

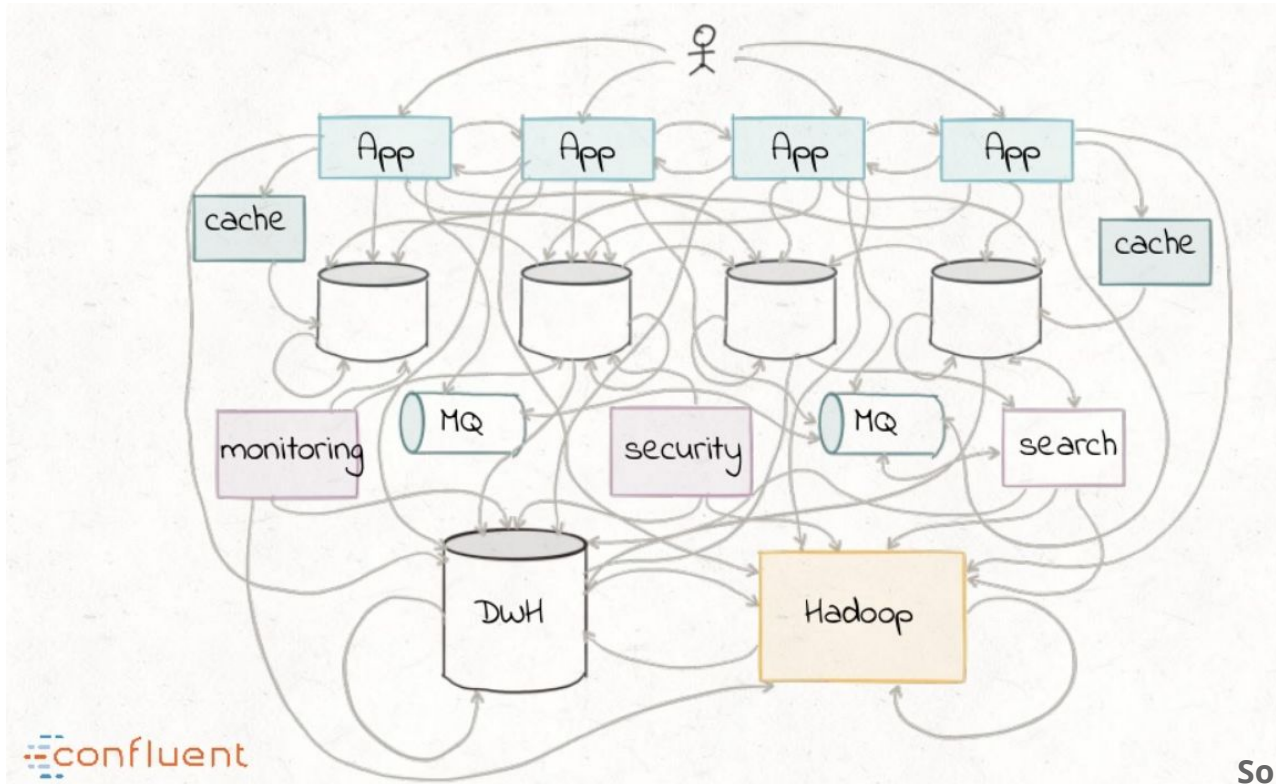
# Amazon Managed Stream for Kafka



Cloud Computing 2019  
René Gómez Londoño  
Ivan Salfati

---

# Motivation



Source: [Confluent](#)



**Mathias Verraes**

@mathiasverraes

Follow



There are only two hard problems in distributed systems:

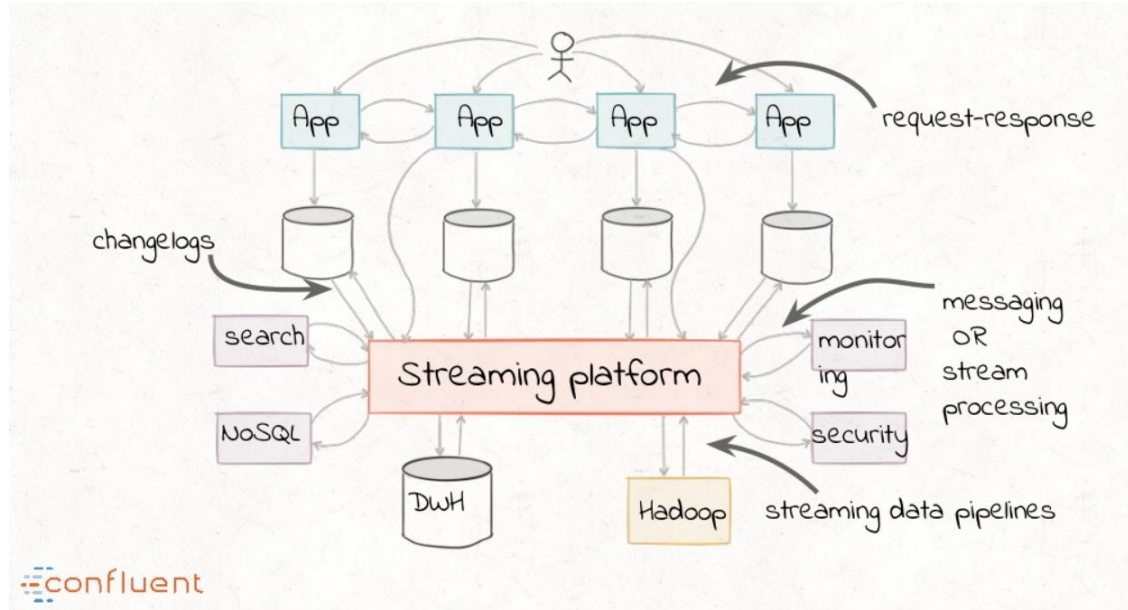
1. Guaranteed order of messages
2. Exactly-once delivery

1:40 PM - 14 Aug 2015

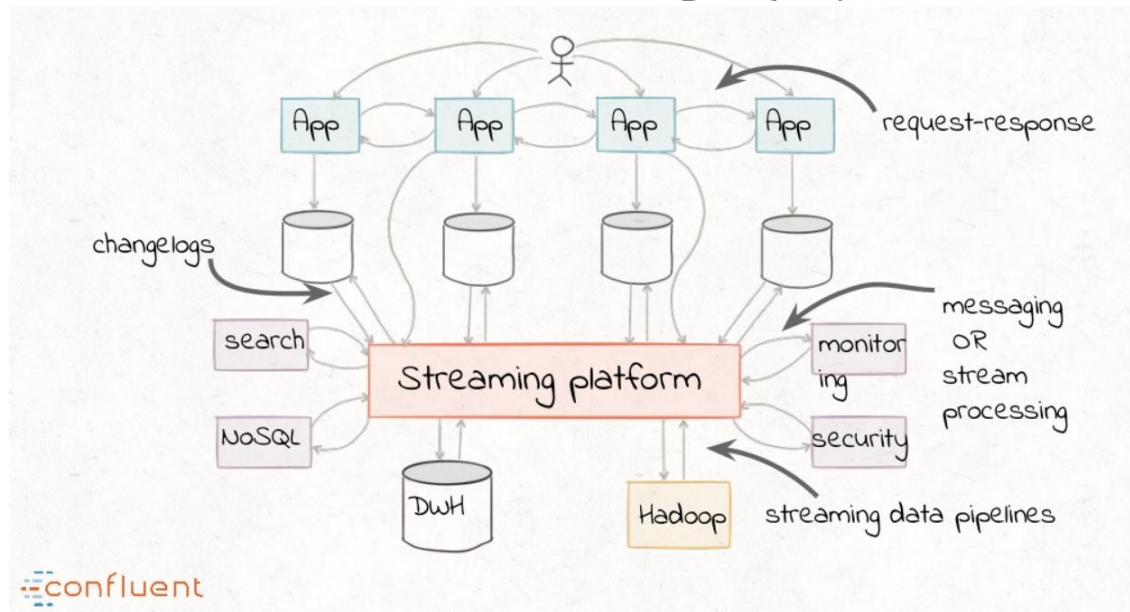
7,196 Retweets 5,658 Likes



# Stream processing

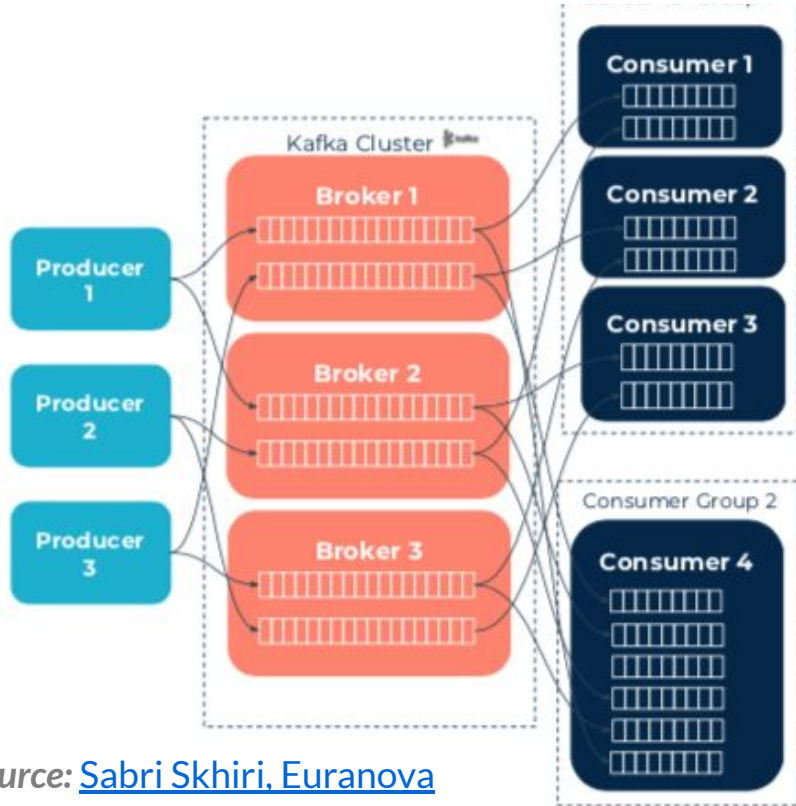


# Service choreography



***Dancers dance following a global scenario without a single point of control***

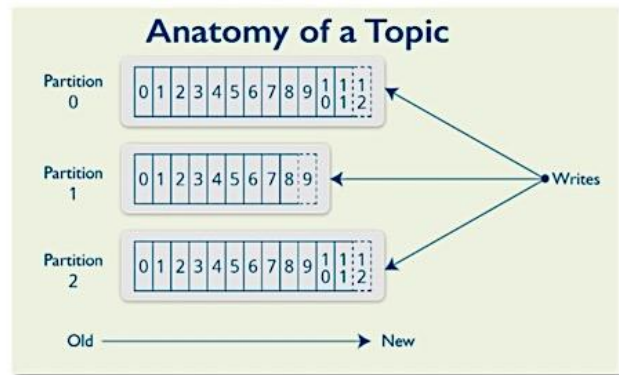
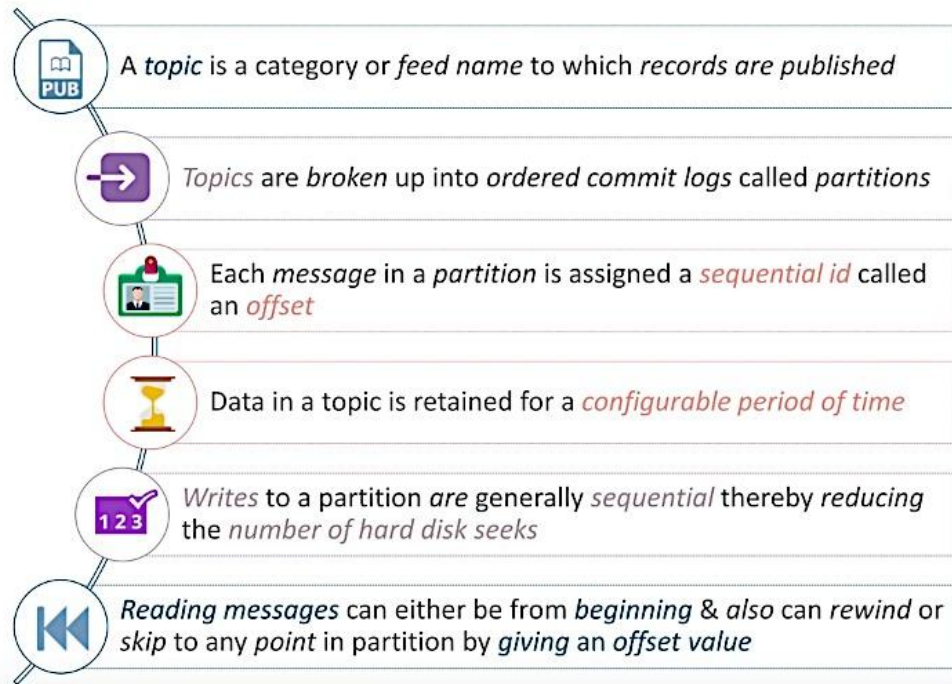
# Why Kafka?



Source: [Sabri Skhiri, Euranova](#)

- **Partitioning** (topics, partitions)
- **Replication** (the **log** is distributed by Kafka)
- **Fault tolerant** (the **log** provides retention)
- **Elastic Scaling** (pub/sub, message queue)

# Concepts

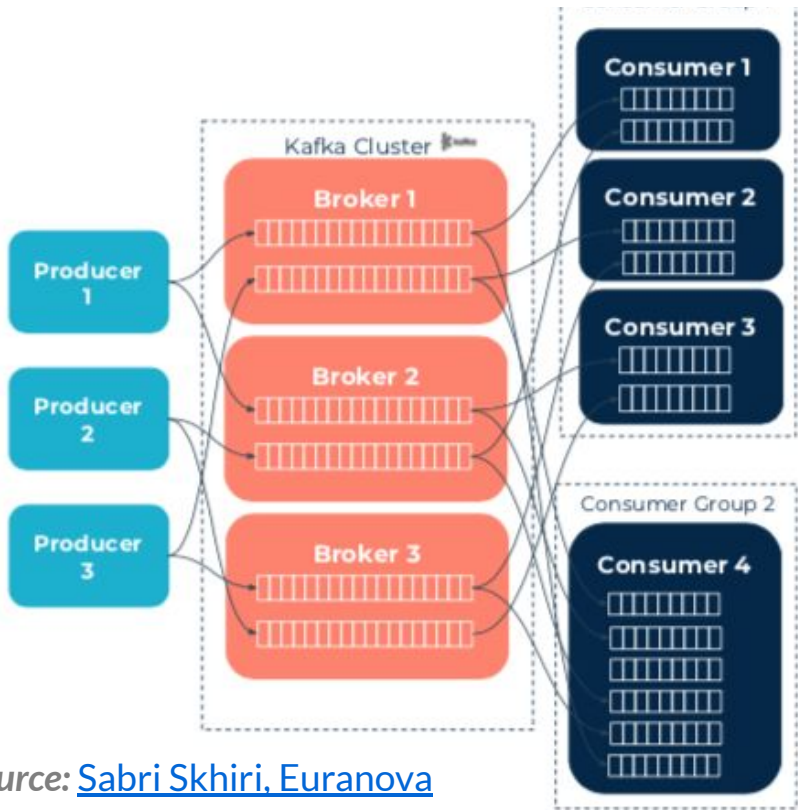


Source: Udemy



# Benefits

- Provides **low latency** & **high throughput**
- Kafka acts as a **buffer** so your **systems won't crash**

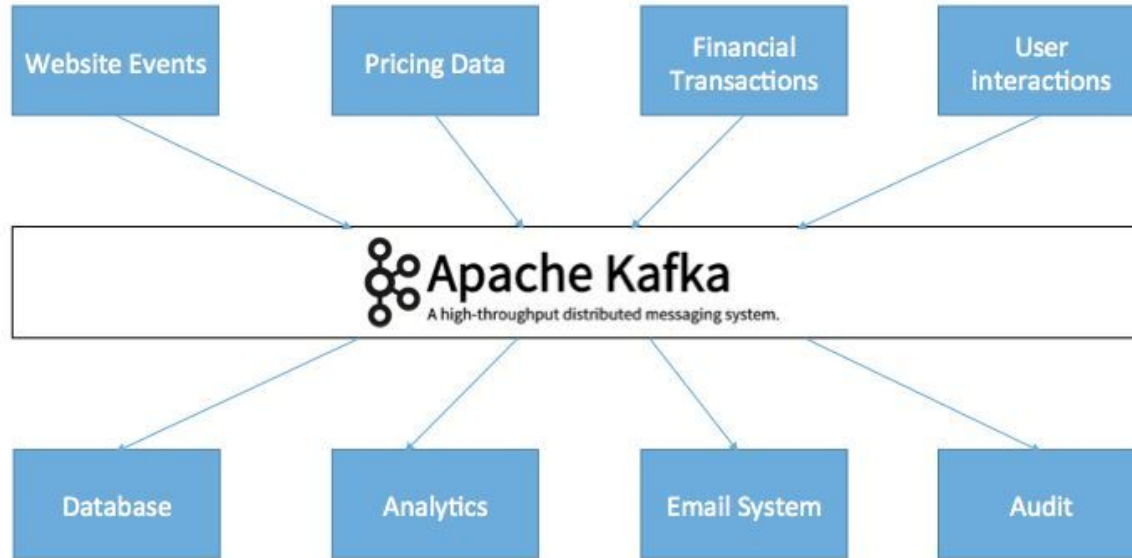


Source: [Sabri Skhiri, Euranova](#)

*Do you want to know more about the log?*  
*The Log: What every software engineer should know  
about real-time data's unifying abstraction*



# Kafka reduces the need for multiple integrations



*Kafka is **difficult to deploy** and manage **without skilled DevOps & Site Reliability Engineers***

# Amazon Managed Streaming for Kafka

Amazon MSK is a **fully managed service** that makes it **easy** for you to **build** and run **applications** that use Apache **Kafka** to process streaming data



# MSK - Benefits

## **Performance tuning**

- Architectures and pricing models tested by other companies

## **Storage options**

- Ephemeral storage (instance store)
- Amazon Elastic Block Store

## **Upgrades**

- Rolling or in-place upgrade
- Downtime upgrade
- Blue/green upgrade

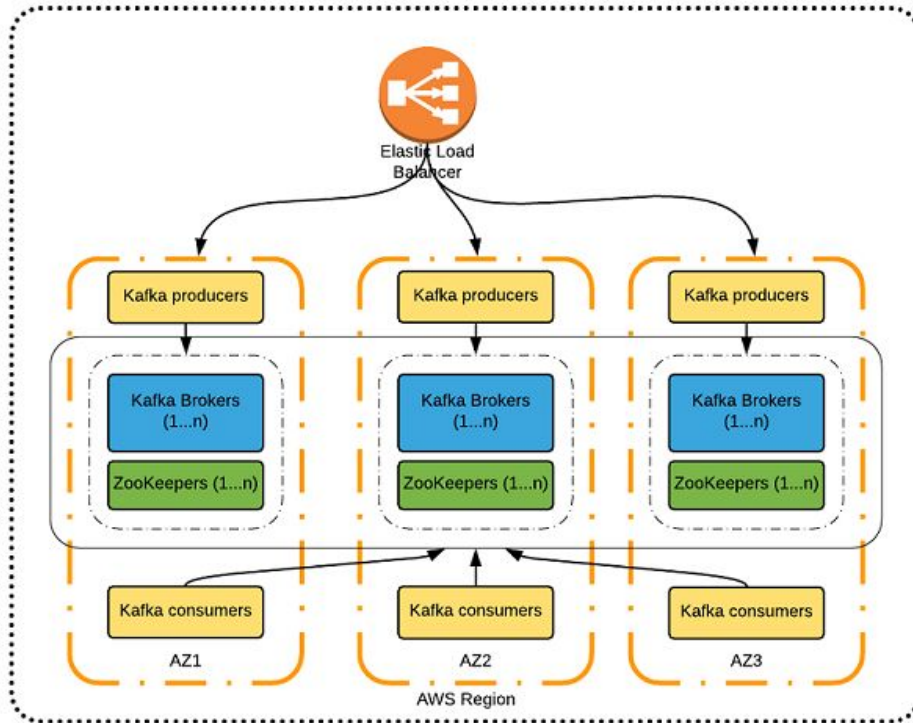
## **Monitoring, security and best practices**

# Best practices (I)

- Kafka provides:
  - High Performance
  - Scalable Solutions
  - Different type instances and storage options
- Not trivial to select the most optimal topology

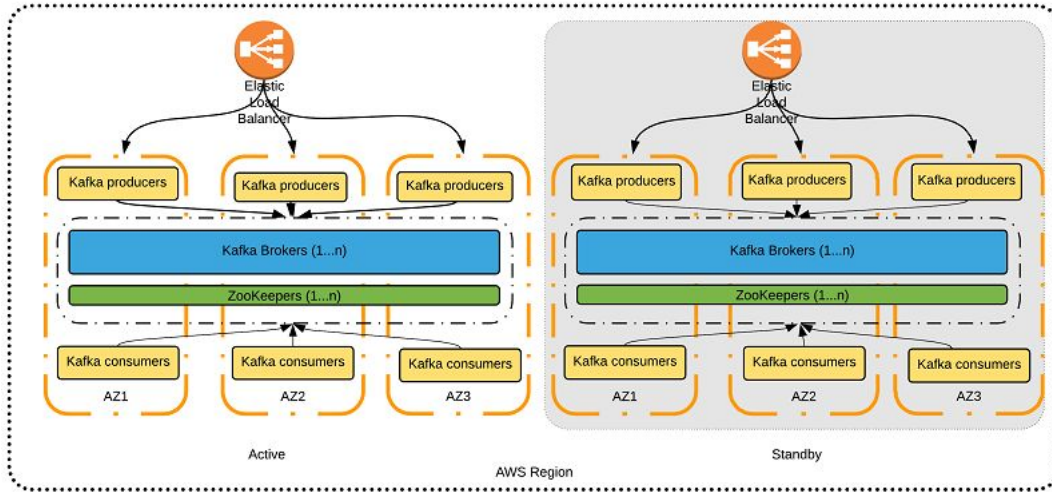


# Best practices




- Deployment and patterns
- Storage
- Instance types
- Networking
- Upgrades
- Performance tuning
- Monitoring
- Security
- Backup and restoration

# Best practices



- Deployment and patterns
- Storage
- Instance types
- Networking
- Upgrades
- Performance tuning
- Monitoring
- Security
- Backup and restoration

# Other cloud providers



**Kafka**  
[Kafka \(Google Click to Deploy\)](#)  
Estimated costs: \$24.67/month

Open source distributed stream processing platform

**LAUNCH ON COMPUTE ENGINE**

CONFLUENT CLOUD

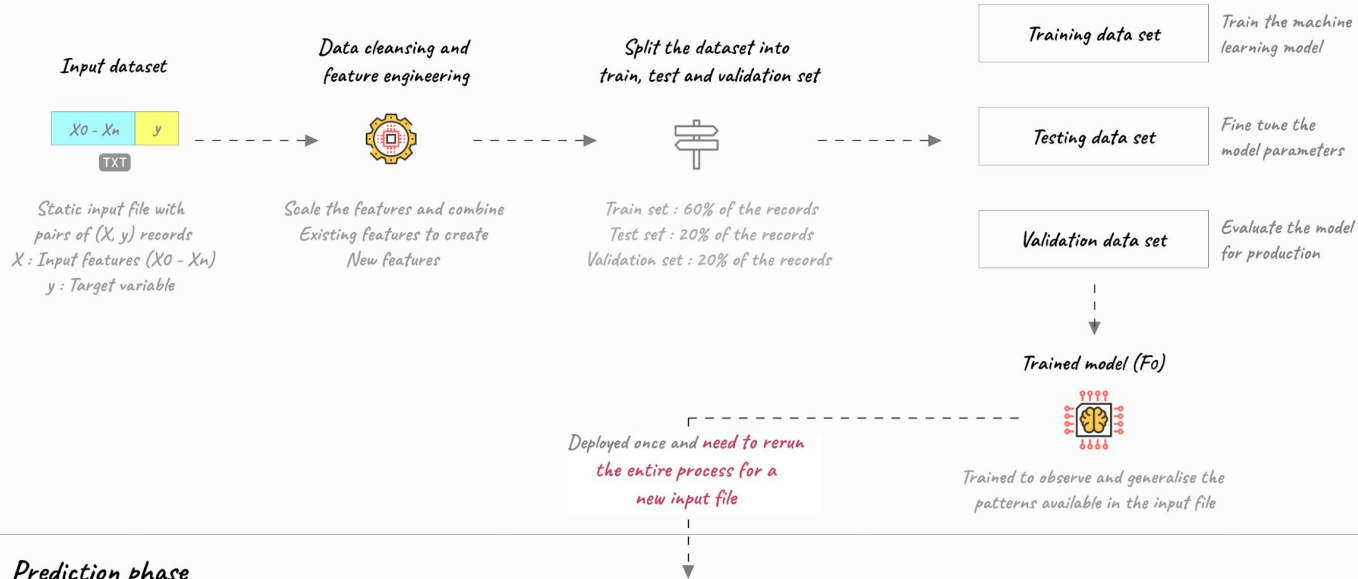
**Apache Kafka  
re-engineered for the  
cloud**



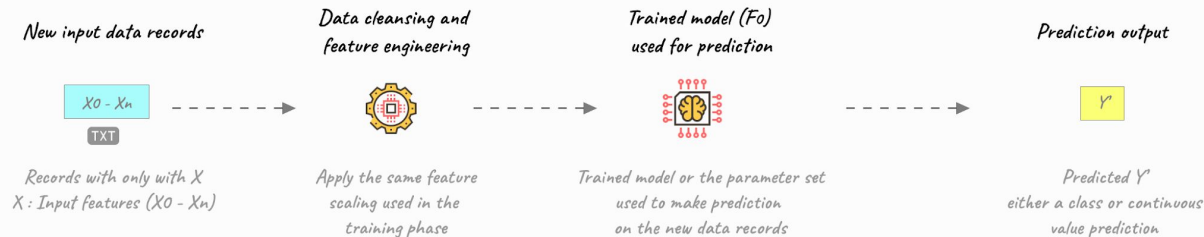
DATA HAS A BETTER IDEA

# Traditional machine learning process (batch)

## Training phase

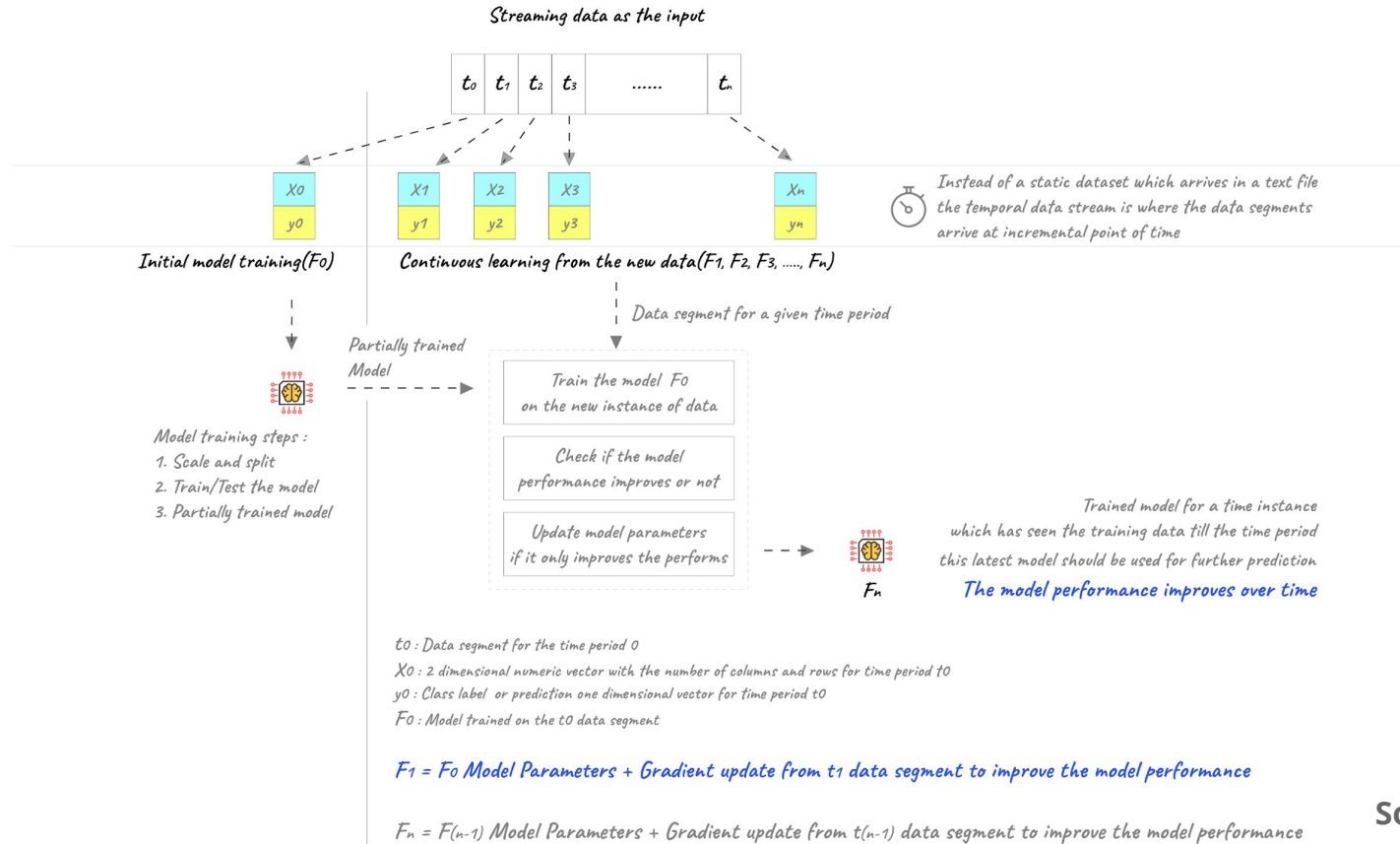


## Prediction phase



Source: [Vidhya](#)

# Online machine learning process (simplified flow)



Source: [Vidhya](https://www.vidhya.com)

Q/A

# References

- **AKF, Apache Kafka Documentation.** Available at: <https://kafka.apache.org/intro>
- **AWS, 2018. What is Streaming Data? – Amazon Web Services (AWS).** Amazon Web Services, Inc. Available at: <https://aws.amazon.com/streaming-data/>
- **Byzek, Y. et al., 2018. Confluent Blog: Apache Kafka Best Practices, Product Updates & More.** Confluent. Available at: <https://www.confluent.io/blog>
- **Data Artisans, 2017. data Artisans - Apache Flink.** Data Artisans. Available at: <https://data-artisans.com>
- **Fowler, M., 2017. What do you mean by “Event-Driven”?** martinfowler.com. Available at: <https://martinfowler.com/articles/201701-event-driven.html>
- **Apache Kafka.** Sharma, Gómez. Advanced Databases. 2018. [http://cs.ulb.ac.be/public/\\_media/teaching/infh415/student\\_projects/2019/kafka.pdf](http://cs.ulb.ac.be/public/_media/teaching/infh415/student_projects/2019/kafka.pdf)
- **Helland, P., 2015. Immutability Changes Everything.** In 7th Biennial Conference on Innovative Data Systems Research (CIDR). 7th Biennial Conference on Innovative Data Systems Research (CIDR). Salesforce, pp. 1–6.
- **Narkhede, N., Shapira, G. & Palino, T., 2017. Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale,** “O’Reilly Media, Inc.”
- **Netflix Technology Blog, 2016. Evolution of the Netflix Data Pipeline – Netflix TechBlog.** Available at: <https://medium.com/netflix-techblog/>
- **Twitter Engineering, Handling five billion sessions a day** – in real time. Twitter Engineering Blog
- **Uber Engineering, Kafka Archives** | Uber Engineering Blog. Uber Engineering Blog. Available at: <https://eng.uber.com/tag/kafka/>
- **Kafka at Scale** | Kai Waehner blog. Available at : [http://www.kai-waehner.de/blog/2018/05/09/deep-learning-at-extreme-scale-%E2%80%A8with-apache-kafka-open-source-ecosystem/netflix\\_linkedin\\_apache\\_kafka\\_at\\_scale/](http://www.kai-waehner.de/blog/2018/05/09/deep-learning-at-extreme-scale-%E2%80%A8with-apache-kafka-open-source-ecosystem/netflix_linkedin_apache_kafka_at_scale/)