Lund&Bendsen
developing developers

Kafka overview

Exercise

Lund&Bendsen
developing developers

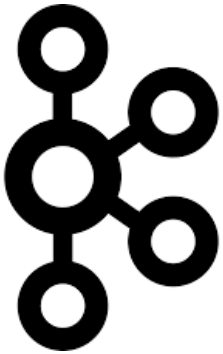
2.1

- Start kafka cluster
  - Make sure DockerDesktop is running
  - Open PowerShell and navigate to <project-folder>/docker
  - Run command **docker run hello-world** to see if docker I up and running
  - Type **docker-compose.exe up –d**
  - This will start a cluster with one zookeeper server and two Kafka brokers in detached mode.
  - Expect 5 min install time!
  - Make sure to keep this window running
  - All services are running in docker containers. The two brokers are exposed on port 29092 and 39092 on the host network
  - Open a new PowerShell window and navigate to <project-folder>/docker
  - Type **docker ps**
  - You should see something like this, showing all running containers:

```
PS C:\LB\LB2628-Kafka> docker ps
CONTAINER ID   IMAGE                              COMMAND                  CREATED       STATUS                   PORTS                                                  NAMES
db3ddc5a68e6   docker-connect-standalone          "start-kafka.sh"         3 hours ago   Up 53 minutes            0.0.0.0:8083->8083/tcp                                 connect-standalone
2b91310a8923   debezium/connect:1.6               "/docker-entrypoint.…"   3 hours ago   Up 53 minutes            8778/tcp, 9092/tcp, 9779/tcp, 0.0.0.0:9090->8083/tcp   postgres-debezium
57b3b8d9658c   tchiotludo/akhq                    "docker-entrypoint.s…"   3 hours ago   Up 3 hours (healthy)     0.0.0.0:8080->8080/tcp                                 akhq
b0beeb3926a1   confluentinc/cp-kafka:latest       "/etc/confluent/dock…"   3 hours ago   Up 53 minutes            9092/tcp, 0.0.0.0:39092->39092/tcp                     kafka2
01e549a6af17   confluentinc/cp-kafka:latest       "/etc/confluent/dock…"   3 hours ago   Up 53 minutes            9092/tcp, 0.0.0.0:29092->29092/tcp                     kafka1
22a21c8052df   confluentinc/cp-zookeeper:latest   "/etc/confluent/dock…"   3 hours ago   Up 3 hours               2888/tcp, 0.0.0.0:2181->2181/tcp, 3888/tcp             zookeeper
79a251ed09e2   postgres                           "docker-entrypoint.s…"   3 hours ago   Up 3 hours               0.0.0.0:5433->5432/tcp                                 postgres
PS C:\LB\LB2628-Kafka>
```

Kafka overview

Lund&Bendsen
developing developers



Kafka was originally built by LinkedIn to handle the growing amount of data streams in the company. Later it was open sourced. It's a distributed log optimized for high-throughput. Today it is used by man companies because of its flexibility and additional features build around it like **Kafka Streams**, a declarative DSL API to simplify stream-processing, **Kafka Connect**, an integration product that provides pluggable connectors to consume and produce to/from external data stores. The name Kafka was chosen by one of the founders of Kafka, Jay Kreps, because he liked the author Franz Kafka, and he thought the product was optimized for writing

Producer: A producer is a component or application that publishes (produces) events or messages to Kafka topics.

Consumer: A consumer is a component or application that subscribes (consumes) events or messages from Kafka topics.

Broker: Kafka clusters consist of one or more Kafka brokers, which are responsible for storing and serving events. Each broker is a Kafka server.

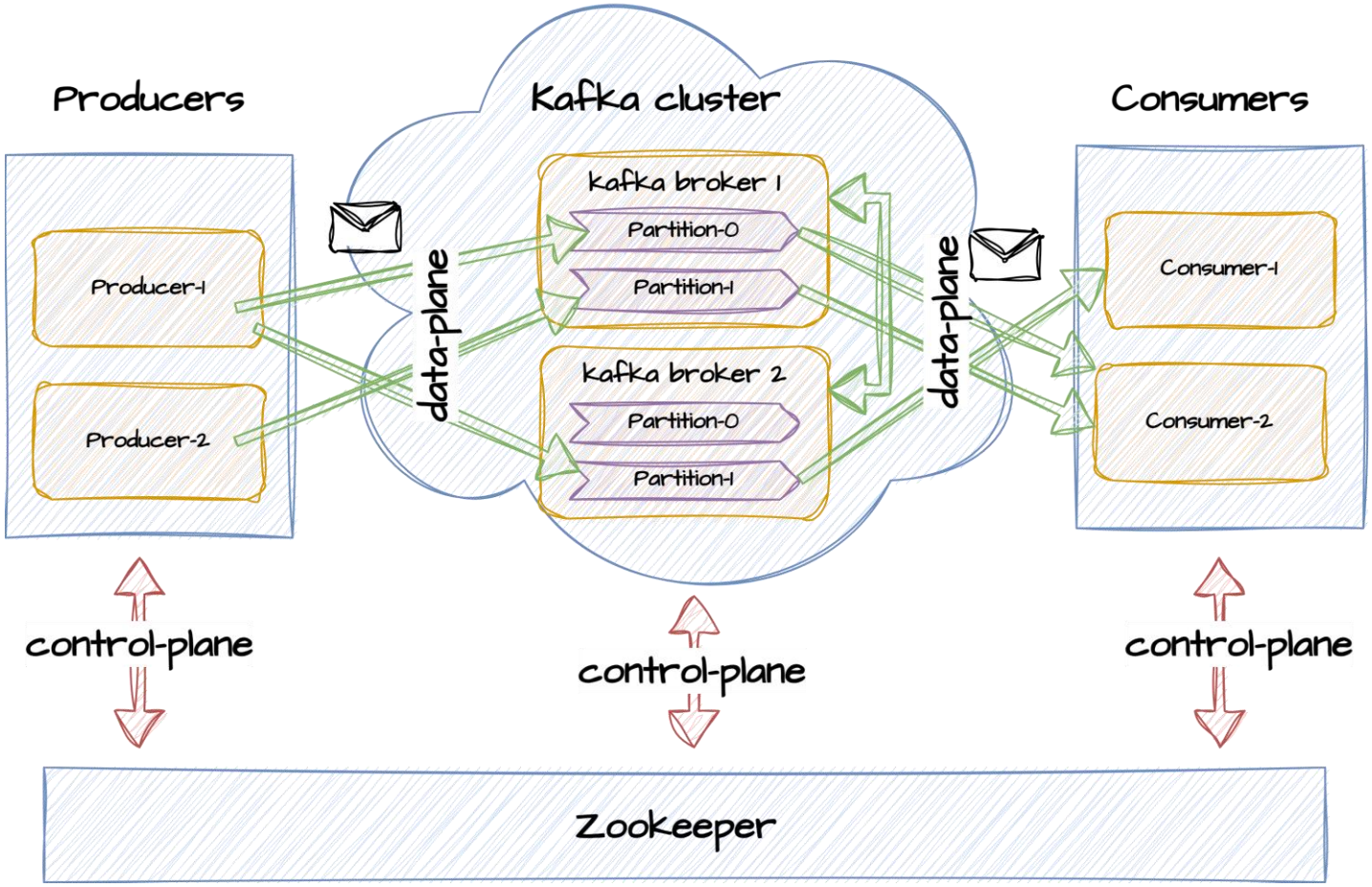
Topic: A topic is a logical channel or category to which events are published by producers and from which events are consumed by consumers. Topics are used to categorize and organize events.

Partition: Each Kafka topic can be divided into multiple partitions. Partitions allow for parallel processing of events within a topic. Each partition is an ordered, immutable sequence of events.

ZooKeeper: ZooKeeper is used for distributed coordination and management of Kafka brokers (control plane). However, Kafka is removing its dependency on ZooKeeper in newer versions in favor of the much faster protocol KRaft.

Kafka overview

Lund&Bendsen
developing developers



Who uses Kafka?

Lund&Bendsen
developing developers

- https://kafka.apache.org/powered-by

Many companies across various industries use Apache Kafka for real-time data streaming and event processing. Some well-known companies that have publicly disclosed their use of Kafka include:

LinkedIn: Kafka was originally developed by LinkedIn, and the company continues to use it extensively for various data processing needs.

Netflix: Netflix uses Kafka for real-time data streaming to enhance its streaming services and content recommendations.

Uber: Uber employs Kafka for real-time data processing and analytics to support its ride-sharing platform.

Airbnb: Airbnb uses Kafka to power its data infrastructure, enabling real-time data processing for various services.

Twitter: Twitter utilizes Kafka to handle a massive amount of real-time data generated by its social media platform.

Pinterest: Pinterest relies on Kafka for real-time data processing and analytics to enhance user experiences.

Walmart: Walmart uses Kafka to process and analyze large volumes of data to improve supply chain management and customer service.

Slack: Slack uses Kafka to support real-time messaging and communication services.

Goldman Sachs: Goldman Sachs utilizes Kafka for real-time data processing in financial services and trading.

The New York Times: The New York Times uses Kafka for real-time analytics and data processing to enhance its digital content delivery.

PostNord ☺

How kafka works

Lund&Bendsen
developing developers

Publishing Events (Producer):

Producers send events to Kafka topics. Each event is associated with a specific topic.

Producers typically batch events for efficiency and send them to Kafka brokers over a network connection.

Producers can choose to send events to a specific partition within a topic or allow Kafka to choose a partition using a partitioning strategy (e.g., round-robin, key-based).

Storing Events (Broker):

Kafka brokers receive and store events in their distributed log data store. Each event is appended to the appropriate partition.

Events are assigned sequential offsets within their partitions. Offsets serve as unique identifiers for events within a partition.

Replication:

Kafka provides data replication for fault tolerance. Each partition has multiple replicas distributed across different brokers.

Replication ensures that events are not lost if a broker fails. One replica is designated as the leader, and others are followers.

Producers send events to the leader replica, and followers replicate the data from the leader.

How kafka works

Lund&Bendsen
developing developers

Consuming Events (Consumer):

Consumers subscribe to Kafka topics and specify the partitions they want to consume from.

Kafka allows multiple consumers to subscribe to the same topic and partition, enabling parallel processing.

Consumers read events from their assigned partitions sequentially based on their offsets.

Retention and Cleanup:

Kafka retains events for a configurable period (retention period) or until a certain size threshold is reached.

Events that have been consumed are not immediately deleted but are marked for deletion based on their offsets.

Kafka uses a garbage collection process to reclaim disk space by deleting expired events.

Scaling:

Kafka is designed to be horizontally scalable. You can add more brokers to a Kafka cluster to increase capacity.

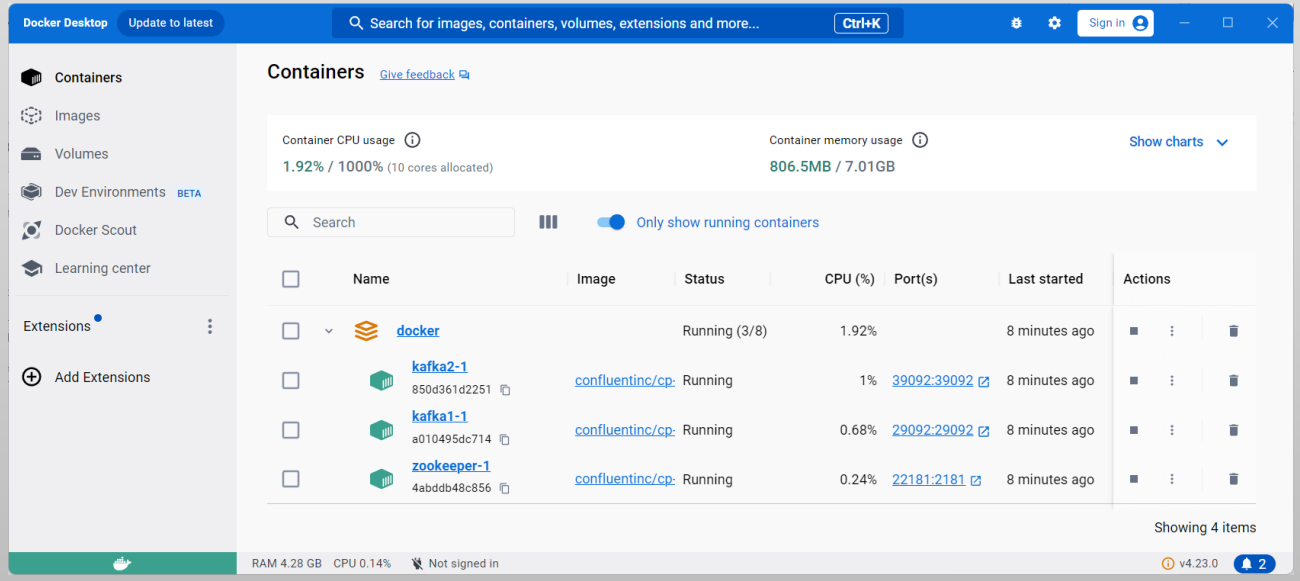
Partitions can be added to topics to distribute the workload and accommodate higher throughput.

Exercise

Lund&Bendsen
developing developers

2.2

- Open Docker Desktop
  - In the taskbar type ‘docker desktop’ and open the application
  - You should see something like the below window showing all running containers.
  - Docker desktop provides a user-friendly interface to see running containers.
  - Here you can start, stop and delete containers
  - Walk-through of the deployment and the containers



Kafka relational diagram

Lund&Bendsen
developing developers