Event streaming platform

Apache Kafka

M. Fernanda Sepúlveda @mf222

Jaime Yañez @JaimePata

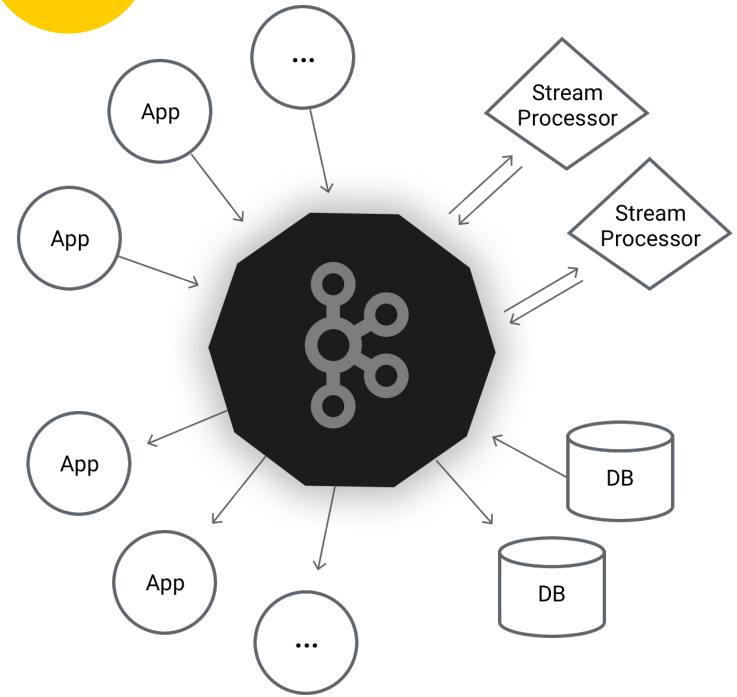
Índice

- ¿Qué es Apache Kafka?
  - ¿Por y para qué?
- ¿Cómo funciona?
- Requisitos no funcionales
- ¿Quienes usan Apache Kafka?
- Live Demo

1

¿Qué es Apache Kafka?

Y qué significa *Streaming Event Platform*



Es un message broker de código abierto

Es un sistema de streaming de mensajes en cola, usando el patrón Pub-Sub, distribuido, persistente y replicado, como si fuese un *transaction log*.

¿Por y para qué Kafka?

Funcionalidades

- Permite publicar y suscribirse a streams de datos.
- Almacena un historial de los datos que se han publicado. Funciona como una especie de log.
- Permite procesar un stream de datos mientras este ocurre.

Capacidades

- Crear pipelines de datos para transmitir información entre aplicaciones.
- Crear aplicaciones que reaccionen en tiempo real a un stream de datos.

2

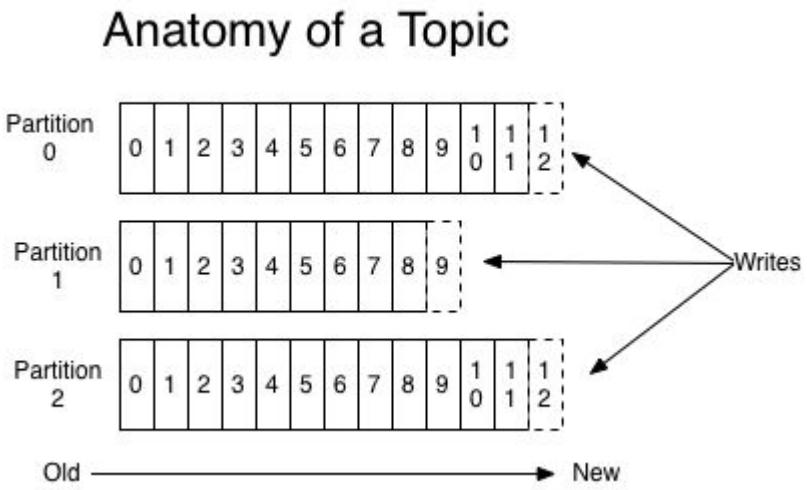
¿Cómo funciona?

Teoricamente, como streaming de datos

¿Cómo funciona?

4 APIS fundamentales

- Producer API: Publica información en un topic.
- Consumer API: Suscripción a topics.
- Streams API: Transforma “input streams” en “output streams” a uno o más topics.
- Connector API: Facilita la creación y utilización de productores y consumidores reutilizables, como por ejemplo, conectores con bases de datos relacionales.

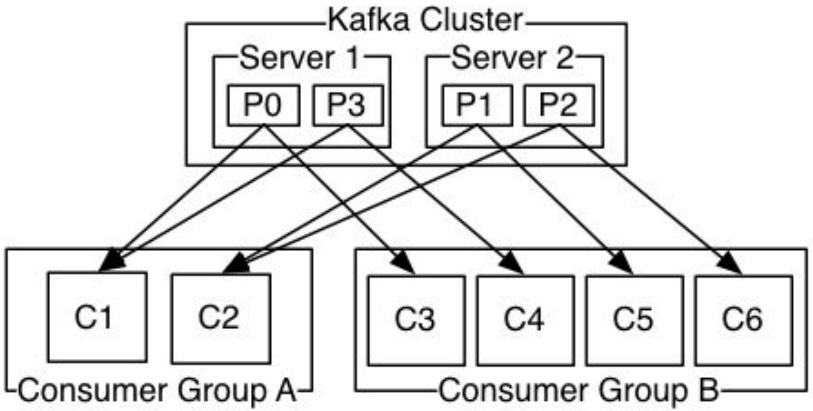


Topics and Logs

Categoría o nombre al cual los clientes pueden suscribirse.

Un topic puede enviar información a múltiples suscriptores.

Un suscriptor puede obtener información de múltiples topics.



Consumers

Los consumidores se dividen en grupos de consumidores de acuerdo al tópico al que están suscritos.

Si todas las instancias que están en un mismo grupo corresponden al mismo tópico entonces el balance de carga es óptimo.

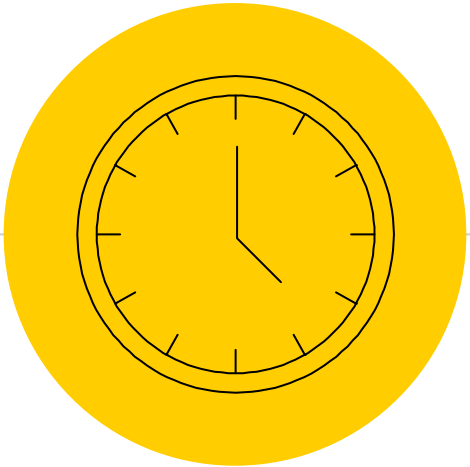
3

Requisitos no funcionales

¿Qué RNF abarca Apache Kafka?

Scalability

Sistema distribuido que escala fácilmente.



Durability

Mensajes enviados persisten en el disco. Permite replicación de forma sencilla.

Reliability

Réplica información. Puede manejar muchos suscriptores simultáneamente.

Performance

Buen performance tanto para publicadores como suscriptores. Paralelismo permite procesar grandes cantidades de información.

4

¿Quienes lo usan?

¿Es usado? ¿Es actual?

Up to 2014 it was all about Hadoop, then it was Spark. Now, it's Hadoop, Spark and Kafka. These are three equal peers in the data-ingestion pipeline in this modern analytic architecture.

“

Brian Hopkins, a vice president and principal analyst with Forrester Research

5 Live Demo

¿Preguntas?

¡Gracias por su atención!