**Apache Kafka**

* **Introduction**

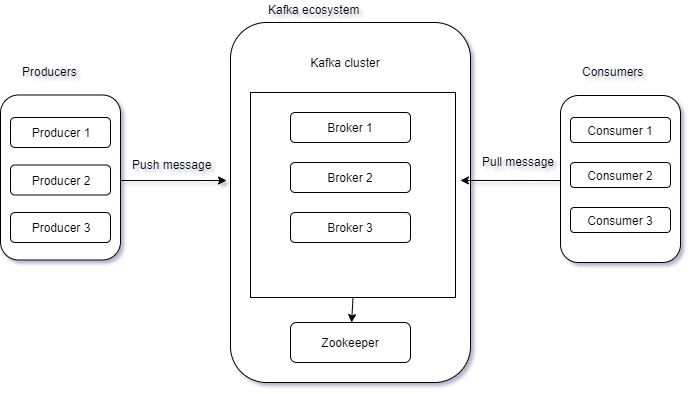
In Big Data, an enormous volume of data is used. Regarding data, we have two main challenges. The first challenge is how to collect large volume of data and the second challenge is to analyze the collected data. To overcome those challenges, you must need a messaging system.

Kafka is designed for distributed high throughput systems. Kafka tends to work very well as a replacement for a more traditional message broker. In comparison to other messaging systems, Kafka has better throughput, built-in partitioning, replication and inherent fault-tolerance, which makes it a good fit for large-scale message processing applications.

Apache Kafka is a community distributed event streaming platform capable of handling trillions of events a day. It is a distributed publish-subscribe messaging system and a robust queue that can handle a high volume of data and enables you to pass messages from one end-point to another. Kafka messages are persisted on the disk and replicated within the cluster to prevent data loss. Kafka is built on top of the ZooKeeper synchronization service.

Thousands of companies are built on Kafka that are Airbnb, Netflix, Linked in, Microsoft, Target etc.

* **Cluster Architecture**



A Brief explanation about Kafka core components are given below,

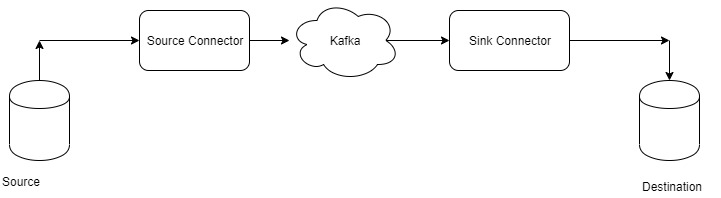
1. **Broker**: Kafka cluster typically consists of multiple brokers to maintain load balance. Kafka brokers are stateless, so they use ZooKeeper for maintaining their cluster state. One Kafka broker instance can handle hundreds of thousands of reads and writes per second and each broker can handle TB of messages without performance impact. Kafka broker leader election can be done by ZooKeeper.
2. **ZooKeeper:** ZooKeeper is used for managing and coordinating Kafka broker. ZooKeeper service is mainly used to notify producer and consumer about the presence of any new broker in the Kafka system or failure of the broker in the Kafka system.
3. **Producers:** Producers push data to brokers. When the new broker is started, all the producers search it and automatically sends a message to that new broker. Kafka producer doesn’t wait for acknowledgements from the broker and sends messages as fast as the broker can handle.
4. **Consumers:** Since Kafka brokers are stateless, which means that the consumer has to maintain how many messages have been consumed by using partition offset. If the consumer acknowledges a particular message offset, it implies that the consumer has consumed all prior messages. The consumer issues an asynchronous pull request to the broker to have a buffer of bytes ready to consume. The consumers can rewind or skip to any point in a partition simply by supplying an offset value. Consumer offset value is notified by ZooKeeper.

The docker-compose-file for kafka cluster setup is https://github.com/confluentinc/cp-all-in-one/blob/7.0.1-post/cp-all-in-one/docker-compose.yml

* **Building Data pipelines**

In this section presumes that you know what is kafka and its core components. This section focus on how to build simple data pipeline using kafka. To Build data pipeline we need

1. Broker
2. Zookeeper
3. Connect
4. Database



Here the requirement is to collect all the data from source side to kafka topic and then move that data to destination side. In between these actions we have to analyse the stored information in kafka topic before writing to destination side.

Let’s see how to solve this,

So our first requirement is to collect all the data from source side, to achieve this we need a “***Source Connector***”. The responsibility of source connector is to collect the data from source data source and push that information to kafka topic. In Data Pipeline Source connectors are the “***Publishers***”. There are two types of source connectors namely JDBC Source connector and Debezium Source Connector. In this looking into JDBC source connector it helps transfer data from database to kafka.

Our next requirement is to transfer data from kafka to external database(destination), to achieve this we need a “***Sink Connector***”. In Data Pipeline Sink connectors are the “***Subscribers***”.

In a nutshell, JDBC source connector helps transfer data from database to kafka, while JDBC sink connector transfer data from kafka to any external databases.