Naya College

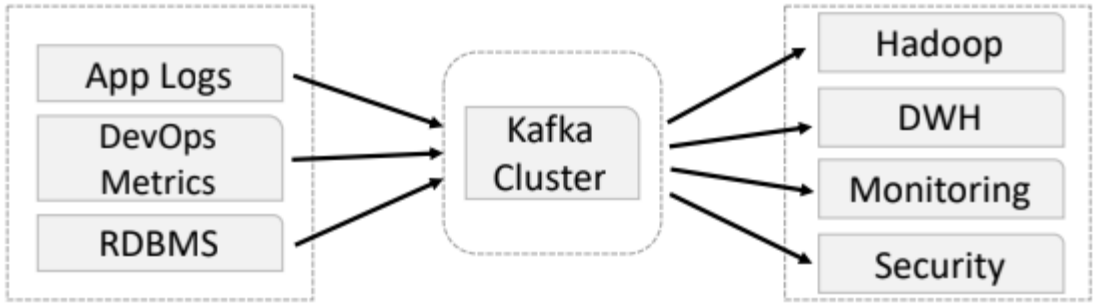
Agenda

- Introduction
- Basics
- Kafka Producer
- Kafka Consumer

Introduction

What is Apache Kafka?

- Apache Kafka is distributed messaging system based on the principle of pub-sub (publish-subscribe) model
- Kafka is incredibly fast, highly scalable, fault-tolerant system
- It can store stream of records and can serve as data platform



Introduction

- Kafka is the most popular platform for streaming data
- Kafka offers a scalable messaging backbone for data integration
- Kafka provide the ability to build streaming data pipelines
- Kafka provides logical functionality for data processing and Transformations
- Kafka is an event / message handler which can address data modifications in database tables or logs files and more

Introduction

Apache Kafka Data Streaming

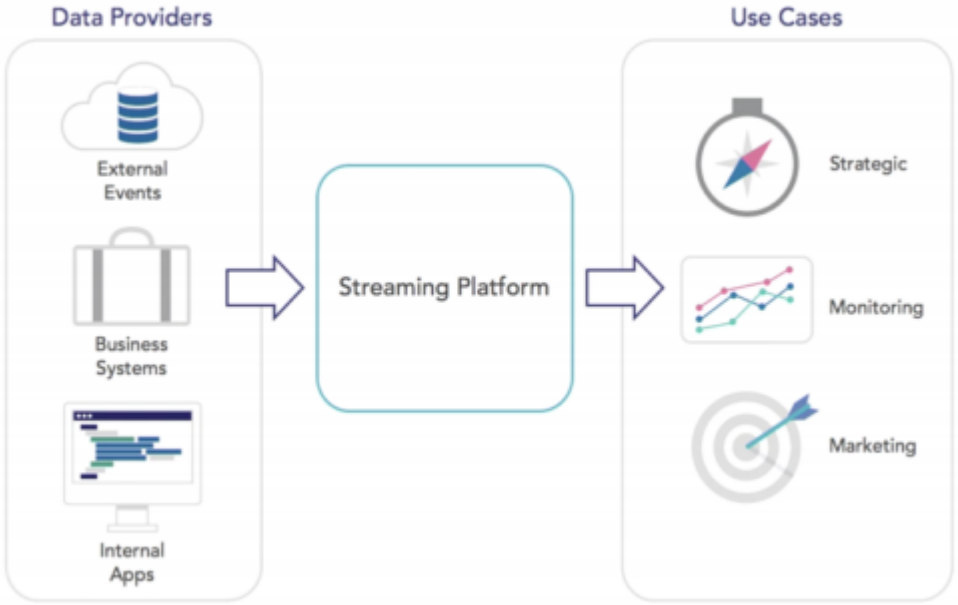
- Provides connections APIs to multiple data sources (Apps | DBs)
- Can process high-volume and high-diversity of data
- Event-Centric Thinking - can process events as they occurs

In a sales web application, a product was viewed...
*Kafka can provide a **Real-Time** product view aggregation*

Introduction

Apache Kafka Architecture

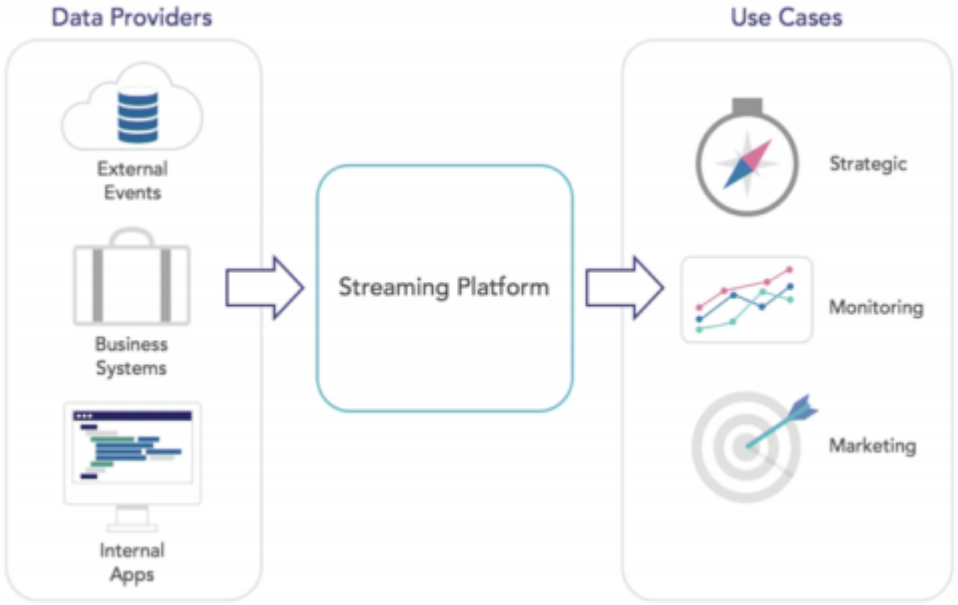
- Kafka uses streaming instead of batch processing
- In Kafka, all the data providers will be known as PRODUCERS
- Data will be ingested into the the streaming platform as it begin generated - real-time



Introduction

Apache Kafka Architecture

- As data being written inside the streaming platform operations can be performed
- Operation such as: orders count | Likes count, all in real-time
- Once the data preparation is done it can be sent to the CONSUMERS



Introduction

Apache Kafka Real Use Case

- Netflix uses Kafka to records events created by users such as:

- All movies that being watched by users (for each user)
- Creating a recommendation system for a specific user
- Saves movie last stop point (”continue from last position”) for
- each user for each movie / program

- Kafka helps Netflix to provide a better user experience

Kafka Producer

- Import Kafka Producer

```
from kafka import KafkaProducer
```

- Create Kafka connector

```
producer = KafkaProducer(bootstrap_servers=['localhost:9092'])
```

- Send Data

```
producer.send(“topic_name”, value=b'Hello, World!')
producer.flush()
```

Kafka Consumer

- Import Kafka Consumer

```
from kafka import KafkaConsumer
```

- Create Kafka consumer with name of topic

```
consumer = KafkaConsumer(“topic_name”)
```

- For any message do something…
  in this example print the data

```
for message in consumer:
    print (str(message.value))
```

Kafka

Kafka

Exresices:

Use example 1, but for each consumed message, print the following information:

- message offset
- exact date and time of it
- the text of the message.

Kafka

Kafka

Exresices:

Repeat the 3rd example, but add a staging step to the HDFS.
On top of the insertion to MySQL, every 10 events a csv file with the last 10 events should be uploaded to /tmp/staging/kafka/ in HDFS.
The file name should reflect the timestamp of the first message in the file.