Kafka

To get data from so many prodeucers and to publish data to so many consumers, we use kafka. Kafka is a streaming service.

Kafka is used to handle live streaming data pipeline.

It is used mainly to:

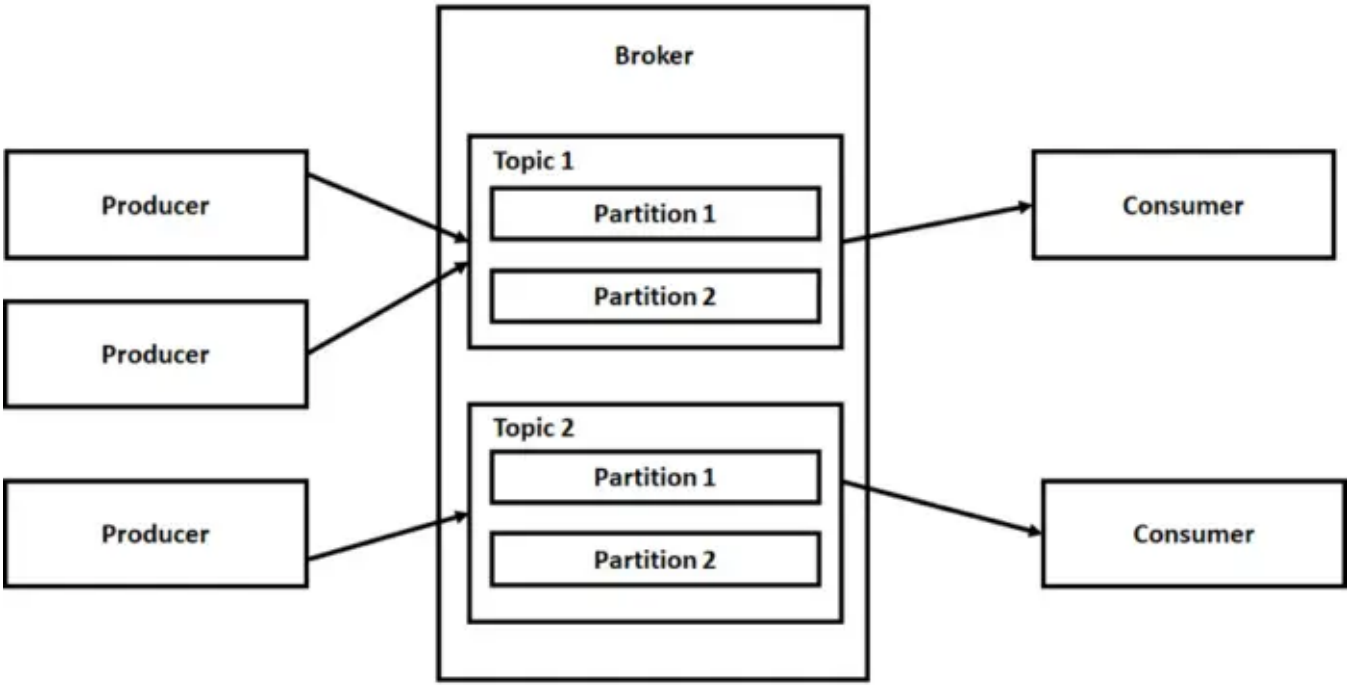
Publish streams of events, continuous import/export of data.

Store stream events.

Process stream of events to occur.

Why we are using kafka, instead we can use mongoDB ?

To receive live streaming data we are using kafka because directly receiving data to mongoDB is not possible. To receive streaming data in database we need to make connection to the database, which will take time again and again.



For pre-processing the data producer write a script and send the produced data in required format to streaming platform(Kafka). Kafka stores the streaming data for some amount of time. Consumer will fetch the data from kafka database by writing a script and stores the data to a database.

By-default kafka stores data for **1 week**. And we can extend the retaintion time. We need to consume the data within that period of time, otherwise kafka automatically deletes the data.

From that database **data Science** or **data analytics** team will collect the data and do some analysis.

Using **kafka** we can setup the **ETL pipeline**. For the above steps from collecting data to storing it to a database in required manner, **big data** team or **data engineering** team is responcible.

Both producer and consumer interact with kafka topics to produce and consume data.

Topics : Kafka topics are virtual groups or logs that hold messages and events in a logical order, allowing users to send and receive data between kafka servers with ease.

Topics are dedicated to every single operation and kafka can have multiple topics. Each Topic has some schema to receive data with device id.

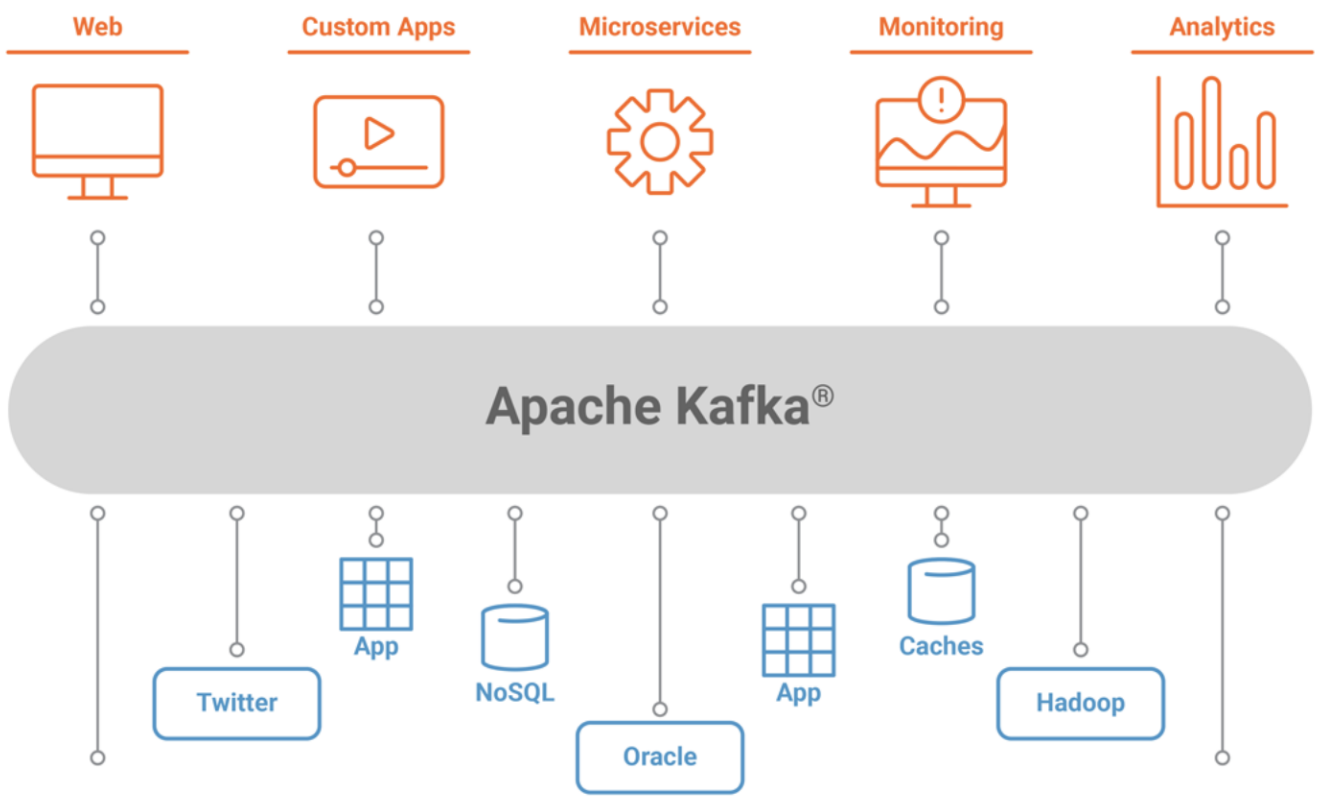
We are using **Confluent** virtual machine to use streaming service of kafka, which is a vendor to manage the operations of kafka with high performance. Confluent is a SAAS, which provide kafka as a service.

Steps in confluent:

1. Create a cluster. (API ket, API Secret, Bootstrap server)
2. Create topics
3. Create Schema API key, API Secret and take Schema End point url -->> To maintain the structure of the data centrally

Kafka can be used directly to create pipeline

Source(devices) >>> Kafka >>> mongoDB



github: https://github.com/saisubhasish (https://github.com/saisubhasish)

In [ ]:

```
1
```