Apache Kafka

Seydou coulibaly

January 17, 2024

Table of Contents

Table of Contents

Where's it come from ?

History

- Data silo at Linkedin company
- Co-developped by : Jay Kreps, Neha Narkhede and Jun Rao
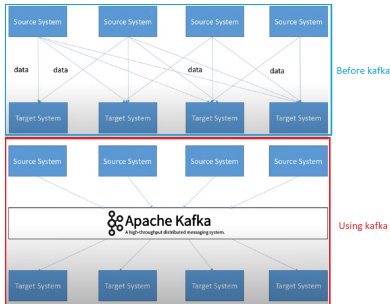- Open sourced since 2011
- Latest version : 3.6.1

What is it ?

kafka (distributed streaming plateform)

distributed : kafka runs as a cluster that can span multiple servers or even multiple data centers

streaming : allows for the developpement of realtime event-driven applications

note : kafka is behind modern day cloud application, users expect a real time experience

@ Stephane Maarek [5]

What is it ?

kafka (distributed streaming plateform)

<u>distributed</u> : kafka runs as a cluster that can span multiple servers or even multiple data centers

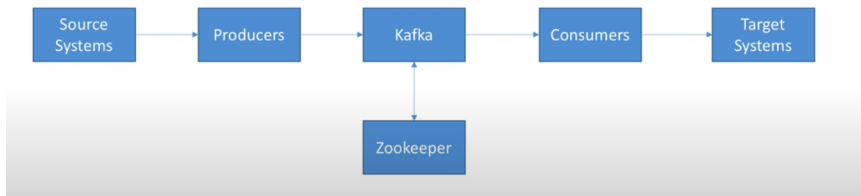<u>streaming</u> : allows for the developpement of realtime event-driven applications

Some use-cases

- Uber : location update in real time
- Netflix : Messaging system, activity tracking
- Linkedin: data gathering to make better connection recommendation in real time
- decoupling system dependencies

Table of Contents

४% kafka.

kafka core (ecosystem)

@ Stephane Maarek [5]

zookeeper is a tool for cluster management
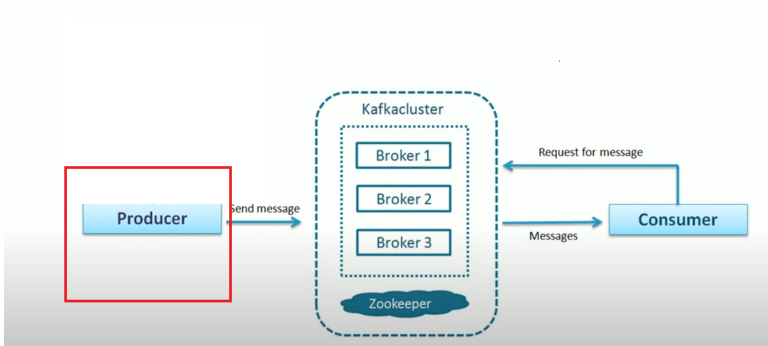
Table of Contents

Broker

Broker is the kafka server, it's receive and store data
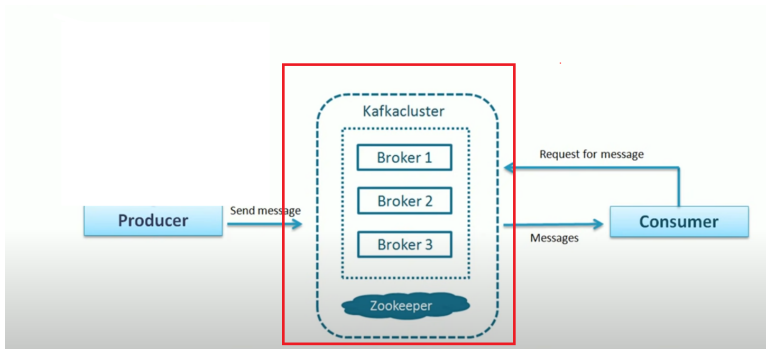
@ learning journal [4]

Producer

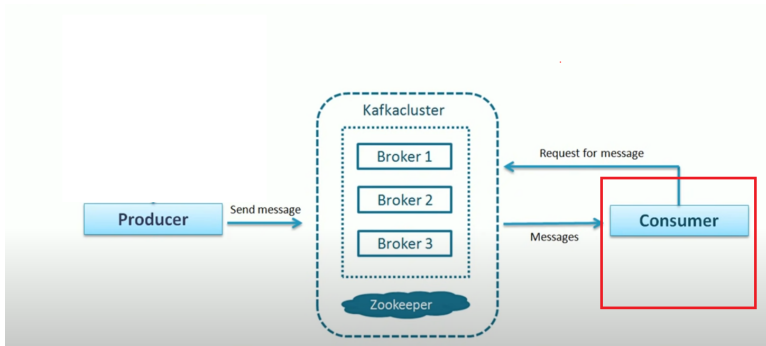## Producer sends stream of records or events

producer has property (acks = [0 | 1 | all])

@ learning journal [4]

Consumer

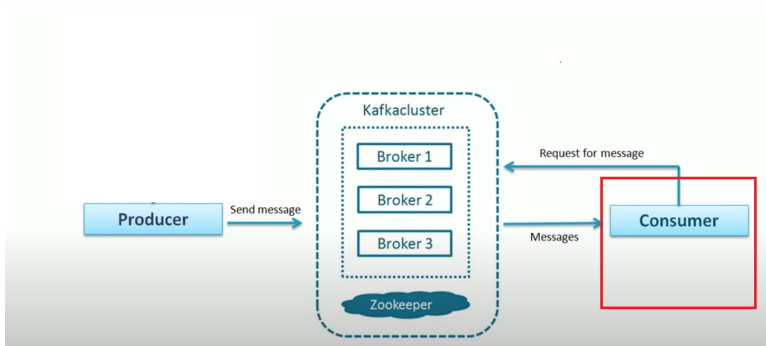## Consumer reads or ingests data from the broker

@ learning journal [4]

consumer group are a group of consumers acting as a single logical unit

Kafka

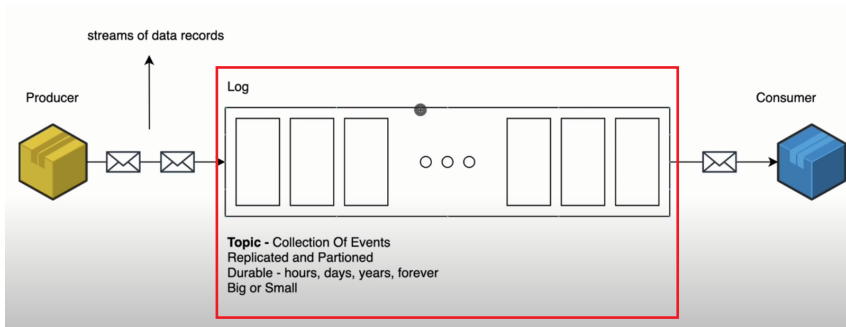## Message's transmission (topic)

@ learning journal [4]

<u>note</u> : messages are store in durable way with retention characteristics
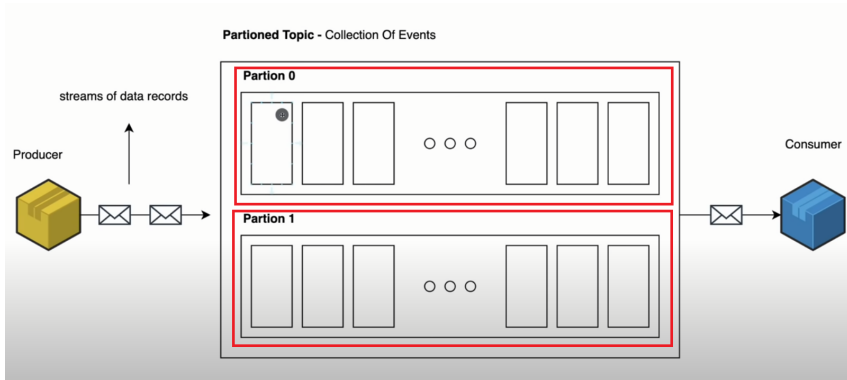
Topic

Topic (or logs) ordered sequence of events

@ amigoscode [1]

Partition

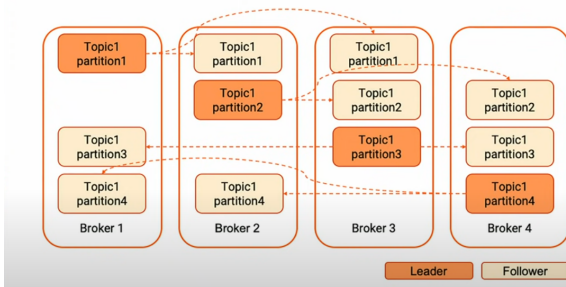## Partition reads or ingests data from the broker

@ amigoscode [1]

remarque : at the consumption, kafka read data in partition1, partition2, partiton3, ..., (etc)

Partition

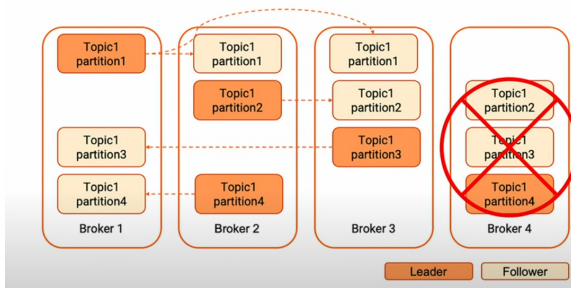## Notion of leader and replicat

Partition Leadership and Replication

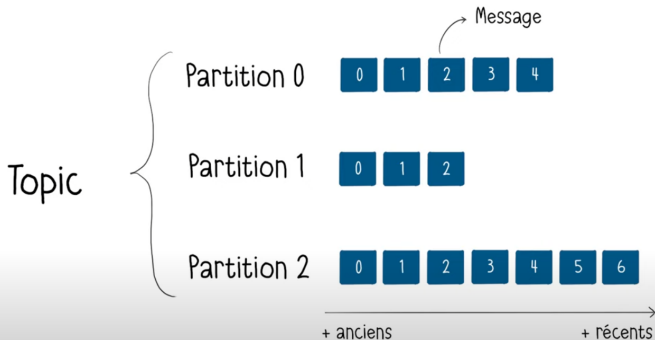@ devoxx FR [3]

Partition

Notion of leader and replicat

@ devoxx FR [3]

Partition

Offset

Some Characteristic of kafka

- Fast
- High accuracy
- Ordered
- Fault tolerant

Some Characteristic of kafka

- Fast
- High accuracy
- Ordered
- Fault tolerant

Advanced concept (API)

- kafka connect : for data integration
- kafka streams : data processing and transformation (grouping, enrich, filtering, aggregating)

Table of Contents

Quickstart

<u>click on the link</u> : apache kafka

Python library

Available libs

- confluent-kafka
- python-kafka

Comparison

	Confluent-kafka	python-kafka
Installation and configuration	-	+
Performance	+	-
Feature completeness	+	- (only essential feature)
Maintenability and support	+	-

Bibliography

📄 Nelson amigoscode.
Apache kafka for beginners, 2022.
https://www.youtube.com/c/amigoscode.

📄 Cookie connecte.
Apache kafka : comprendre l'essentiel en 4 minutes, 2022.
https://www.youtube.com/watch?v=kCLjBcKrbfM.

📄 Nicolas Romanetti Florent Ramiere, Jean-Louis Boudart.
Apache kafka, de haut en bas, 2019.
https://www.youtube.com/watch?v=KWAELycyxTc.

📄 Learning Journal.
Kafka tutorial, 2016.
https://www.youtube.com/@ScholarNest.

📄 Stephane Maarek.
Apache kafka for beginners, 2019.

End

Thank you for your attention !