Introduction to Apache Kafka

Martin Brucker

June 20. 2024

Table of Contents

Introduction

- **Apache Kafka**: A distributed streaming platform designed for high-throughput, fault-tolerant, and real-time data processing.
- Combines the features of both a **messaging system** and a **log system**.
- Originally developed by LinkedIn and later open-sourced.

Distributed Message Brokers and Log Systems

- **Message Brokers**:
  - Queues and topics for asynchronous communication.
  - Point-to-point and publish-subscribe models.
- **Log Systems**:
  - **Distributed commit log**: Stores all published messages.
  - **Append-only storage**: Immutable logs for durability.

IT Architecture of Kafka

- **Topics**: Logical channels for data streams.
- **Partitions**: Segments of topics stored on different servers.
- **Producers**: Data publishers that write to topics.
- **Consumers**: Data subscribers that read from topics.

Client Libraries

- **Java**: Official client library (highly recommended).
- **Python**: Confluent's Python client (for Python enthusiasts).
- **Other languages**: Community-supported libraries (e.g., Go, .NET).

Typical Use Cases

- **Real-time analytics**: Process and analyze data as it arrives.
- **Log aggregation**: Centralize logs from various services.
- **Event sourcing**: Capture and replay events for stateful systems.
- **Stream processing**: Transform and enrich data streams.

Installation and Configuration

- **Download and install Kafka**: Obtain the Kafka distribution.
- **Configure Kafka properties**:
  - Specify server settings (e.g., broker ID, port).
  - Define topics and their partitions.
- **Start Kafka server and Zookeeper ensemble**:
  - Kafka relies on Zookeeper for coordination.
  - Start both services to enable Kafka functionality.

Programming with Kafka

- **Produce messages** using Kafka producers.
- **Consume messages** using Kafka consumers.
- Handle message **serialization and deserialization**.

Conclusion

- Apache Kafka is a powerful tool for building scalable and reliable data pipelines.
- Dive deeper into its documentation and explore advanced usage scenarios.