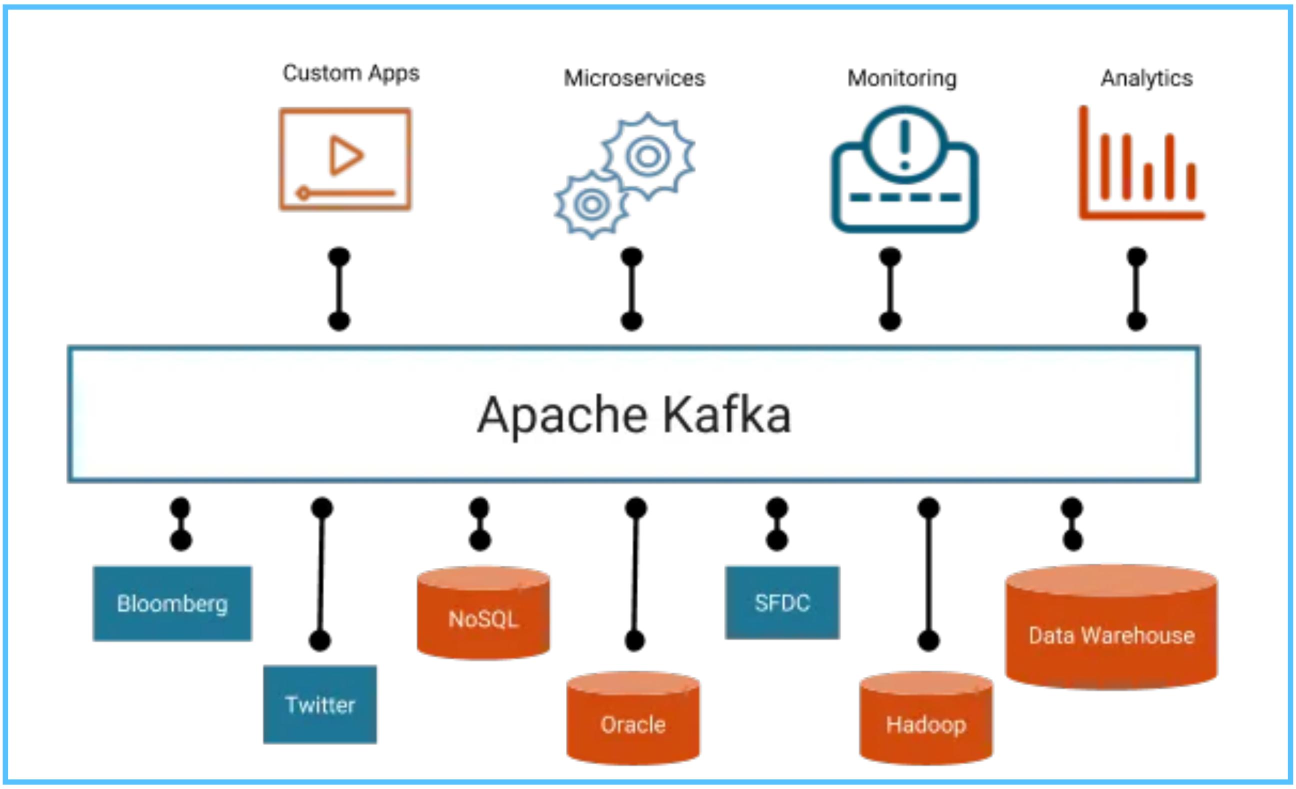
Overview on Apache Kafka

- *Apache Kafka is a popular **event streaming platform** used to collect, process, and store streaming event data or data that has no discrete beginning or end.*
- *It is an **Open source** with high performance and low latency.*
- *Apache Kafka makes possible a new generation of distributed applications capable of scaling to handle billions of streamed events per minute.*
- *Streaming data is data that is continuously generated by thousands of data sources, which typically send the data records in simultaneously.*



- *It is a distributed data store optimized for ingesting and processing streaming data in real-time.*

- *It combines **messaging**, **storage**, and **stream processing** to allow storage and analysis of both historical and real-time data.*
- *A streaming platform will handle constant influx of data, and process the data sequentially and incrementally.*
- *Kafka provides three main functions to its users:*
    - *Publish and subscribe to streams of records*
    - *Effectively store streams of records in the order in which records were generated*
    - *Process streams of records in real time*
- *Apache Kafka is used by many major companies such as Netflix, Apple, Uber, Spotify, and LinkedIn.*

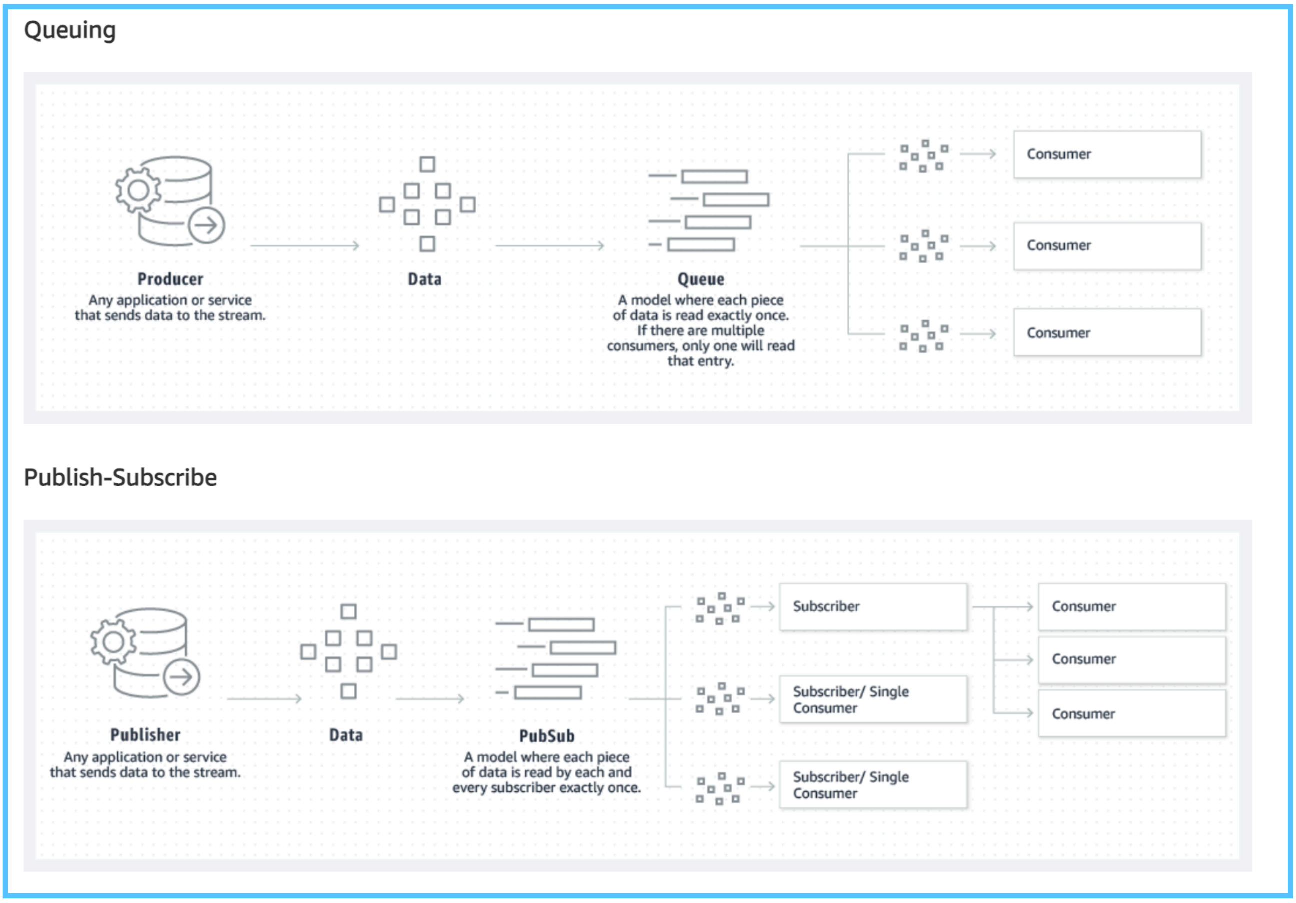
What Can We Do With Apache Kafka

Apache Kafka is used in many use cases like

- *Fraud and anomaly detection*
- *Recommendation engine*
- *Monitoring / Metrics*
- *Activity Tracking*
- *Integrate systems*
- *Real-time stream processing*

How does Apache Kafka work

*Kafka combines two messaging models, **queuing** and **publish-subscribe**, to provide the key benefits of each to consumers. Kafka enables streaming event processing through five core functions:*


Queuing
Producer
Any application or service that sends data to the stream.
Data
Queue
A model where each piece of data is read exactly once. If there are multiple consumers, only one will read that entry.
Consumer
Consumer
Consumer
Publish-Subscribe
Publisher
Any application or service that sends data to the stream.
Data
PubSub
A model where each piece of data is read by each and every subscriber exactly once.
Subscriber
Subscriber/ Single Consumer
Subscriber/ Single Consumer
Consumer
Consumer
Consumer

Publish

- *A data source can publish or place a stream of data events into one or more Kafka topics, or groupings of similar data events.*
- *For example, you can take data streaming from an IoT device—say a network router—and publish it to an application that does predictive maintenance to calculate when that router is likely to fail.*

Consume

- *An application can subscribe to, or take data from, one or more Kafka topics and process the resulting stream of data.*
- *For example, an application can take data from multiple social media streams and analyze it to determine the tenor of online conversations about a brand.*

Process

- *Kafka Streams API can act as a* ***stream processor****, consuming incoming data streams from one or more topics and producing an outgoing data stream to one or more topics.*

Connect

- *You can also build reusable producer or consumer connections that link Kafka topics to existing applications.*
- *There are hundreds of existing connectors already available, including connectors to key services like Dataproc, BigQuery, and more.*

Store

- *Apache Kafka provides durable storage.*
- *Kafka can act as a "source of truth," being able to distribute data across multiple nodes for a highly available deployment within a single data center or across multiple availability zones.*

Benefits of Apache Kafka

Kafka is open source

- *This means its source code is freely available to anyone to take, modify, and distribute as their own version, for any purpose.*
- *There are no licensing fees or other restrictions.*
- *Kafka also benefits from having a global community of developers.*

Scale and speed

- *Kafka not only scales with ever-increasing volumes of data, but provides that data across the business in real time.*

Scalable

- *Kafka's partitioned log model allows data to be distributed across multiple servers, making it scalable beyond what would fit on a single server.*

Fast

- *Kafka decouples data streams so there is very low latency, making it extremely fast.*

Durable

- *Partitions are distributed and replicated across many servers, and the data is all written to disk.*

What is Apache Kafka used for?

- *Kafka is used to build **real-time streaming data pipelines** and **real-time streaming applications**.*
- *For example, if you want to create a data pipeline that takes in user activity data to track how people use your website in real-time, Kafka would be used to ingest and store streaming data while serving reads for the applications powering the data pipeline.*
- *Kafka is also often used as a **message broker solution**, which is a platform that processes and mediates communication between two applications.*