Apache Kafka

Degendra Sivakoti

MSc IT | Islington College

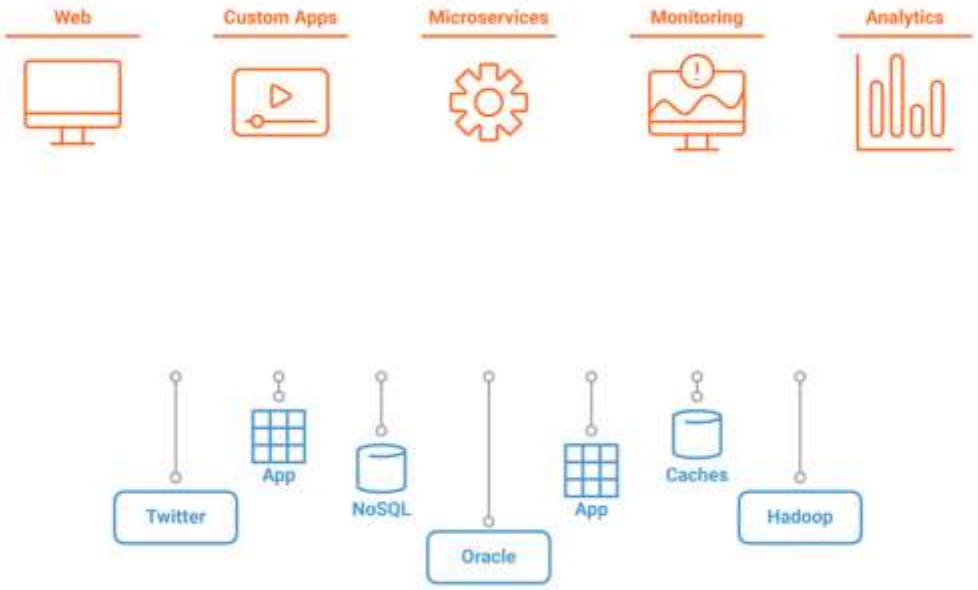
Apache Kafka

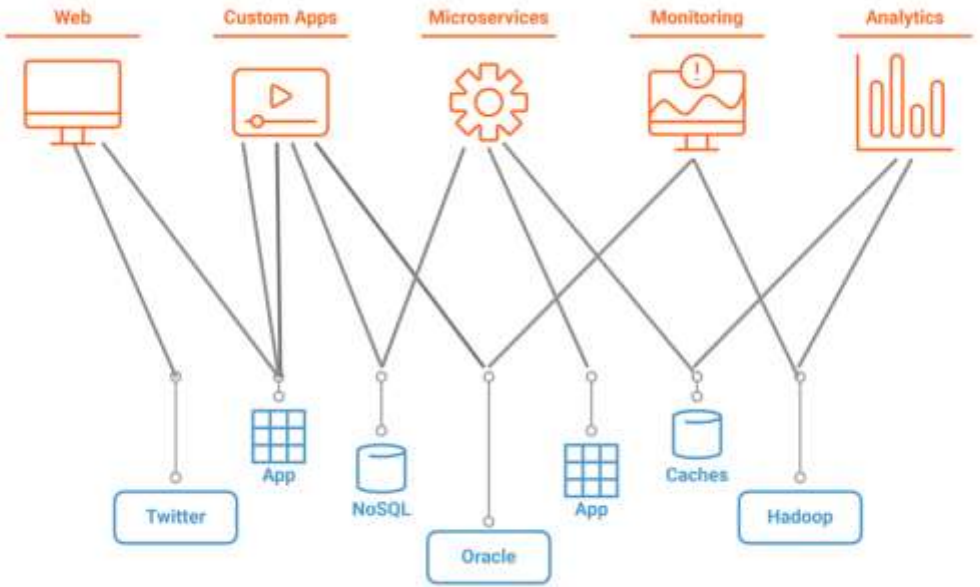
Apache Kafka® is *an open sourced distributed streaming platform*.

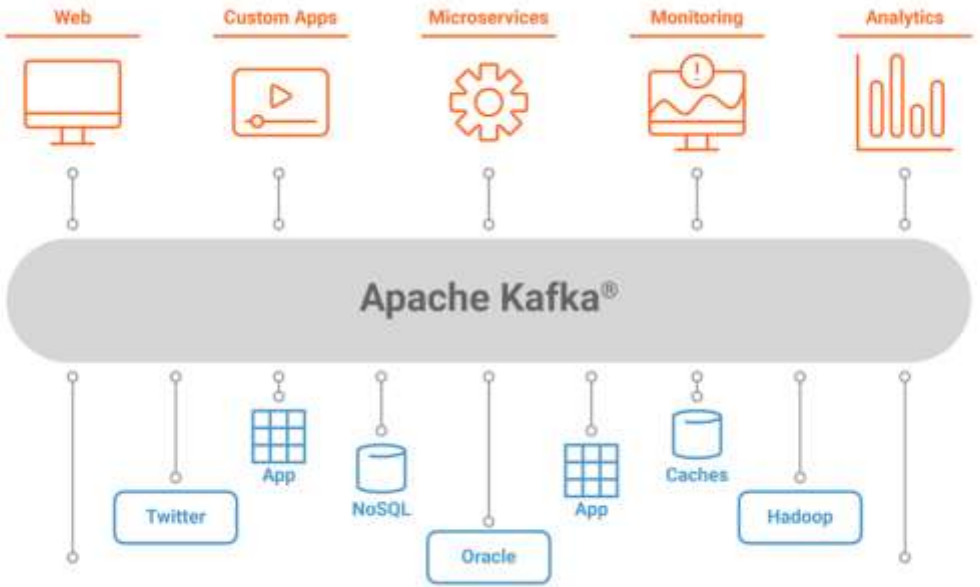
History

- Developed by LinkedIn / Microsoft
- Open sourced in early 2011
- Donated to Apache Software Foundation on 2012
- Latest release and stable version 2.5.0
- Written in Scala/ Java

Why Kafka?


Web
Custom Apps
Microservices
Monitoring
Analytics
Twitter
App
NoSQL
Oracle
App
Caches
Hadoop


Web
Custom Apps
Microservices
Monitoring
Analytics
Twitter
App
NoSQL
Oracle
App
Caches
Hadoop

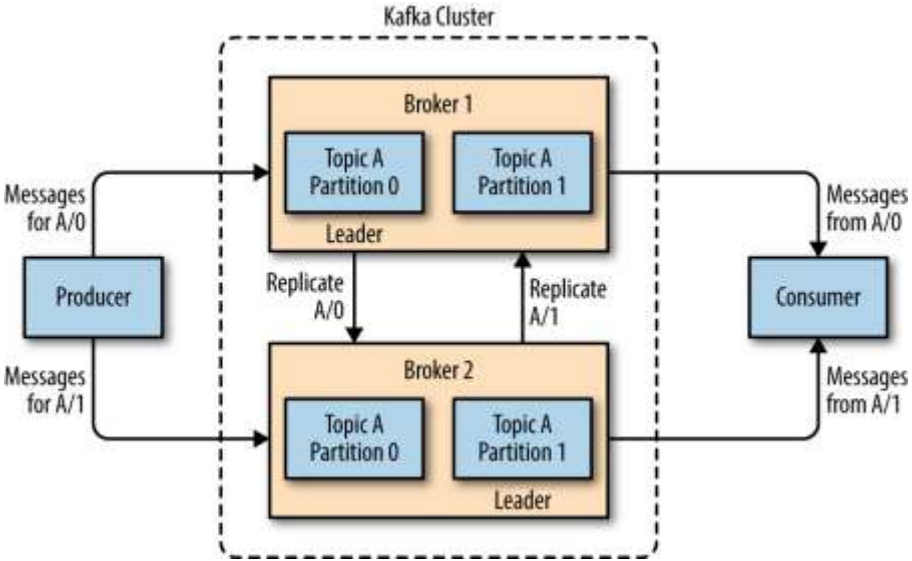

Web
Custom Apps
Microservices
Monitoring
Analytics
Apache Kafka®
Twitter
App
NoSQL
Oracle
App
Caches
Hadoop

Major Features

- Kafka as Messaging System
- Kafka for Stream Processing

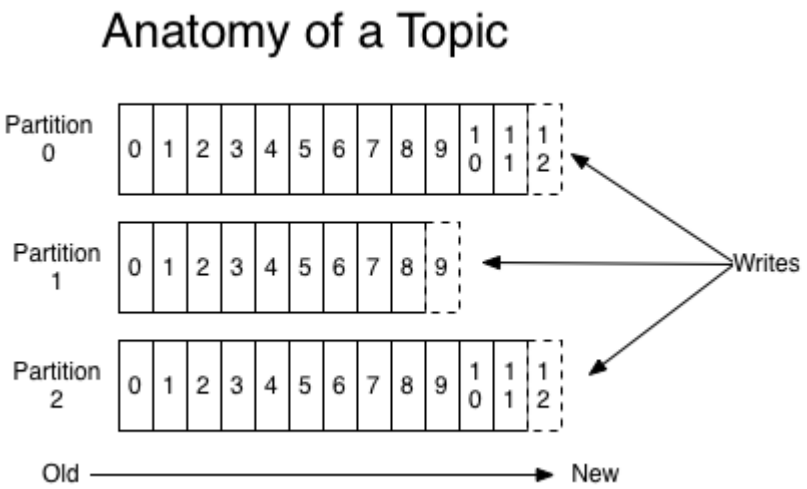
Core Components

- Zookeeper
  - Core dependency as of version 2.5.0
- Broker
  - Kafka server
- Producer
  - Produces message to topic partition
- Consumer
  - Listens to topic partition
- Topic
  - Partition



Topic

- A topic can have one or more partitions
- Topics are always multi-subscriber
- **Partitions (logs)**:
  - One leader and zero or more followers
  - Distributed

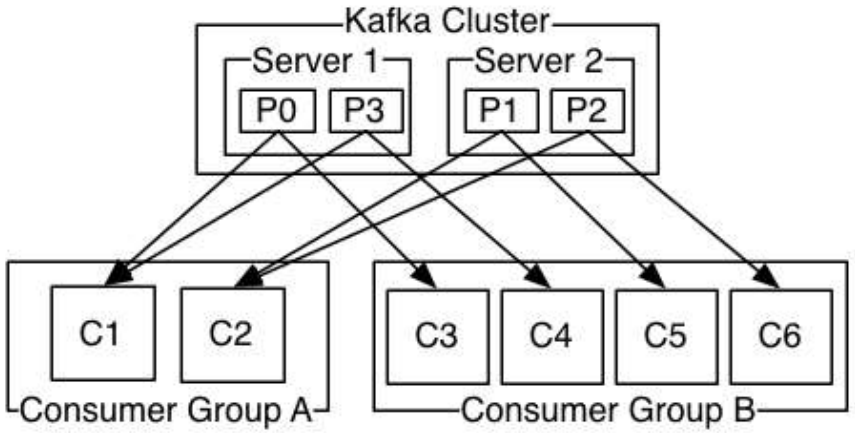


Producer

- Produce message to a topic
- Select topic partition
- Normally, uses round-robin technique to balance load

Consumer

- Consumer listens to one or more topics
- Consumer belongs to a **consumer group**
- Only on consumer from a consumer group can consume message from a topic partition.
- **Rebalancing**
  - More consumers in a group than partition means **idle** consumers



Example

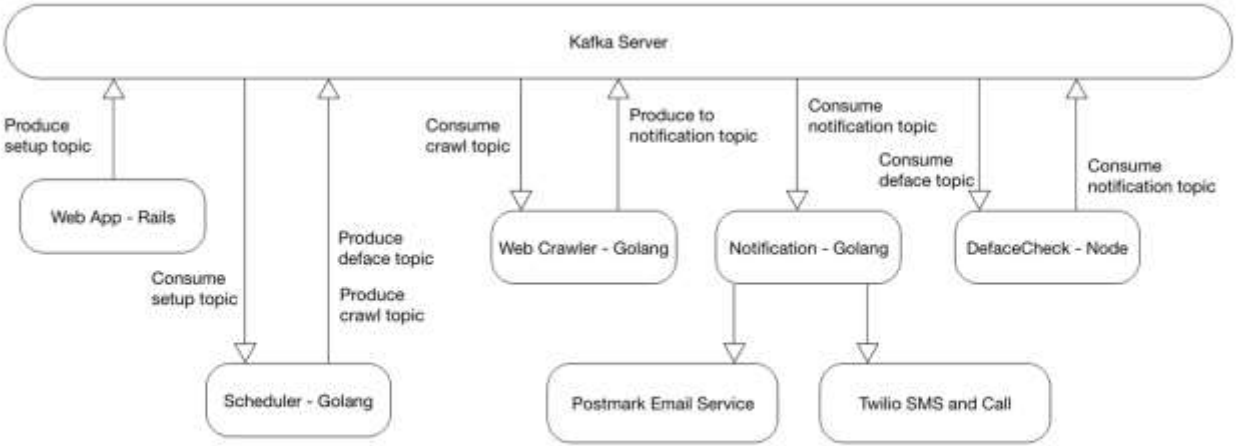
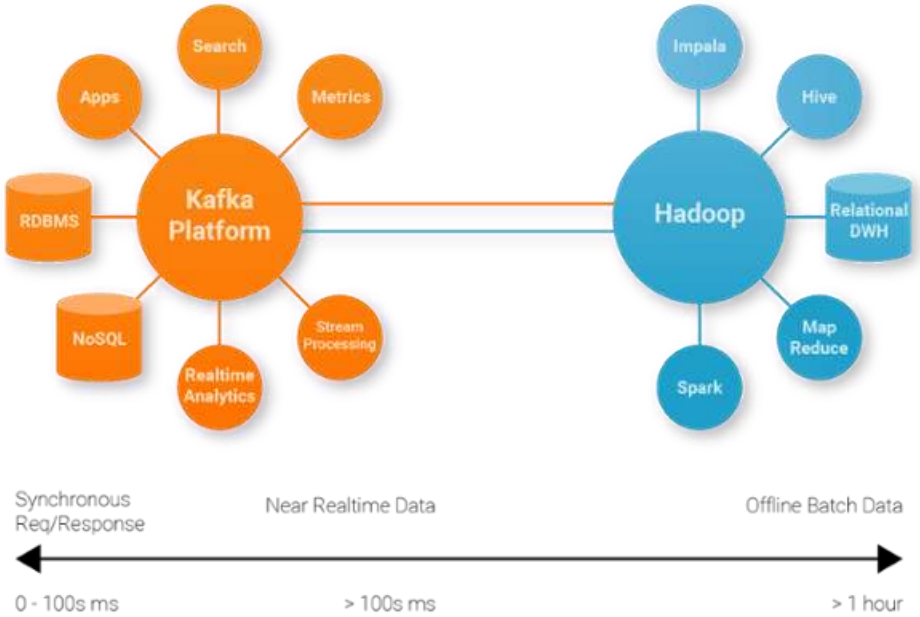


Fig. SiteHawk.io - microservices architecture

Kafka and BigData

- Process existing big data
  - Sitting on HDFS
  - Sitting in a database
- How does new data get into your cluster?
  - New log from web servers
  - New sensor data from IoT systems
  - New stock trades
- Use for data ingestion in Hadoop system

Kafka and Hadoop Ecosystem



Questions?

Thank you!