Labs 6–8 Summary: Airflow Orchestration, Kafka, and Structured Streaming

Z5008 Big Data Lab · Even Semester 2026 · Dr. Innocent Nyalala

1. Batch DAGs with Apache Airflow

Airflow expresses pipelines as Directed Acyclic Graphs with explicit task dependencies, idempotent retries, backfills, and lightweight metadata passing via XComs.

2. Kafka and Spark Structured Streaming

Kafka decouples producers and consumers via append-only partitioned logs indexed by consumer-group offsets. Spark Structured Streaming treats live streams as unbounded tables, combining event-time tumbling/sliding windows with watermarks $W = \max(t_{\text{event}}) - \Delta_{\text{late}}$ and write-ahead checkpoints for fault-tolerant exactly-once sinks.