End-to-End Data Pipeline Development

This assignment challenges you to construct a comprehensive data pipeline that mirrors real-world enterprise architectures. You'll gain hands-on experience with the complete data lifecycle—from generation through to reporting—whilst mastering industry-standard tools including Apache NiFi, Kafka, and Hadoop. The objective is to develop both technical proficiency and architectural thinking necessary for production-scale data engineering.

Core Pipeline Components and Technical Requirements

01

Data Generation

Design and implement a system to generate large volumes of synthetic data, such as application logs or network traffic captures. Your generator must produce realistic, high-throughput data streams that simulate production environments—aim for sustained generation rates that stress-test downstream components.

02

Data Ingestion and Transfer

Configure Apache NiFi to establish robust network paths for data transfer to designated storage locations. Implement appropriate flow control, back-pressure handling, and monitoring to ensure reliable ingestion even under peak loads.

03

Stream Processing with Kafka

Integrate Apache Kafka as your distributed streaming platform to decouple data producers from consumers. Configure topics, partitions, and replication factors appropriate for your data volume whilst ensuring fault tolerance and scalability.

04

Data Transformation

Implement either ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform) methodology to process raw data. Leverage Apache Hadoop for distributed storage and processing, applying transformations that clean, enrich, and structure your data for analytical queries.

05

Reporting and Querying

Develop mechanisms to report and query processed data over configurable time windows—last 24 hours, 7 days, or 30 days. Implement efficient indexing and partitioning strategies to enable rapid query response times across large datasets.