The Hoang

Data Engineering in MLOps

Understand and Implement Production-Grade Machine Learning Operations

Learning Objectives

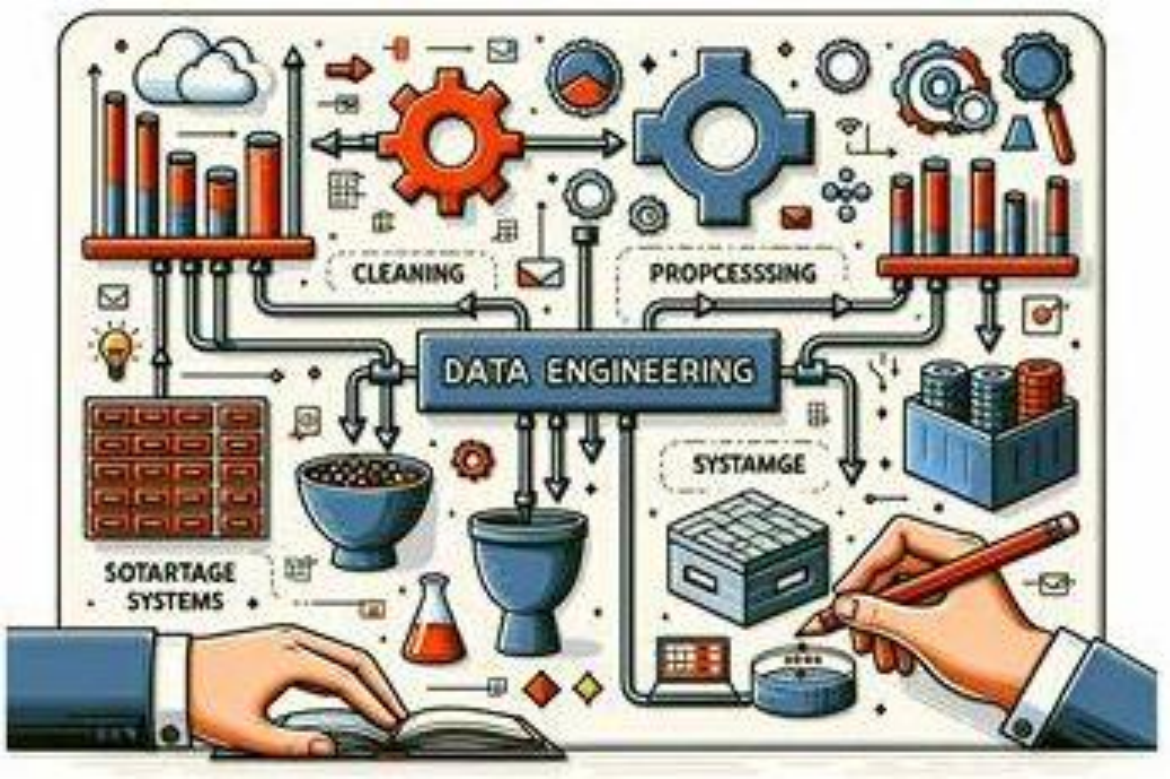
- Introduction of Data Engineering
- Data Engineering: Core Components
- Data Engineering in Machine Learning, MLOps
- Hands-On Practice

What is Data Engineering?

Definition

Data Engineering is the practice of designing, building, and maintaining data pipelines that prepare data for analysis and machine learning models.

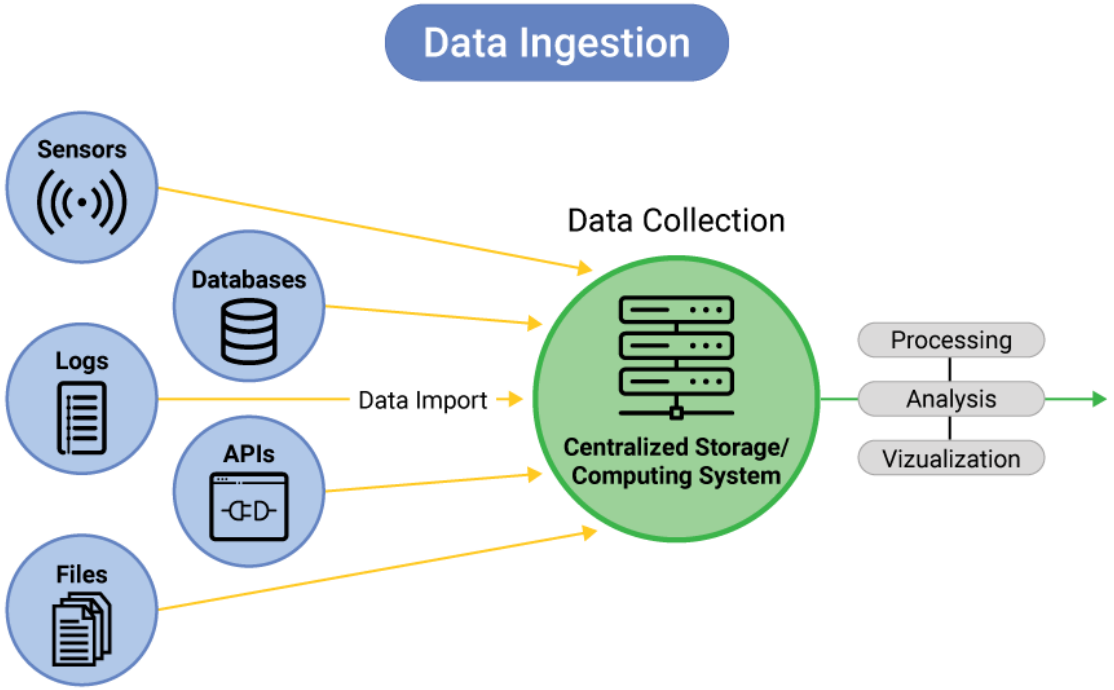
It involves data ingestion, transformation, storage, and retrieval.



Core Components – Data Ingestion

This is the process of collecting data from various sources.

- Batch Ingestion → Data collected at intervals (e.g., da logs, files).
- Streaming Ingestion → Real-time data processing (e.g Kafka, Kinesis).
- ETL (Extract, Transform, Load) vs. ELT (Extract, Load Transform).



Common Tools:

- Batch: Apache Nifi, Airflow, dbt, Talend
- Streaming: Apache Kafka, Apache Flink, AWS Kinesis
- Cloud Storage: AWS S3, Google Cloud Storage, Azure Blob

Core Components – Data Storage

Once data is ingested, it needs to be stored efficiently.

- Databases (OLTP) → For transactional data (e.g., MySQL, PostgreSQL).
- Data Warehouses (OLAP) → For analytics (e.g., Snowflake, BigQuery, Redshift).
- Data Lakes → For raw & semi-structured data (e.g., AWS S3, Delta Lake).
- Lakehouse → Hybrid of Data Lakes & Warehouses (e.g., Databricks, Iceberg).

Storage Formats:

- Structured → SQL databases, Parquet, ORC
- Semi-structured → JSON, Avro, XML
- Unstructured → Logs, images, videos
- Feature Stores (e.g., Feast, Tecton) store ML features for reuse.

Core Components – Data Processing

Once stored, data often needs cleaning & transformation.

- Batch Processing → Large data at scheduled times (e.g., Apache Spark, Pandas).
- Stream Processing → Real-time processing (e.g., Apache Flink, Kafka Streams).

ETL vs. ELT for ML:

- ETL: Preprocess data before storing it.
- ELT: Store raw data and transform it later using SQL or Spark.

ML Pipelines often use ELT because raw data can be reused for different tasks.

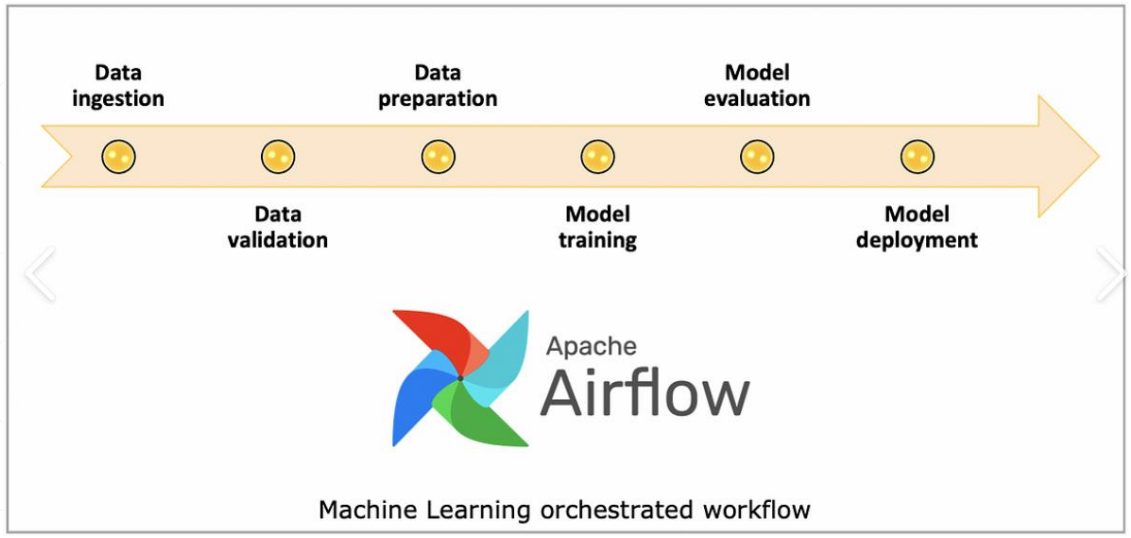
Workflow Orchestration

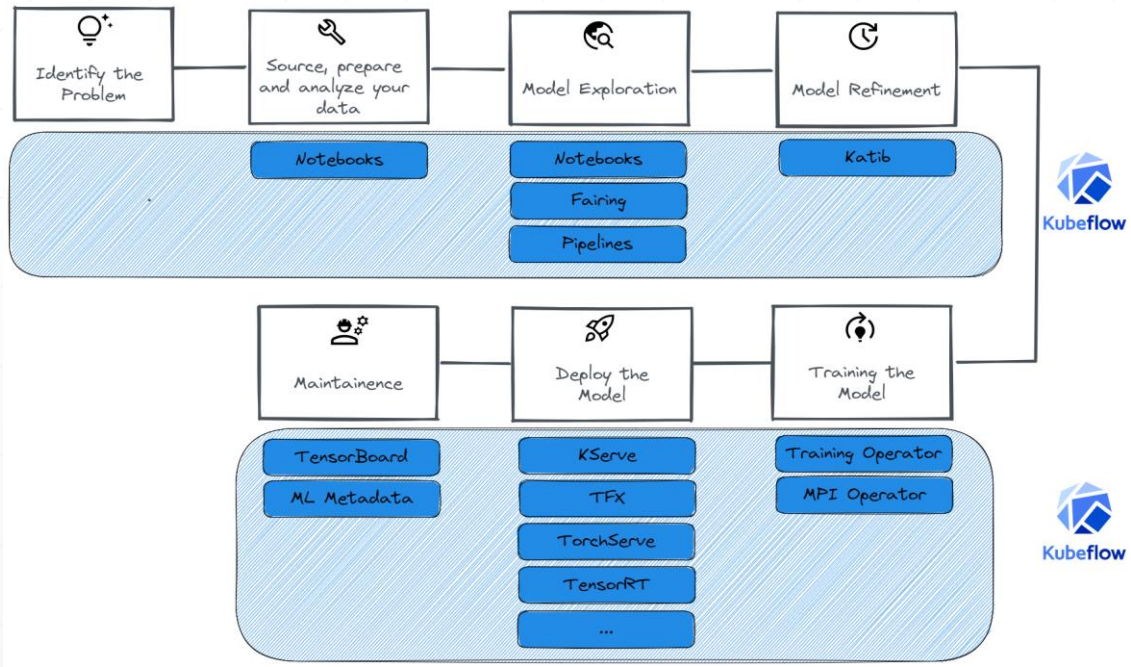
Data Engineering workflows need to be automated.

- Apache Airflow: Popular for scheduling pipelines.
- Prefect, Dagster: Modern alternatives.
- KubeFlow Pipelines: Designed for ML workflows.

For ML & MLOps:

- Automate data preparation for model training.
- Run feature engineering jobs periodically.
- Integrate with ML pipelines (e.g., Airflow + Kubeflow).





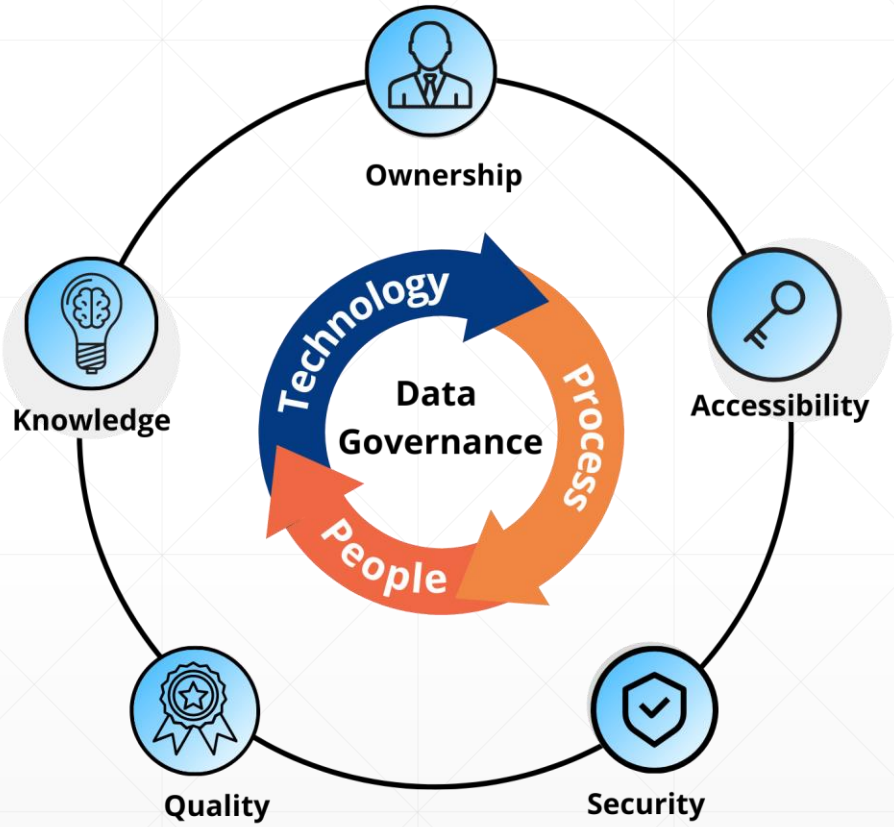
Data Quality & Governance

ML models rely on high-quality data.

- Data Validation → Great Expectations, Deequ
- Data Lineage → Apache Atlas, DataHub
- Schema Management → Apache Avro, Protobuf
- Feature Store Governance → Feast

Key for ML:

- Detect data drift and anomalies.
- Ensure consistent feature engineering.



Building a Data Pipeline for ML – 1/2

Step 1: Data Collection

- Collect structured & unstructured data.
- Use Kafka/Kinesis for real-time ingestion.

Step 2: Data Storage

- Store raw data in S3, Delta Lake, or BigQuery.
- Use Feature Stores for ML features.

Step 3: Data Processing

- Use Apache Spark, dbt, or Pandas for transformation.
- Stream processing with Apache Flink/Kafka Streams.

Building a Data Pipeline for ML – 2/2

Step 4: Data Validation

- Use Great Expectations or Deequ to test data quality.

Step 5: Workflow Automation

- Use Apache Airflow or Prefect to schedule jobs.

Step 6: Model Training & Deployment

- Use Kubeflow, MLflow, or TFX for ML lifecycle.
- Monitor data drift to trigger retraining.

Key Data Engineering Skills for MLOps

Category	Tools & Frameworks
Programming	Python, SQL, Scala
Data Storage	PostgreSQL, Snowflake, BigQuery, Delta Lake
Data Processing	Apache Spark, Pandas, dbt, Flink
Data Orchestration	Apache Airflow, Prefect, Dagster
Streaming	Apache Kafka, AWS Kinesis
Feature Store	Feast, Tecton, Hopsworks
Data Validation	Great Expectations, Deequ
MLOps	MLflow, Kubeflow, TFX

Recommended Learning Resources

Books

- Designing Data-Intensive Applications – Martin Kleppmann
- Fundamentals of Data Engineering – Joe Reis & Matt Housley
- Data Pipelines with Apache Airflow – Bas Harenslak

Courses

- Google Cloud Data Engineer Certification (Coursera)
- Data Engineering Zoomcamp (Free, by DataTalksClub)
- Spark & Airflow for Data Engineering (Udemy)

Hands-On Learning

- Build a real-time analytics pipeline using Kafka & Flink.
- Create a feature store and integrate it into an ML model.
- Automate data pipelines with Airflow & MLOps tools.

Hands-On Practice

1. Data Ingestion & Storage
2. Data Processing & Transformation
3. Data Orchestration with Apacheflow