

Apache Kafka Analytics

Duration: 3 Days

Description

The **Apache Kafka Analytics Training** aims to teach the core features of Apache Kafka, processing real-time data streams, and integrating with analytical systems. This 3-day training program is designed for analytics professionals, data engineers, and software developers who want to understand how Kafka is used in data collection, processing, storage, and analysis workflows

Audience

- Teams starting or working on Kafka-based projects
- Data Scientists
- Data Engineers
- Data Analysts

Outline

Day 1: Apache Kafka and Data Integration

Introduction to Apache Kafka

- Kafka's Role in Data Analytics
- Advantages of Kafka for Real-Time Data Analytics
- Key Kafka Components (Producer, Consumer, Topic, Partition, Broker)

Data Integration with Kafka Connect

- Kafka Connect Architecture and Use Cases
- Source and Sink Connectors
- Using JDBC Connector for Data Ingestion and Delivery
- HDFS and Object Storage Sink Connector Usage
- Writing and Configuring Custom Connectors

Querying Data with Kafka SQL

- What is KSQLDB?
- Concepts of Streams and Tables in KSQL
- Data Filtering and Transformation with KSQL

- Stream-Stream and Stream-Table Joins
- Creating Real-Time Dashboards with KSQL

Day 2: Real-Time Data Processing with Apache Flink and Spark

Processing Kafka Data with Apache Flink

- Overview of Flink Architecture
- Flink and Kafka Integration
- Stream Processing with Flink
- Flink Windowing Concepts and Use Cases
- Stateful Processing with Flink

Processing Kafka Data with Apache Spark

- Introduction to Spark Streaming Architecture
- Reading Data from Kafka with Spark Structured Streaming
- Window Aggregation and Filtering with Spark Streaming
- Combining Stream and Batch Processing with Spark
- Checkpointing and Fault Tolerance Management

Day 3: Advanced Applications and Performance Optimization

Advanced Kafka SQL and Stream Processing

- Aggregation and Window Functions with KSQL
- Persistent Queries in KSQL
- Optimizing Performance with Partitioning and Parallelism

Advanced Applications with Flink

- Managing State Backends in Flink
- Event Time and Watermark Usage
- Detecting Data Anomalies and Alerting with Flink

Advanced Applications with Spark

- ML Pipeline Integration with Spark Streaming
- Real-Time Analytics Dashboard with Spark and Kafka
- Spark Streaming Performance Optimization

Kafka Ecosystem Tools for Data Analytics

- Kafka Monitoring Tools (Confluent Control Center, Prometheus, Grafana)
- Using Schema Registry and Avro/Protobuf Serialization
- Log Compaction and Cleanup Policy Management

Prerequisites

Basic Java programming knowledge is required.