ONE PAGE SUMMARY

IT & Software › Data Engineering › Data Pipelines

Introduction to Data Pipelines

Master the flow of data with effective data pipelines.

DEFINITION

> Data pipelines are automated processes that move and transform data from one system to another. They ensure data is accessible and usable for analysis and decision-making.

KEY CONCEPTS

Data Ingestion

The process of collecting and importing data from various sources into a system.

Data Transformation

Modifying data into a suitable format or structure for analysis, including cleaning and aggregating.

Data Storage

The method of saving data in databases or data lakes for future access and analysis.

Workflow Orchestration

Managing and scheduling the sequence of tasks in a data pipeline to ensure smooth execution.

Monitoring and Maintenance

Regularly checking the pipeline's performance and fixing issues to ensure reliability.

EXAMPLES

→ Extracting sales data from a CRM and loading it into a data warehouse.

→ Transforming raw log files into structured datasets for analysis.

→ Automating the daily update of customer information in a reporting dashboard.

MEMORY TIPS

★ Think of a pipeline as a water pipe: data flows through it smoothly.

★ Remember 'I-T-S-M' for Ingestion, Transformation, Storage, Monitoring.

★ Visualize a factory assembly line where raw materials become finished products.

COMMON MISTAKES

✗ Neglecting to validate data after transformation.

✗ Overcomplicating the pipeline with unnecessary steps.

✗ Failing to monitor pipeline performance regularly.

QUICK RECAP

Data pipelines automate the flow of data from sources to storage, transforming it along the way. Key processes include ingestion, transformation, and monitoring to ensure data quality and accessibility.