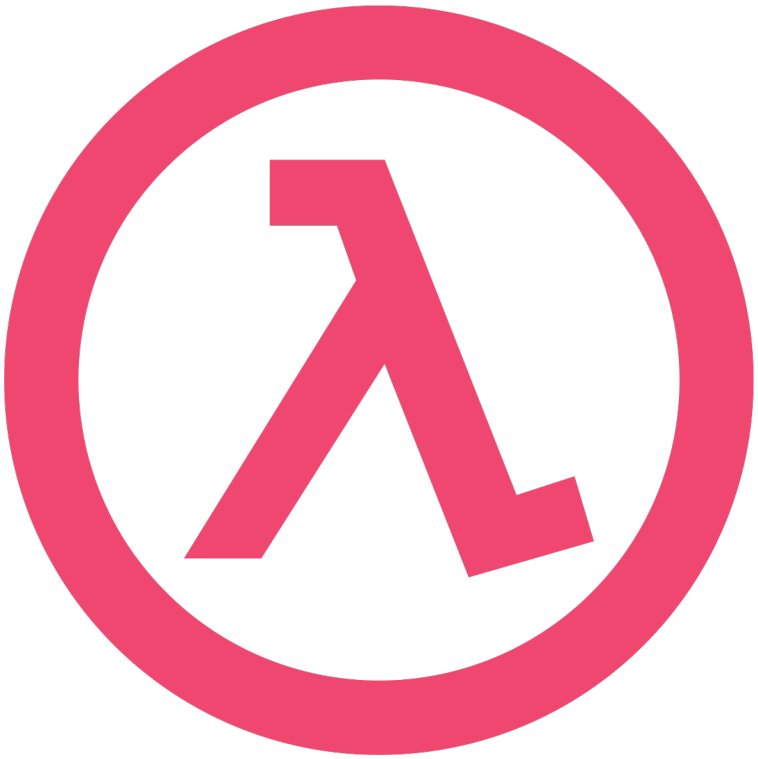
Data Warehouse

A large, centralised repository for storing and managing structured data, optimised for high-performance analytics and reporting.



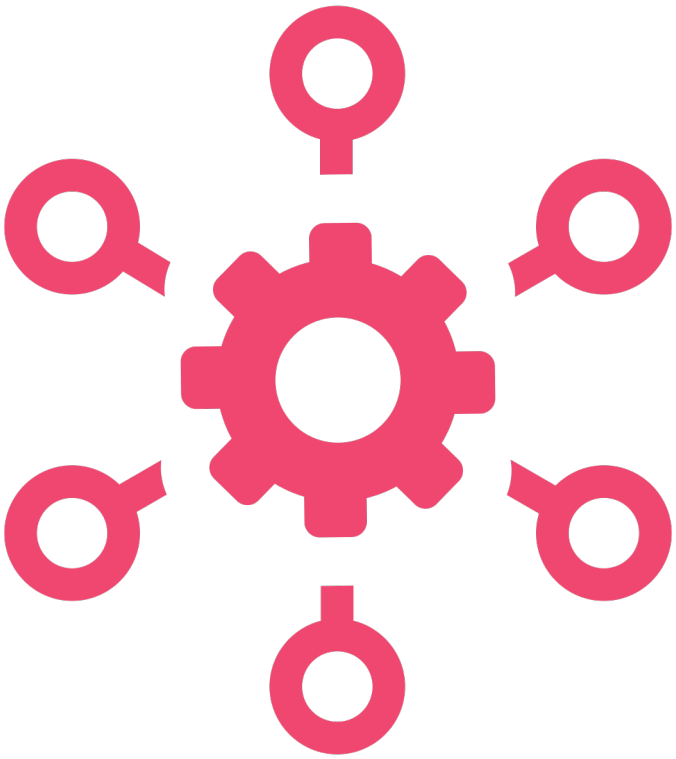
Lambda Architecture

A data processing architecture that combines batch and stream processing for fault-tolerant, scalable, and real-time data analytics.



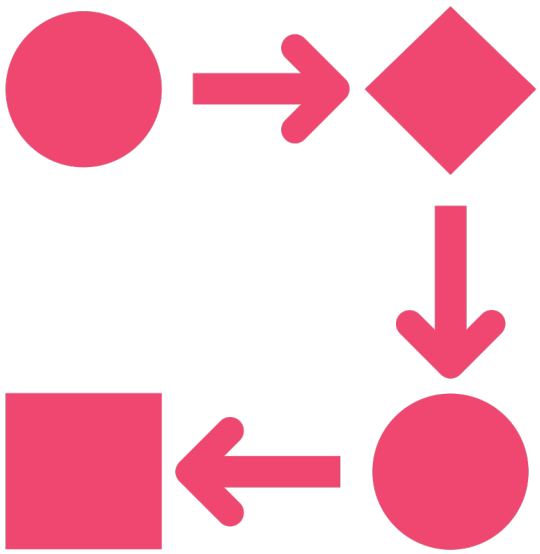
Kappa Architecture

A data processing architecture that simplifies Lambda Architecture by only using stream processing for both real-time and historical data.



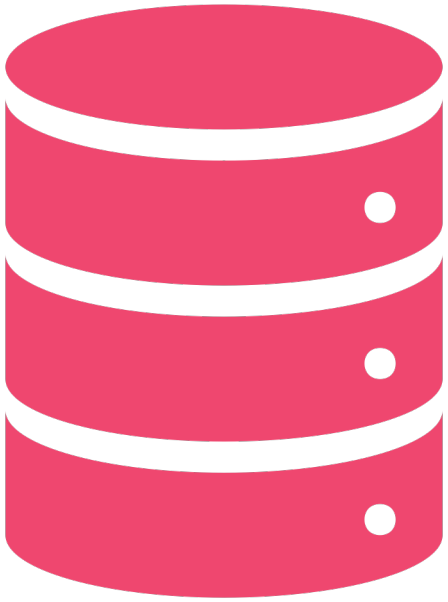
Microservices Architecture

A design approach that structures applications as a collection of small, independently deployable services, allowing for greater flexibility and scalability.



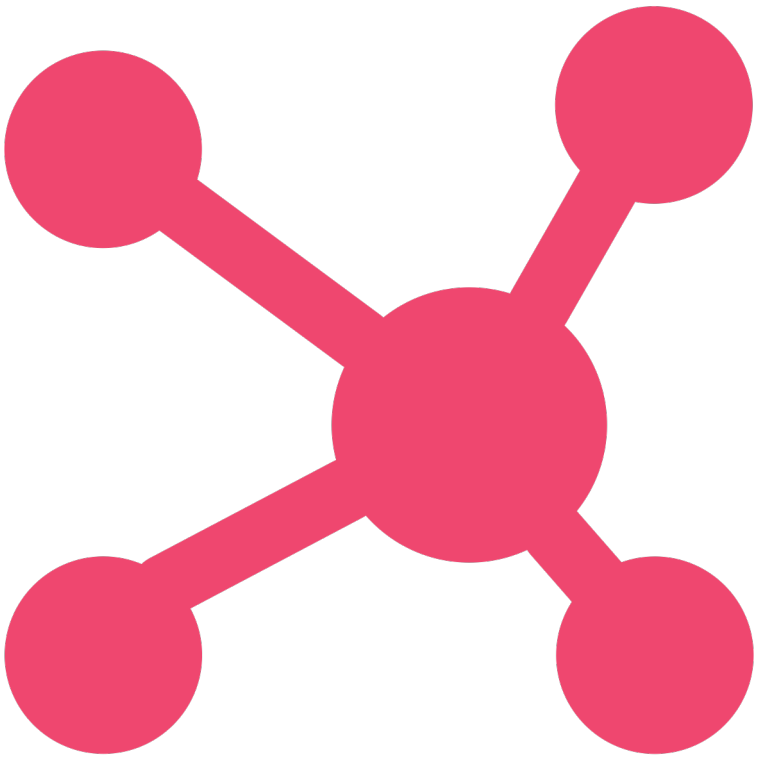
Event-driven Architecture

A software design pattern that promotes the production, detection, and reaction to events, enabling loose coupling and high scalability in distributed systems.



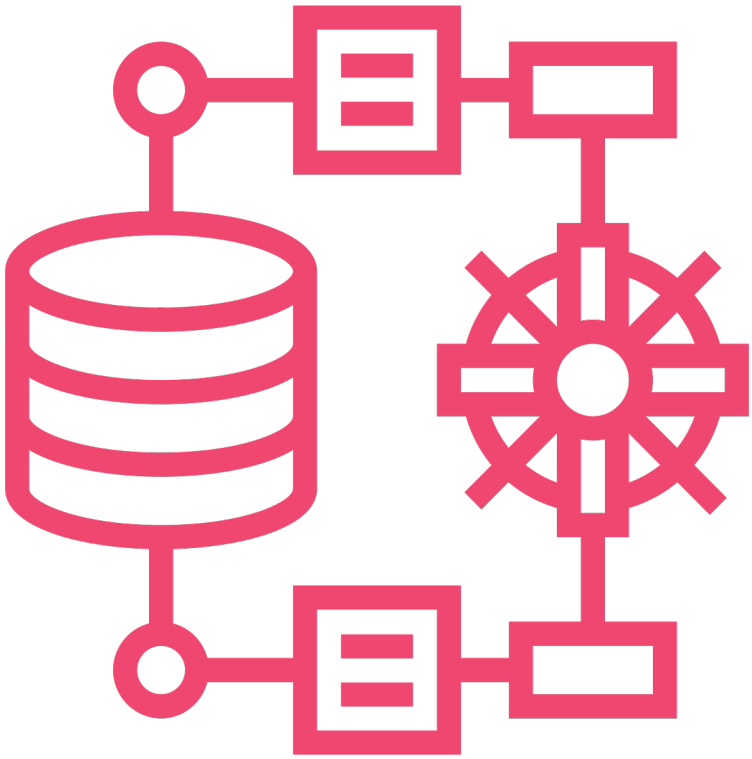
Polyglot Persistence Architecture

A data storage strategy that uses multiple types of databases to store and manage data according to its specific needs, optimizing performance, and reducing complexity.



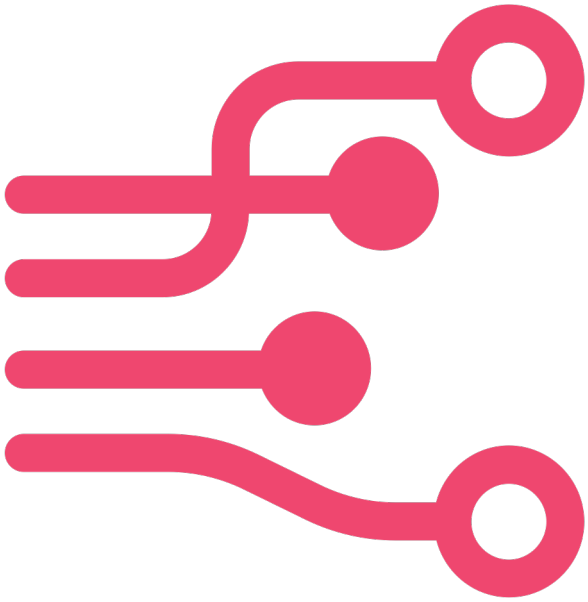
Data Mesh

A decentralized approach to data architecture and data platform thinking, focusing on domain-oriented data ownership, self-serve data infrastructure, and product-oriented data delivery.



Data Vault

A hybrid data modeling and storage methodology that combines aspects of 3NF and star schema to create a scalable, flexible, and auditable data warehousing solution.

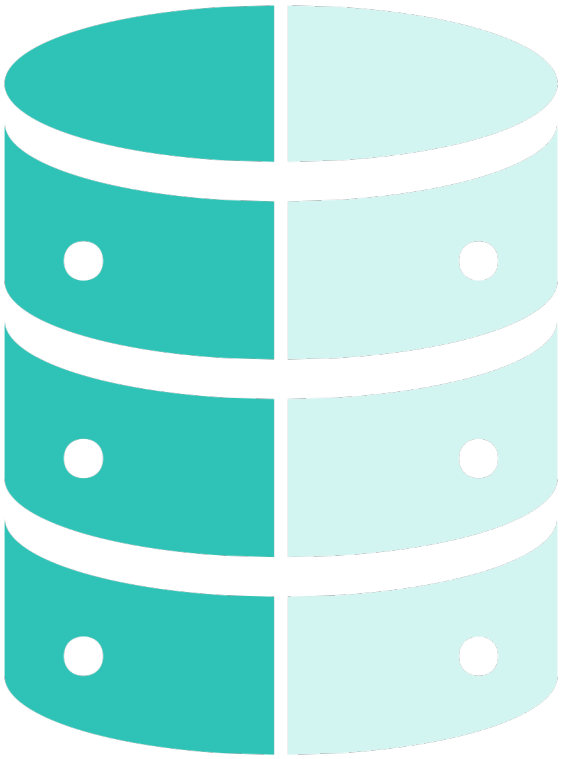


Streaming-first

An approach that prioritizes real-time data processing and analysis by utilizing event streaming technologies and stream processing engines for handling data.

Storage Patterns

Sharding	Partitioning	Replication
Federated Storage	Object Storage	Columnar Storage
Time-series Storage	Graph Storage	In-Memory Storage
	Hybrid Storage	

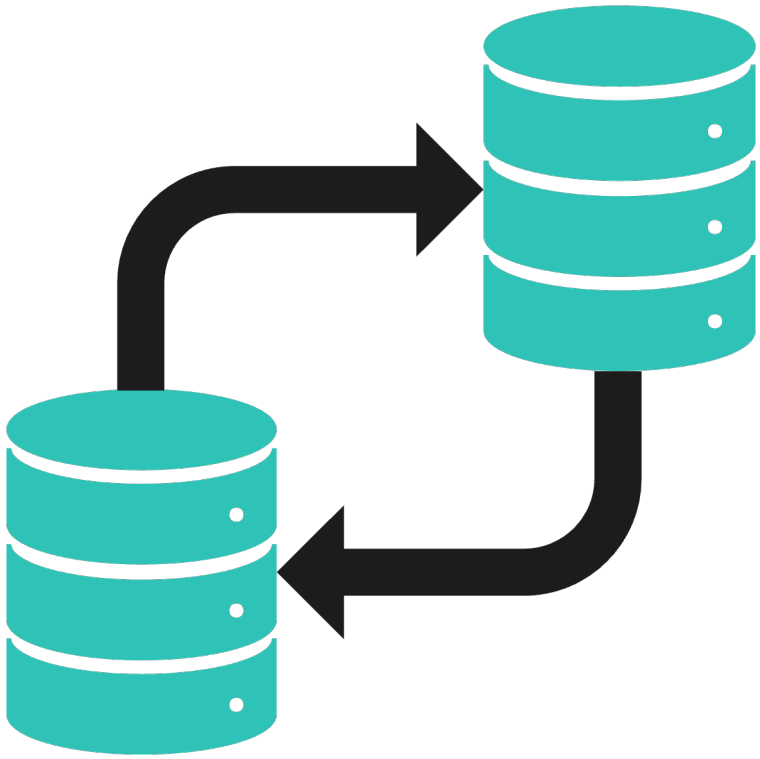


Sharding

A method of distributing data across multiple database servers to improve performance and scalability.

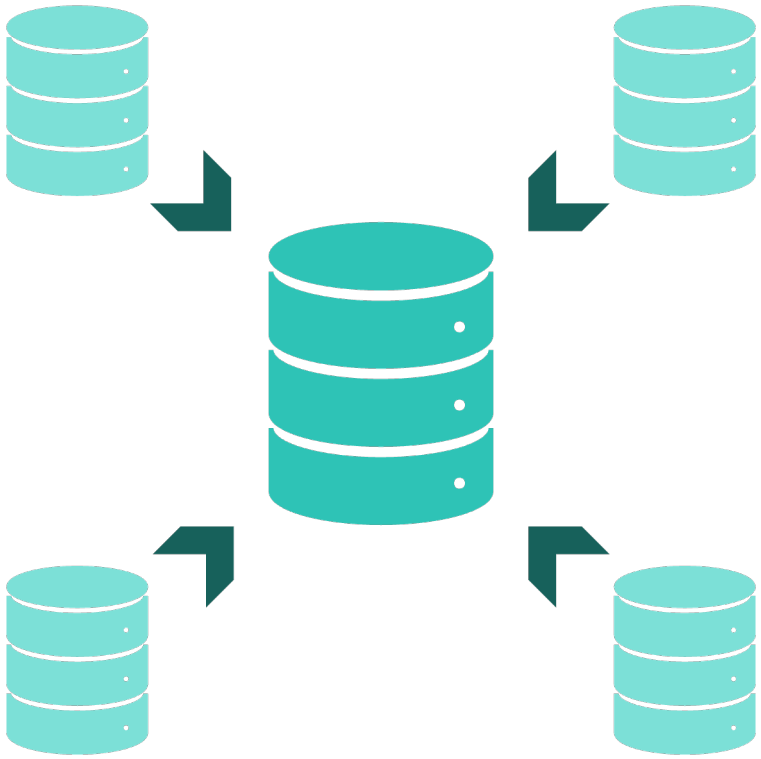
Partitioning

The process of dividing a large table into smaller, more manageable pieces to improve query performance and maintainability.



Replication

The process of copying data from one database to another to ensure data availability, redundancy, and load balancing.

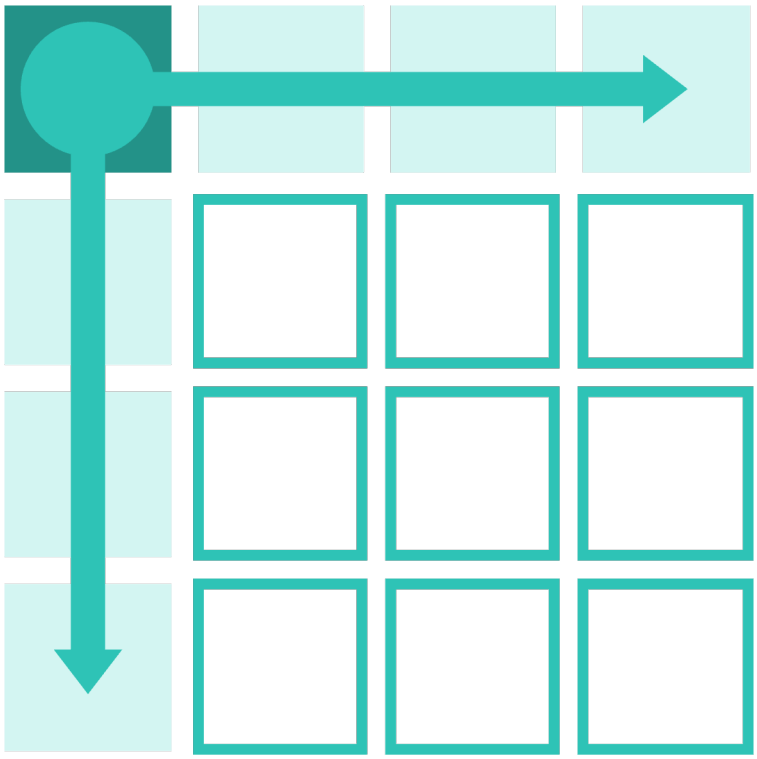


Federated Storage

A storage architecture that integrates multiple storage systems under a unified management framework, enabling seamless data sharing and access across systems.

Object Storage

A scalable storage architecture that manages data as objects, rather than as files or blocks, providing high performance and reliability for large-scale unstructured data storage.

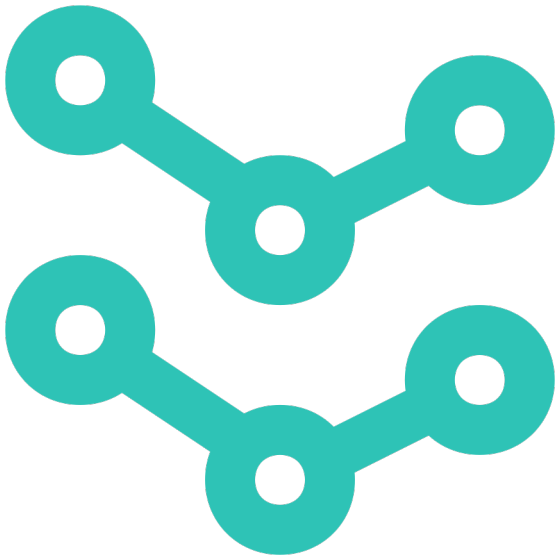


Columnar Storage

A storage format that stores data by column, rather than by row, which is particularly suited for analytics and data warehousing workloads.

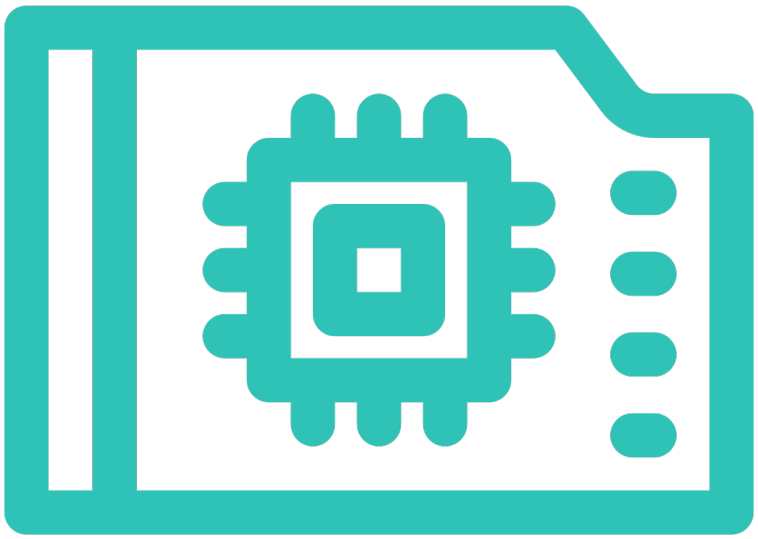
Time-series

A specialized storage system designed to handle time-stamped data, such as sensor data, stock prices, or logs, efficiently and effectively.



Graph Storage

A storage system optimized for storing and querying graph data, which represents entities and their relationships in a highly interconnected structure.



In-memory Storage

A storage architecture that stores data in the main memory (RAM) instead of disk storage, providing significantly faster access and processing speeds.

Hybrid Storage

A storage solution that combines different storage types, such as on-premises and cloud storage, to optimize cost, performance, and flexibility.

Integration Patterns

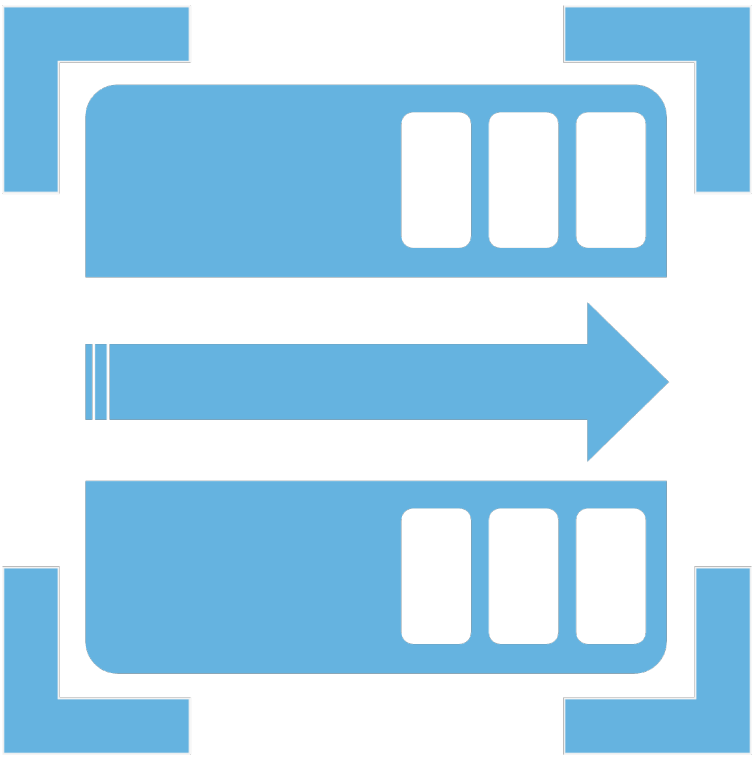
Extract, Transform, Load (ETL)	Extract, Load, Transform (ELT)	Change Data Capture (CDC)
Data Federation	Data Virtualization	Data Replication
Data Synchronization	Data Preparation	Publish/Subscribe Pattern
	Request/Reply Pattern	

Extract, Transform, Load (ETL)

A process of extracting data from source systems, transforming it into a desired format or structure, and loading it into a target system, such as a data warehouse.

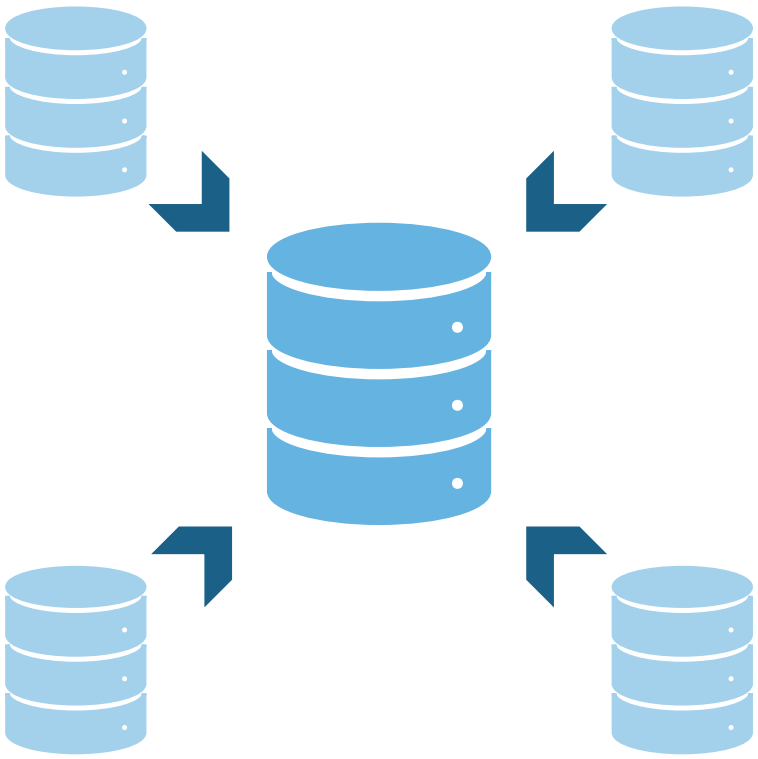
Extract, Load, Transform (ELT)

A variation of ETL, where data is first loaded into the target system and then transformed, leveraging the processing capabilities of modern data storage systems.



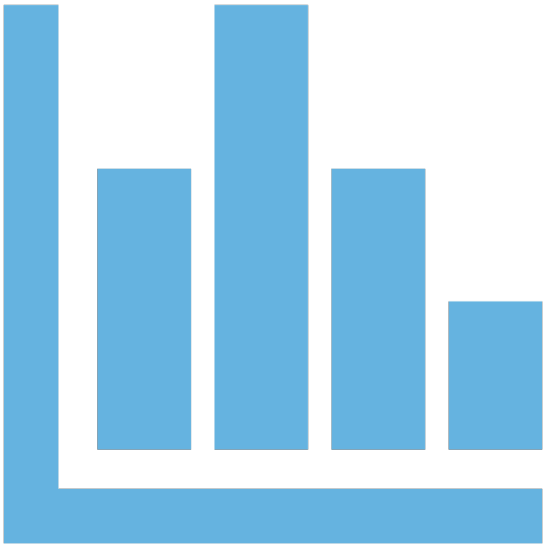
Change Data Capture (CDC)

A technique for capturing and processing changes in source data, enabling incremental updates to target systems, and reducing the need for batch data transfers.



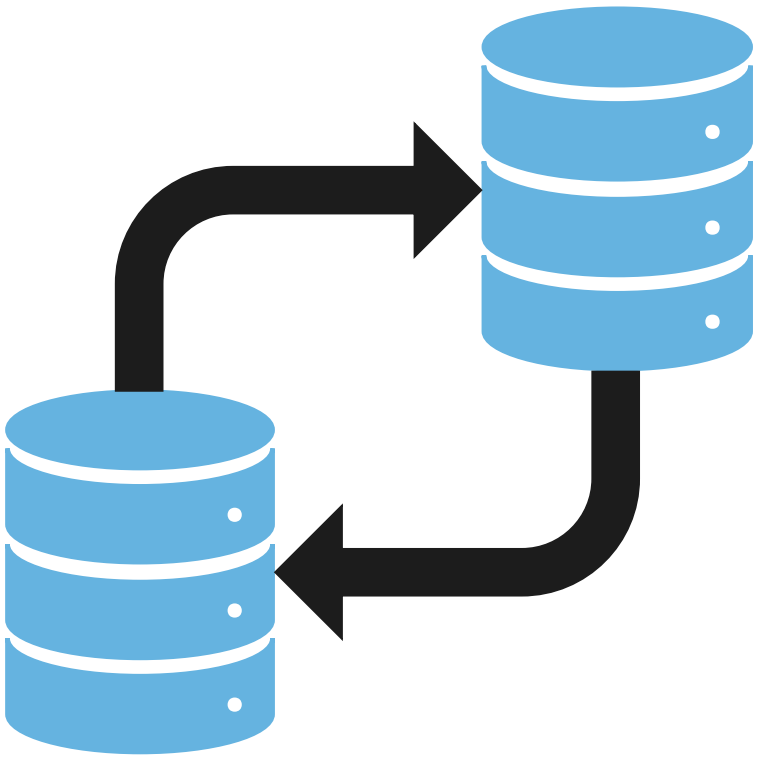
Data Federation

A technique for integrating data from disparate sources without physically moving or copying the data, providing a unified view of data across systems.



Data Visualization

An approach that abstracts underlying data sources, enabling users and applications to access and manipulate data without needing to know its physical location or structure.

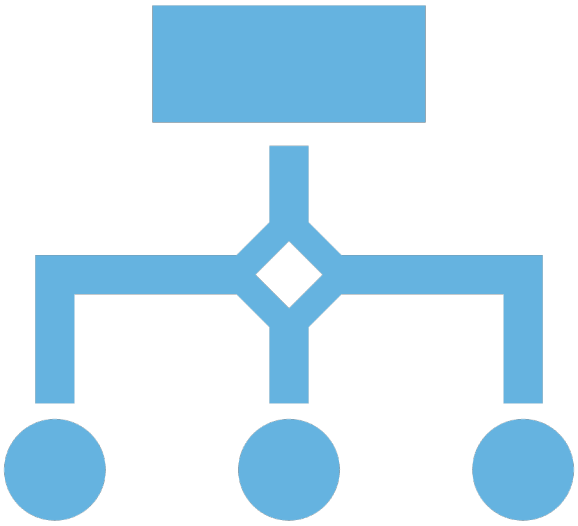


Data Replication

The process of copying data from one database to another to ensure data availability, redundancy, and load balancing.

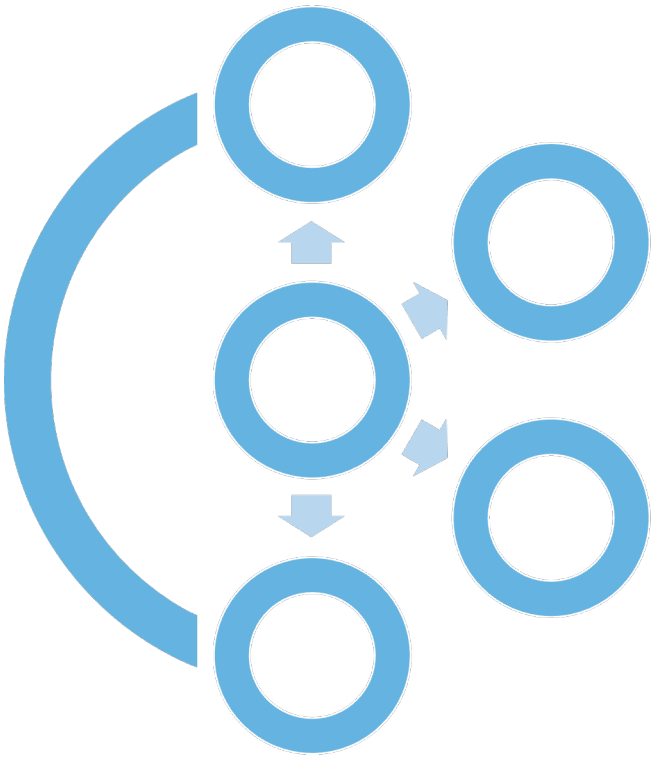
Data Synchronization

The process of keeping data in multiple locations or systems consistent and up-to-date by propagating changes between systems.



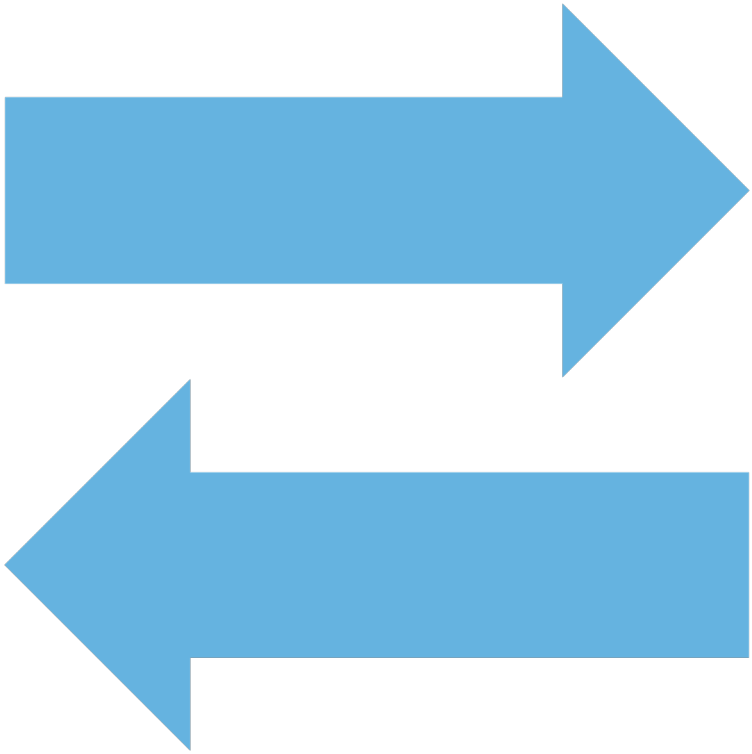
Data Preparation

The process of cleaning, transforming, and enriching data to make it suitable for analysis, reporting, or further processing.



Publish/Subscribe

A messaging pattern that decouples data producers and consumers by using an intermediary (message broker) to manage data distribution, enabling scalable and flexible data integration.



Request/Reply

Pattern: A messaging pattern where a data consumer sends a request to a data producer and waits for a response, allowing for synchronous communication and data exchange.

Data Analytics

Descriptive Analytics	Diagnostic Analytics	Predictive Analytics
Prescriptive Analytics	Real-time Analytics	Batch Analytics
Text Analytics	Geospatial Analytics	Sentiment Analysis
	Network Analytics	

Descriptive Analytics

The analysis of historical data to understand past events and trends, often presented through visualizations, reports, or dashboards.

Diagnostic Analytics

The process of examining data to determine the causes of past events or trends, often using techniques such as data mining, correlations, or anomaly detection.

Predictive Analytics

The use of data, statistical algorithms, and machine learning techniques to predict future events or trends based on historical data.

Prescriptive Analytics

The process of recommending actions or decisions based on data analysis, often using optimization, simulation, or machine learning algorithms.

Real-time Analytics

The analysis of data as it is generated or received, providing immediate insights and enabling rapid decision-making or automated actions.

Batch Analytics

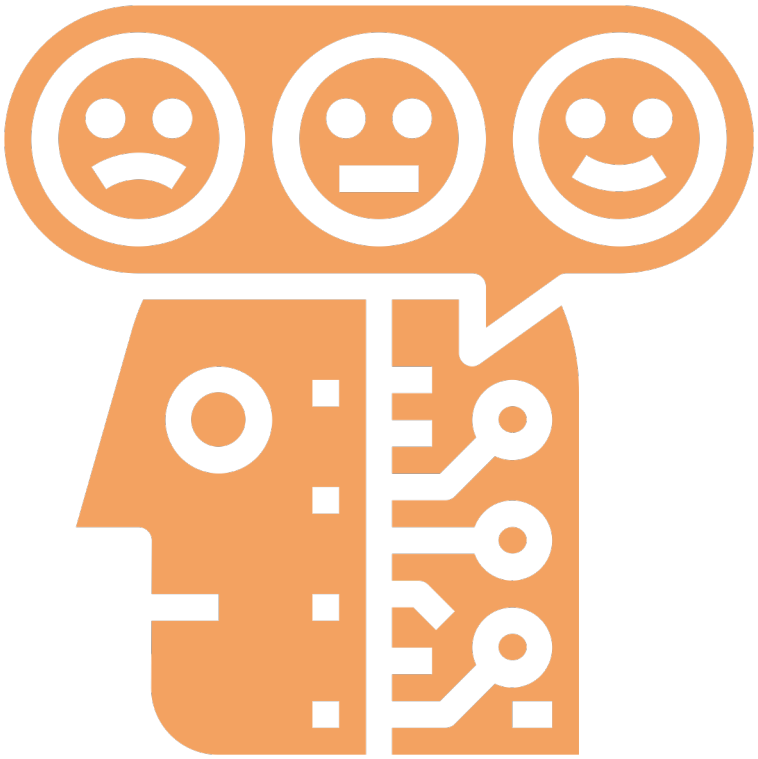
The processing and analysis of large volumes of data in batches, often scheduled at regular intervals or on-demand, providing insights for decision-making or reporting.

Text Analytics

The process of extracting meaningful information and insights from unstructured text data, using techniques such as natural language processing, sentiment analysis, or text mining.

Geospatial Analytics

The analysis of geographically-referenced data, often involving the visualization, exploration, and interpretation of spatial relationships and patterns.



Sentiment Analytics

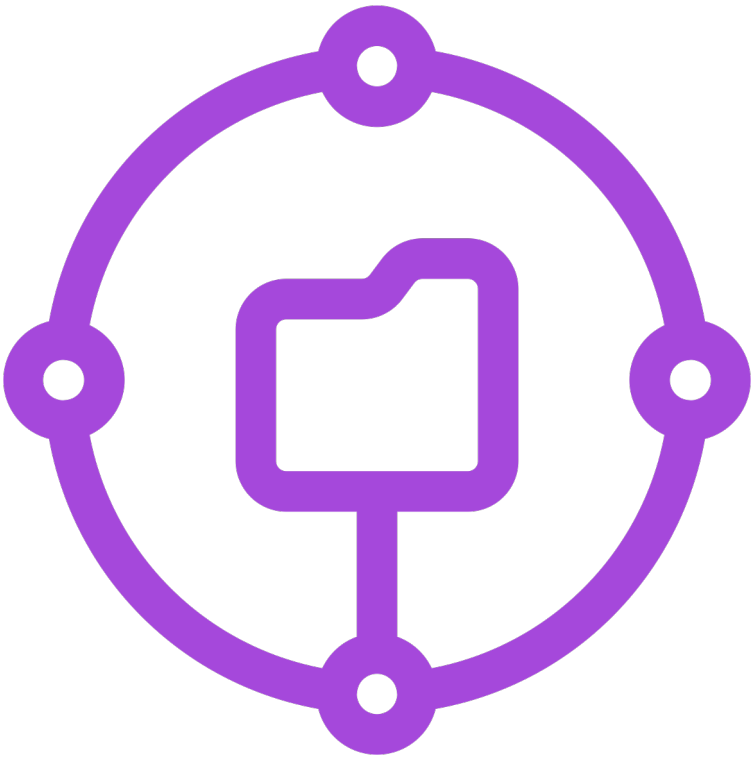
A technique that uses natural language processing (NLP) to determine the sentiment or emotion expressed in textual data, such as reviews, social media posts, or customer feedback.

Network Analytics

The analysis of network data to uncover patterns, trends, and insights related to the relationships and interactions between nodes (entities) in a network.

Data Management

Master Data Management (MDM)	Reference Data Management (RDM)	Metadata Management
Data Catalog	Data Lineage	Data Versioning
Data Provenance	Data Lifecycle Management	Data Virtualization
	Data Profiling	



Master Data Management (MDM)

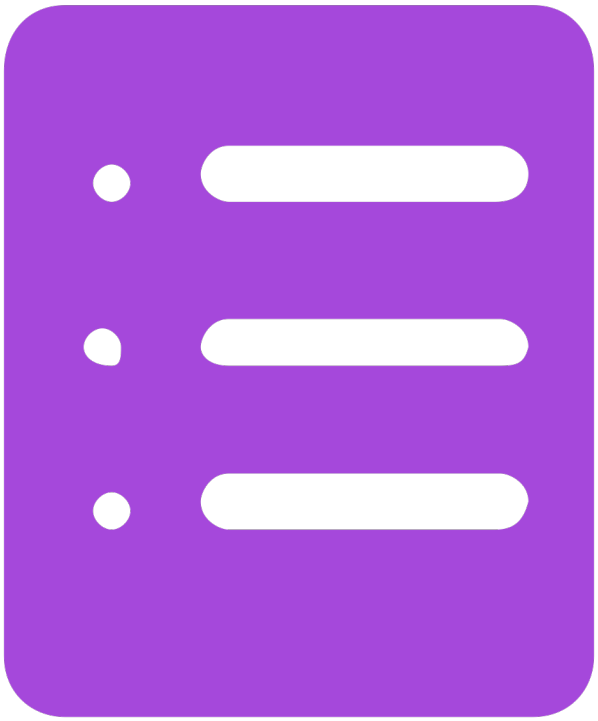
A process of creating a single, consistent, and authoritative source of truth for critical business data, improving data quality and consistency.

Reference Data Management (RDM)

The practice of managing shared data used across multiple systems, such as codes, categories, or hierarchies, to ensure consistency and accuracy.

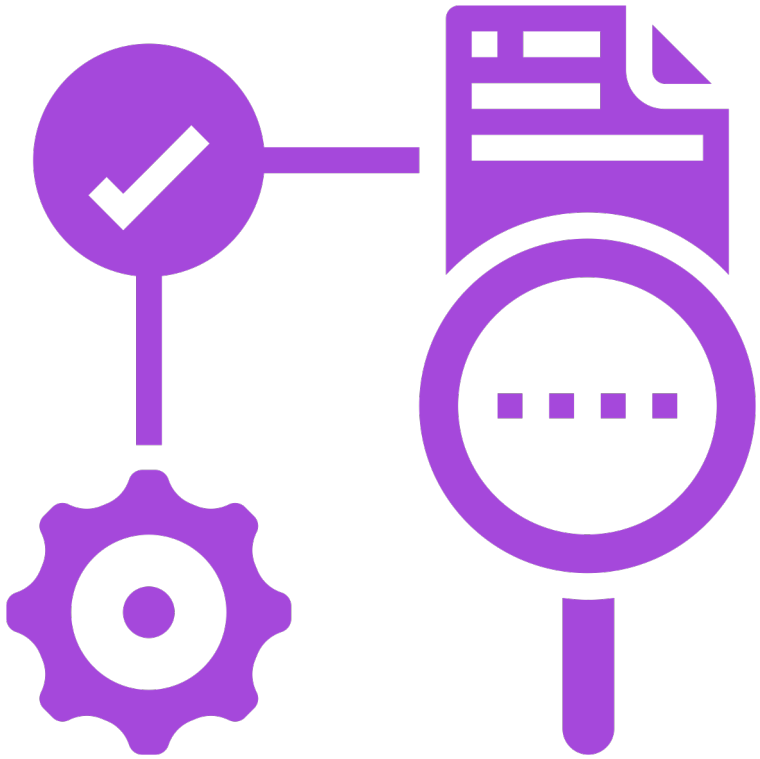
Metadata Management

The process of creating, maintaining, and using metadata (data about data) to facilitate data understanding, discovery, and governance.



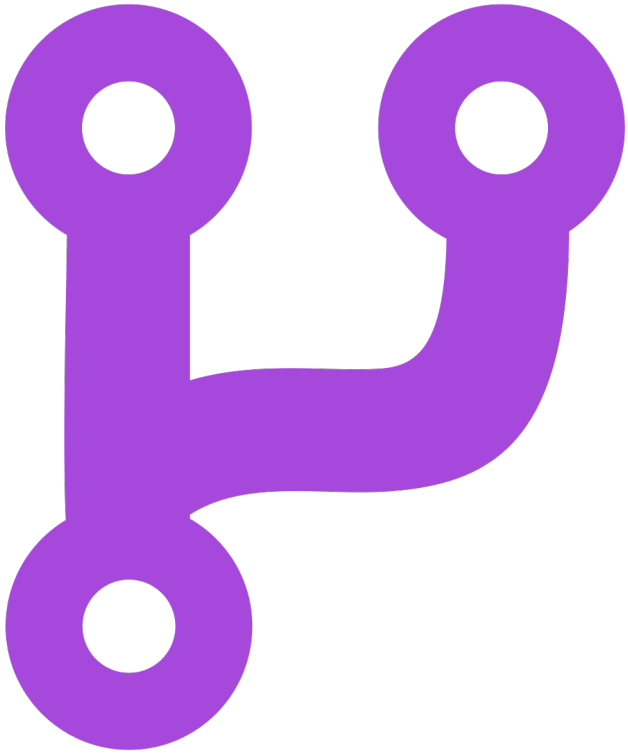
Data Catalog

A centralized repository that provides a searchable and discoverable inventory of an organization's data assets, including datasets, reports, and dashboards.



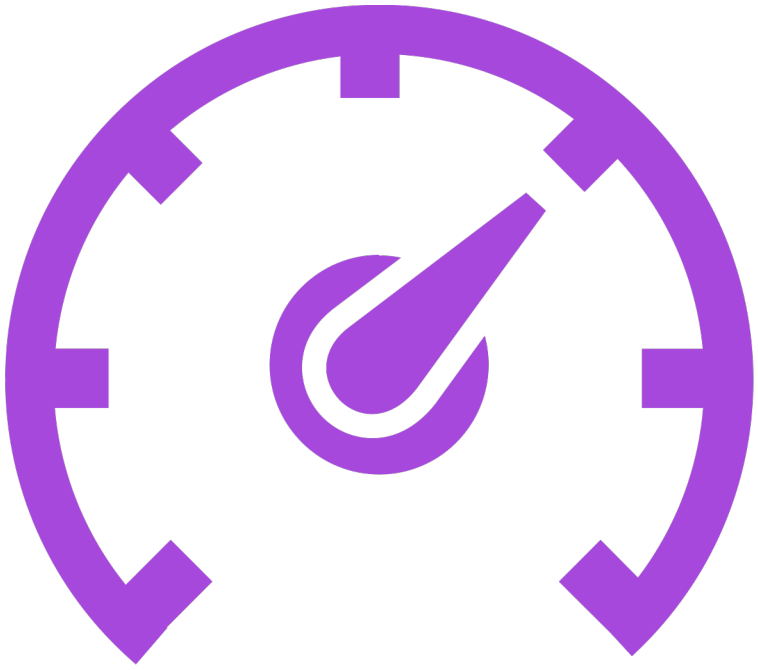
Data Lineage

The practice of tracking the flow of data through systems and processes, including its origin, transformations, and dependencies, to ensure data accuracy and compliance.



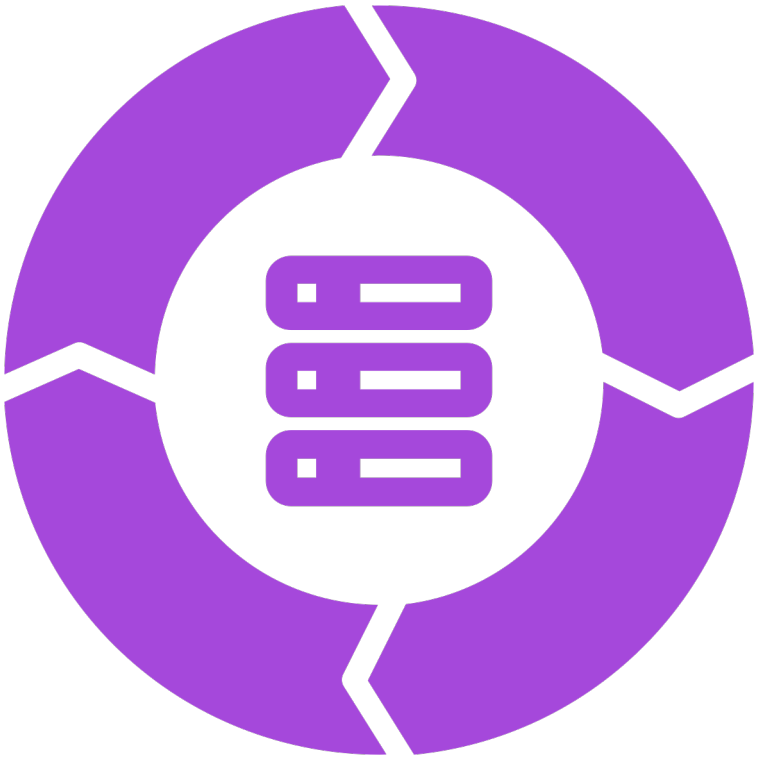
Data Versioning

The process of tracking and managing changes to data over time, allowing for data rollback, recovery, and auditing.



Data Performance

The process of documenting the origin, history, and processing of data, ensuring data trustworthiness, and enabling traceability and reproducibility.



Data Life Cycle Management

A comprehensive approach to managing data throughout its lifecycle, from creation to archival or deletion, ensuring data availability, compliance, and cost optimization.

Data Virtualization

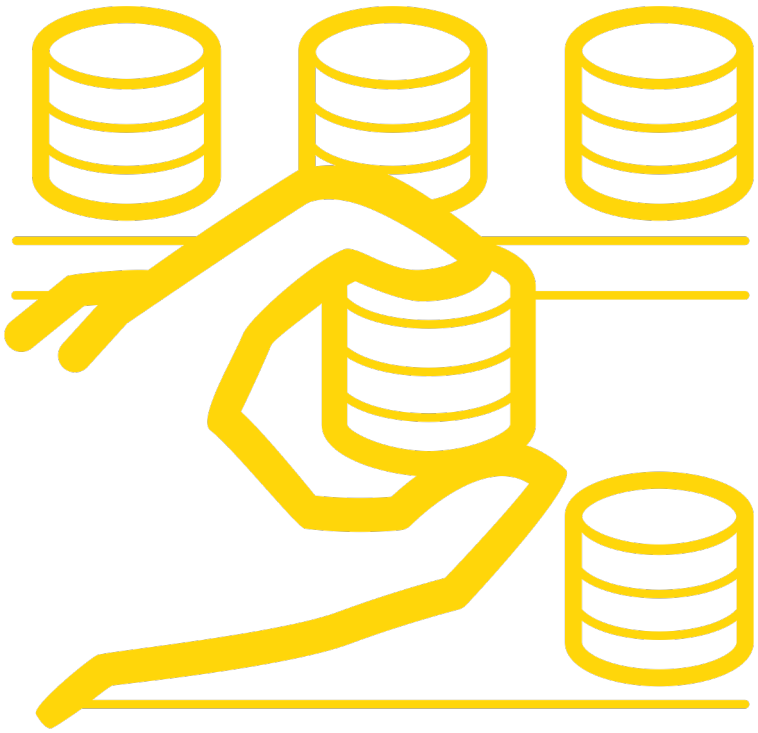
A technique that abstracts the underlying data sources, allowing users and applications to access and manipulate data without needing to know its physical location or structure.

Data Profiling

The process of examining, analyzing, and assessing data quality by collecting statistics, identifying patterns, and detecting anomalies or inconsistencies.

Data Governance

- Data Stewardship
- Data Quality Management
- Data Policy Management
- Data Classification
- Data Retention and Archiving
- Data Privacy Compliance
- Data Lineage and Provenance
- Data Cataloging and Discovery
- Data Risk Management
- Data Ownership



Data Stewardship

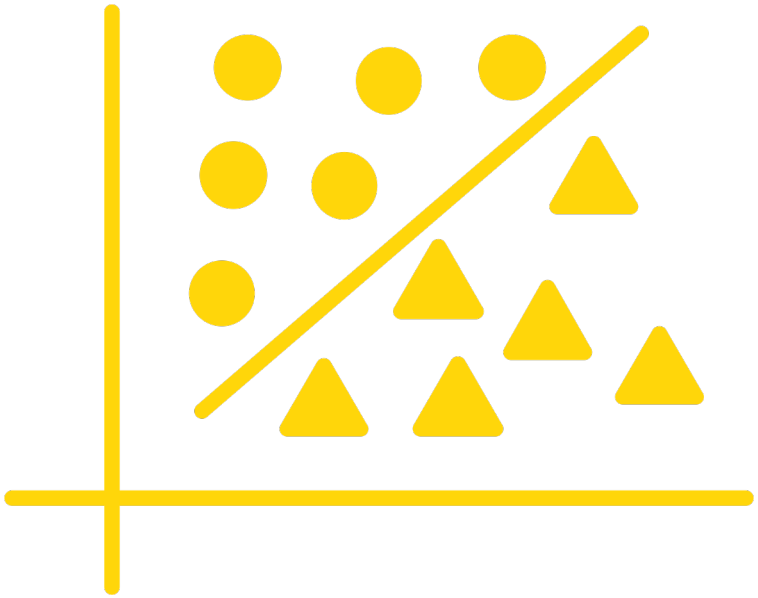
The practice of managing and overseeing an organization's data to ensure its quality, consistency, and compliance with policies and regulations.

Data Quality Management

The process of measuring, monitoring, and improving the accuracy, completeness, consistency, and timeliness of data, ensuring its trustworthiness and reliability.

Data Policy Management

The development, implementation, and enforcement of data policies, standards, and procedures that govern the use, storage, and management of data.



Data Classification

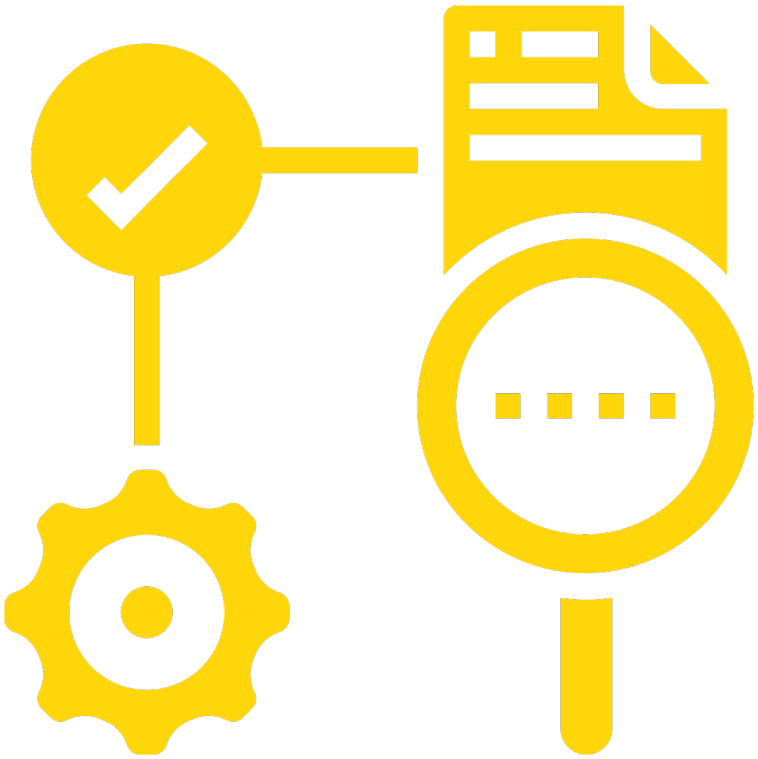
The process of categorizing data based on its sensitivity, value, or risk, enabling the implementation of appropriate security and compliance measures.

Data Retention and Archival

The practice of defining and implementing policies for retaining, storing, and disposing of data based on legal, regulatory, and business requirements.

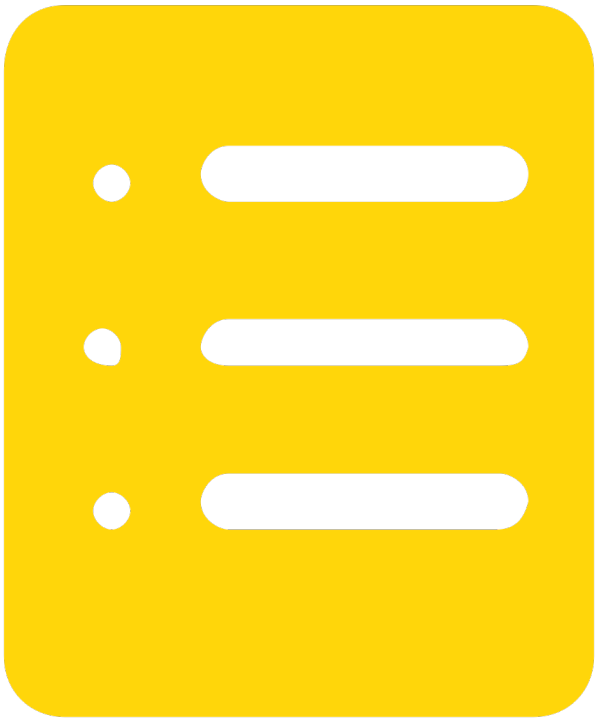
Data Privacy Compliance

The process of ensuring that data management practices adhere to applicable privacy laws, regulations, and standards, such as GDPR or CCPA.



Data Lineage and Provenance

The practice of tracking the flow of data through systems and processes, including its origin, transformations, and dependencies, to ensure data accuracy and compliance.



Data Cataloging and Discovery

The process of creating, maintaining, and using a centralized repository that provides a searchable and discoverable inventory of an organization's data assets.

Data Risk Management

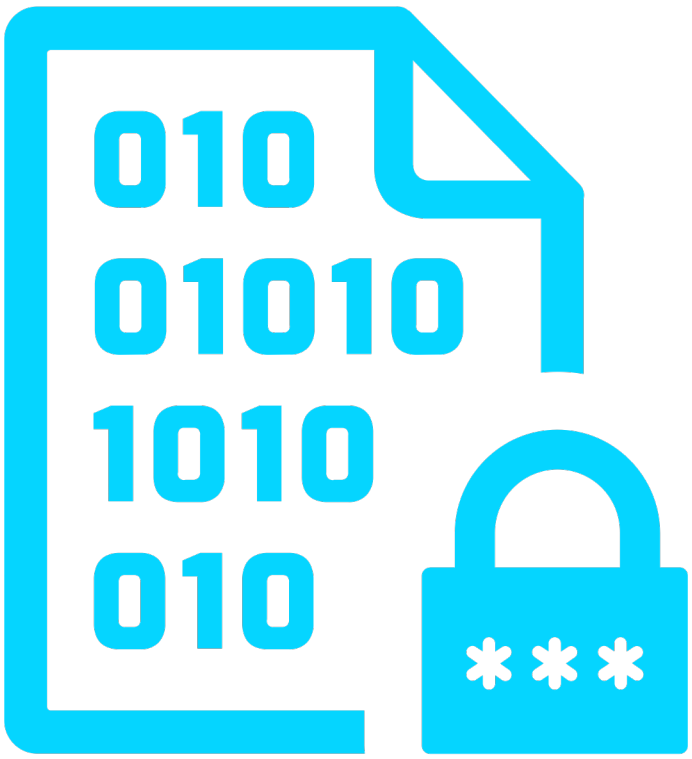
The process of identifying, assessing, and mitigating data-related risks, such as data breaches, data corruption, or regulatory non-compliance.

Data Ownership

The practice of assigning responsibility and accountability for data assets to specific individuals or teams, ensuring proper data management, quality, and governance.

Data Security

Data Encryption	Data Masking	Data Tokenization
Data Access Control	Data Auditing	Data Anonymization
Data Pseudonymization	Data Security Monitoring	Database Activity Monitoring
	Data Loss Prevention	

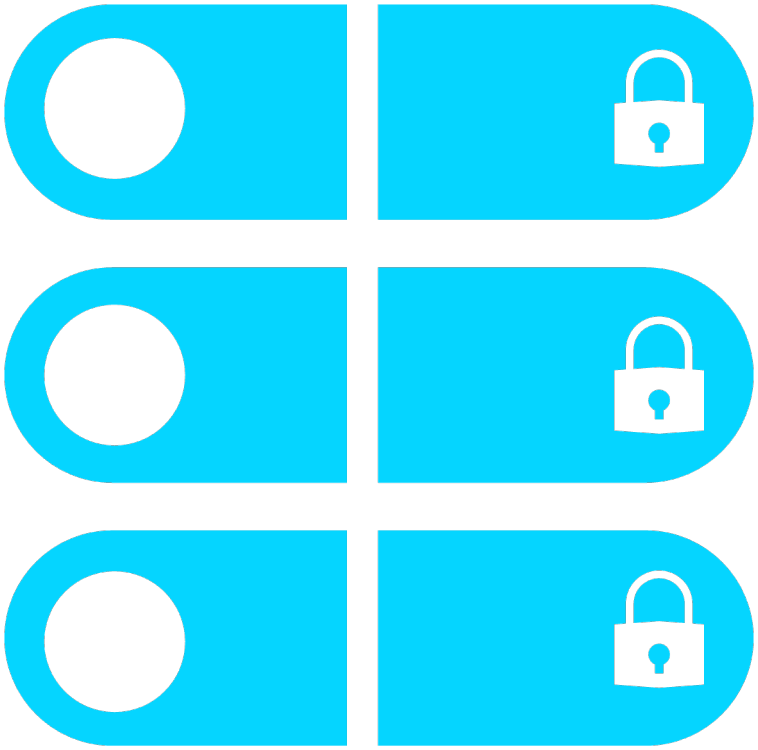


Data Encryption

The process of encoding data to protect it from unauthorized access, both when stored on disk (at-rest) and when transmitted across networks (in-transit).

Data Masking

The technique of obscuring sensitive data by replacing it with fictitious or scrambled data, ensuring that sensitive information is not exposed to unauthorized users.



Data Tokenization

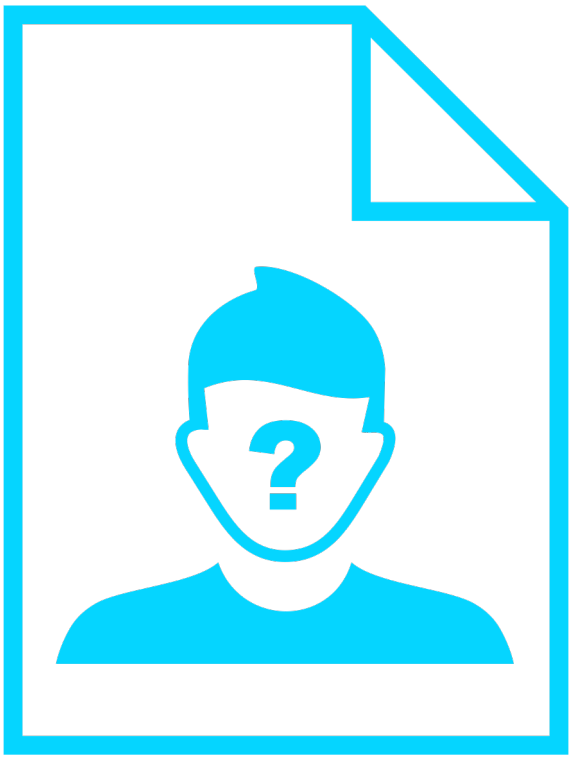
A method of substituting sensitive data with non-sensitive tokens, allowing the original data to be protected while still enabling some operations and analytics to be performed.

Data Access Control

The process of defining and enforcing policies that determine who can access, modify, or delete data, based on roles, responsibilities, and security requirements.

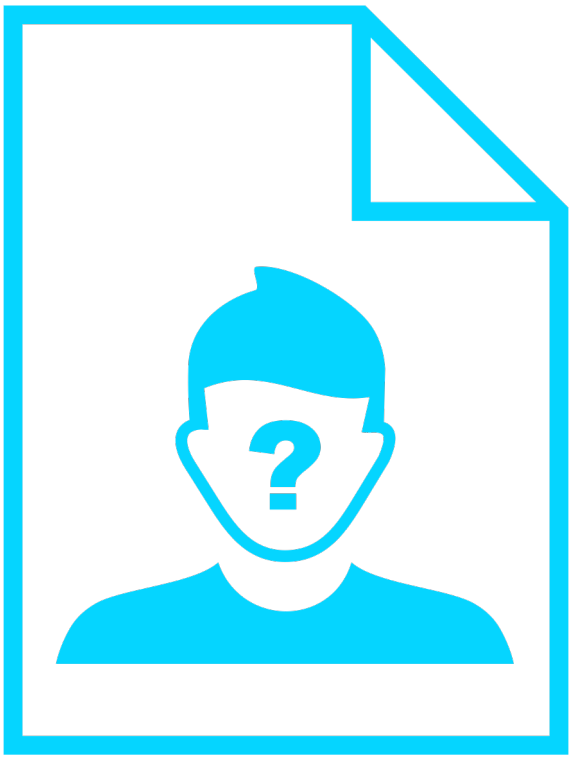
Data Auditing

The practice of monitoring, recording, and analyzing data-related activities and events, enabling the detection of unauthorized access, data breaches, or compliance violations.



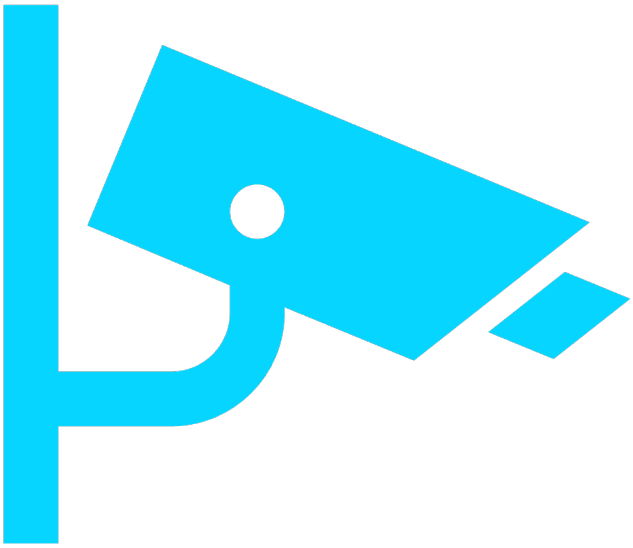
Data Anonymization

The process of removing personally identifiable information (PII) from datasets to protect individual privacy while preserving the utility of the data for analysis.



Data Pseudonymization

A technique that replaces sensitive data with pseudonyms or artificial identifiers, reducing the risk of re-identification while maintaining some level of data usability.



Data Security Monitoring

The practice of continuously monitoring and analyzing data systems, networks, and activities for potential security threats, vulnerabilities, or breaches.

Data Activity Monitoring

The continuous monitoring and analysis of database activities and transactions to detect and prevent unauthorized access, data leaks, or policy violations.

Data Loss Prevention

A set of tools and practices designed to protect sensitive data from unauthorized access, leakage, or theft, by monitoring, detecting, and blocking potential data breaches.

Thank You

Deepak Bhardwaj