DATA PIPELINE

What is a Data Pipeline?

A **Data Pipeline** is a set of automated processes that extract data from various sources, transform it into a usable format, and load it into a destination system. It ensures the continuous flow of data from source to target, supporting both batch and real-time processing.

Key Characteristics:

- Automated: Reduces manual intervention
- Scalable: Handles increasing data volumes
- Reliable: Ensures data integrity and consistency
- Flexible: Supports multiple data formats and sources

What is ETL?

ETL stands for:

- **Extract**: Pulling data from source systems
- **Transform**: Cleaning, enriching, and reshaping data
- **Load**: Storing data into target systems

ETL is a subset of data pipeline operations, typically used for batch processing and data warehousing.

Data Pipeline Architecture

Source Systems

- Relational Databases (MySQL, PostgreSQL, Oracle)
- NoSQL Databases (MongoDB, Cassandra)
- APIs (REST, GraphQL)
- Files (CSV, JSON, XML)
- Cloud Storage (AWS S3, Azure Blob, Google Cloud Storage)

ETL Components

- **Extract Layer**: Connectors to source systems, data ingestion tools (e.g., Apache Sqoop, Talend)
- **Transform Layer**: Data cleaning, enrichment, business logic implementation
- **Load Layer**: Writing to target systems, partitioning, indexing

Target Systems

- Data Warehouses (Snowflake, Redshift, BigQuery)
- Data Lakes (Hadoop, Azure Data Lake)
- BI Tools (Power BI, Tableau, Looker)

Orchestration & Scheduling

- Apache Airflow
- AWS Glue
- Azure Data Factory
- Prefect

Data Flow Example

- Extract customer data from CRM and sales databases
- Transform by cleaning, joining with product data, and calculating KPIs
- Load into a Snowflake data warehouse for dashboard reporting

---

Data Quality and Validation

Maintaining data quality is critical. Common validation steps include:

- Schema validation
- Null checks
- Duplicate detection
- Range checks
- Referential integrity

Monitoring and Logging

Monitoring ensures pipeline health and quick issue resolution.

- Job status tracking
- Alerting on failures
- Performance metrics (latency, throughput)
- Audit logs

Tools: Prometheus, Grafana, ELK Stack, Datadog

Security and Compliance

- Data encryption (at rest and in transit)
- Access control and IAM policies
- Data masking for sensitive information
- Compliance with GDPR, HIPAA, SOC 2

Best Practices

- Use modular ETL components for reusability
- Implement version control (Git) for pipeline scripts
- Document data sources, transformations, and dependencies
- Schedule jobs during off-peak hours for batch loads
- Use idempotent operations to avoid duplicate loads

Tools and Technologies

Category	Tools/Technologies
ETL Frameworks	Talend, Apache NiFi, dbt
Orchestration	Airflow, Prefect, Luigi
Storage	S3, HDFS, Azure Blob
Processing	Apache Spark, Pandas, SQL
Monitoring	Prometheus, Grafana

Category	Tools/Technologies
Data Warehousing	Snowflake, BigQuery

Use Cases

- Customer 360 View: Integrate data from CRM, support, and sales
- Marketing Analytics: Analyze campaign performance across channels
- Financial Reporting: Consolidate transactional data for compliance
- Machine Learning: Prepare training datasets from raw logs

Future Enhancements

- Integration with streaming platforms (Kafka, Flink)
- Support for real-time ETL
- Automated schema evolution
- Enhanced data lineage and cataloging
- Incorporation of AI/ML for anomaly detection

Glossary

- **ETL**: Extract, Transform, Load
- **Data Lake**: Centralized repository for raw data
- **Data Warehouse**: Structured storage optimized for analytics
- **Orchestration**: Automated scheduling and management of workflows
- **Data Lineage**: Tracking data origin and transformations