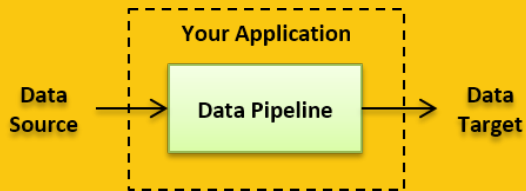


Data Pipeline



Iman Dehdar
Scala Developer & Data Engineer

Data Engineer

Scala

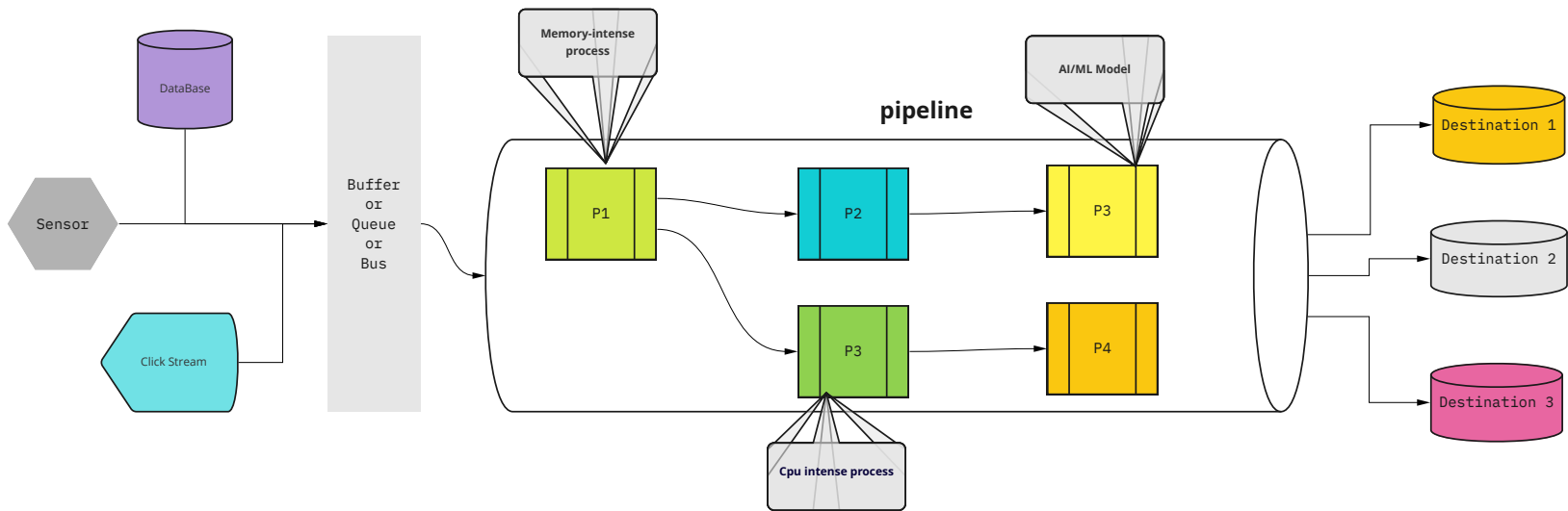
Python Java SQL

Big Data Engineer

MLOps

Machine Learning Engineer

Batch Data Pipeline
Event-Driven Pipeline
Lambda Data Pipeline



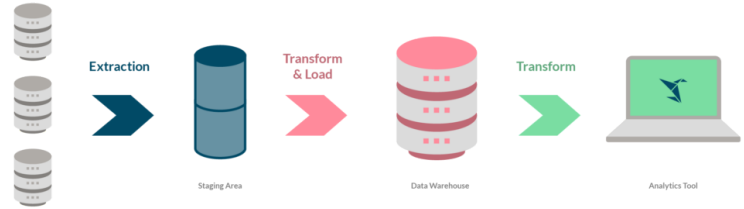
Agenda

- ETL Pipeline - Batch
- Event-Driven Pipeline - Stream
- Machine Learning Pipelines - Mix
- Data Architecture

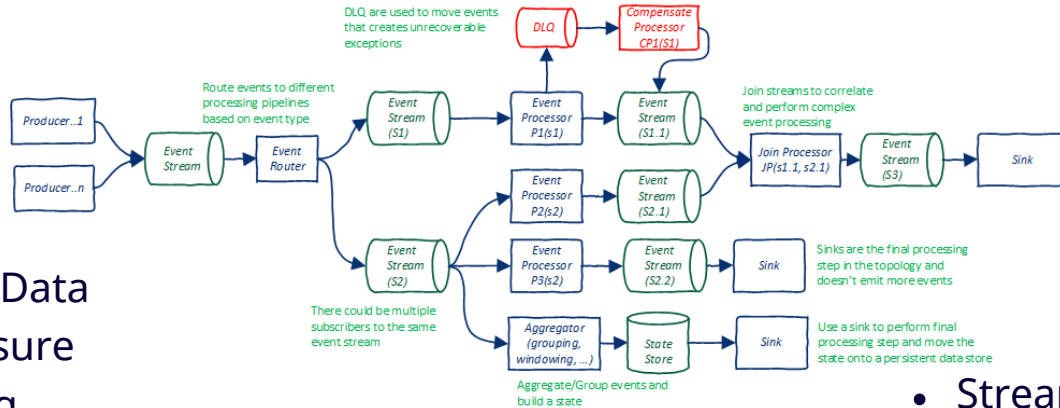
ETL Pipeline

Extract, Transform, Load

- Batch
- Governance & Staging
- Data validation & Alerting
- Data Transformation & Cleaning(Glue, Spark, Hadoop, ...)
- Cloud Storage(S3, Ceph, Delta Lake, ...)
- Data Warehouse(Redshift, Databricks Lakehouse, MongoDB, ...)



Event-Driven Pipeline

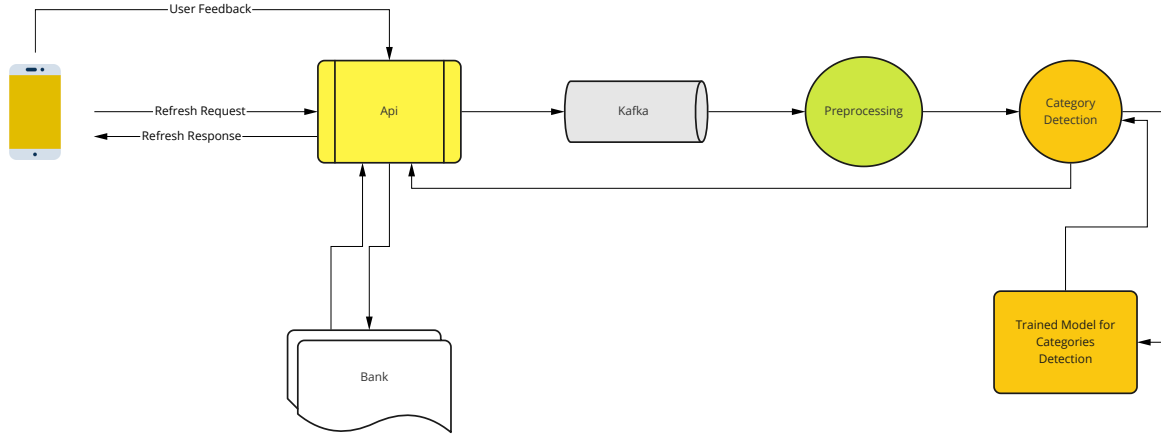


- Stream of Data
- Back-Pressure
- Windowing

- Stream Joins
- Stateful Aggregation

- Streaming Libraries (Akka Stream, Kafka Stream, ...)
- Streaming System (Spark, Flink, ...)
- Message broker (kafka, RabbitMQ, ...)

Machine Learning Pipeline

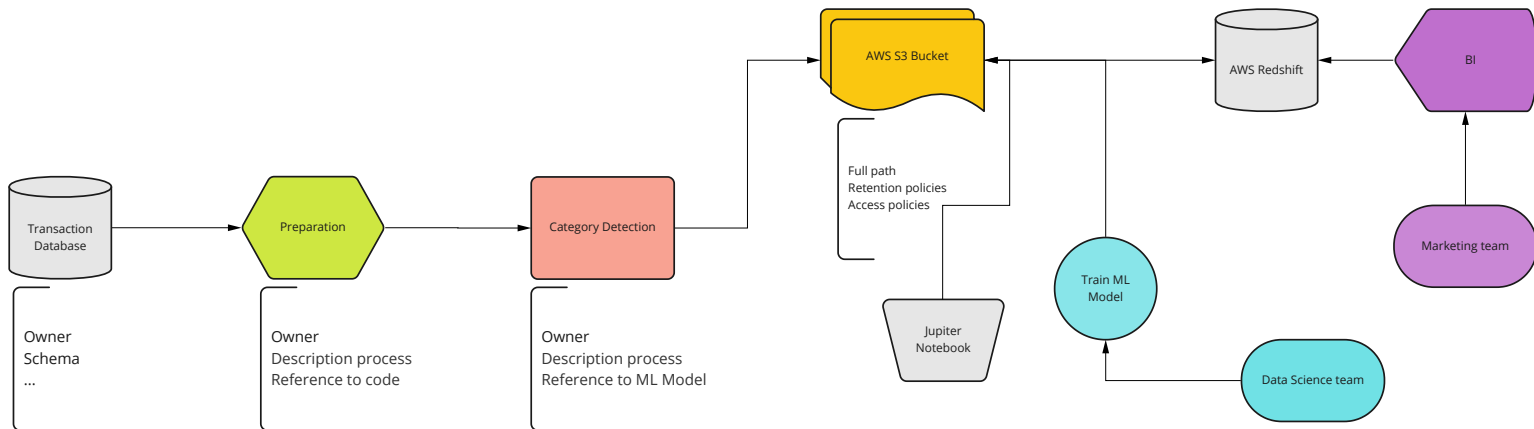


Data Architecture

"A modern business strategy depends on data."

- A blueprint for managing data
 - Data Flow
 - Data Process
 - Stakeholders & Business goal
 - Policies & governance

- How
 - From on-premise to cloud-based
 - from batch to read-time
 - From pre-integrated to modular
 - From enterprise warehouse to domain-based



Q&A