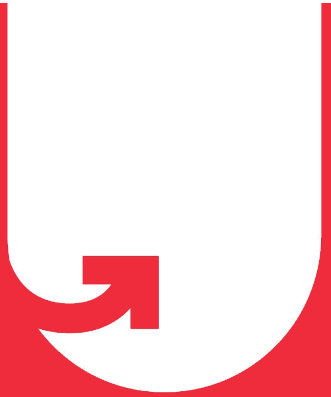
upGrad

Real-Time Data Streaming with Apache Kafka

About

upGrad



upGrad

Course: Data Engineering - II

Lecture On: Real-Time Data Streaming with Apache Kafka

Instructor: Vishwa Mohan

BATCH PROCESSING

01 Data is collected over a period of time.

02 This time period can be an hour, a day or a week.

03 After data is collected, it is processed

04 Examples: Banks issuing monthly statements, electricity bills, sales of a particular item over a period of time.

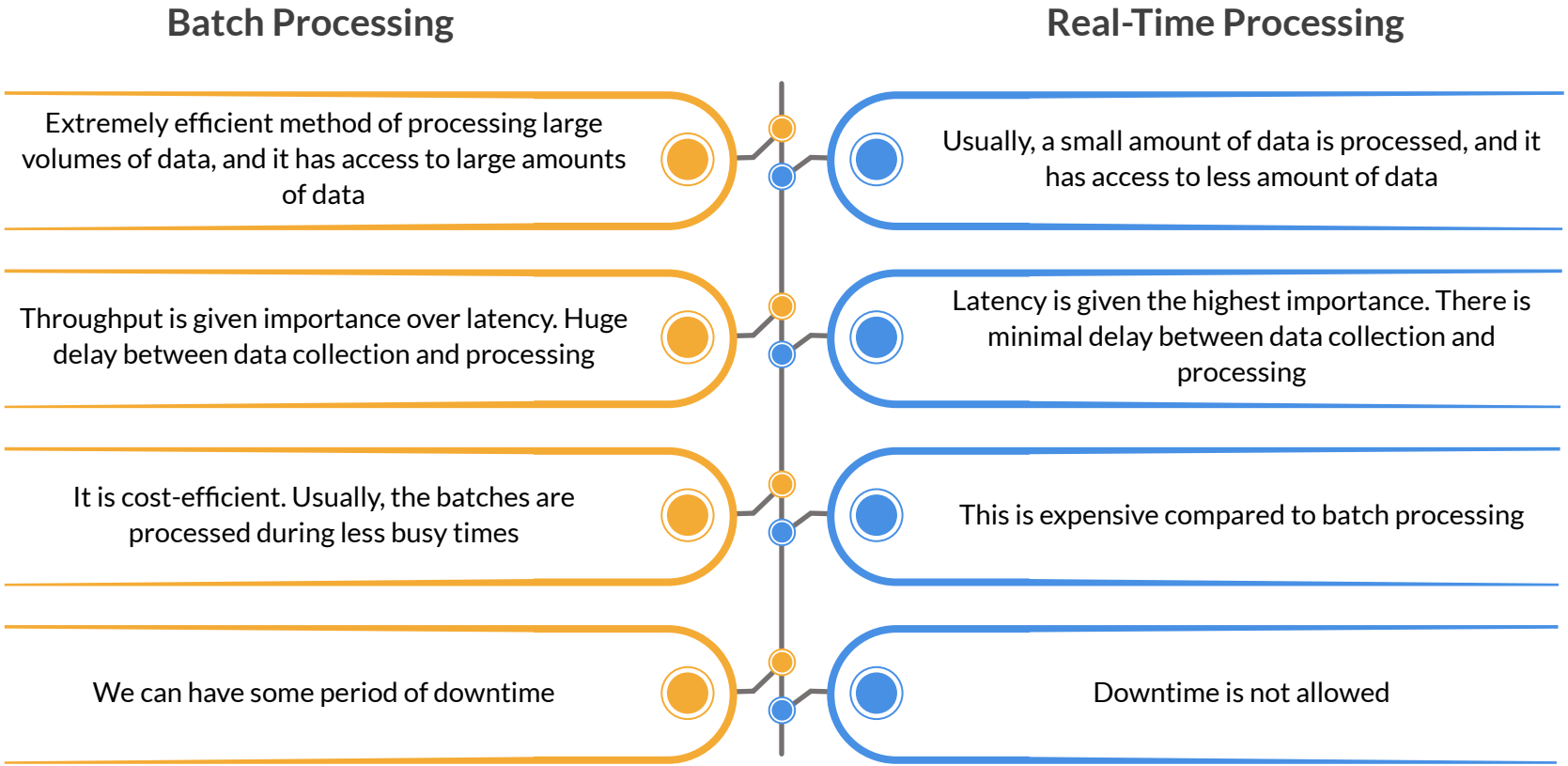
REAL-TIME PROCESSING

01 It allows the processing of data in real time.

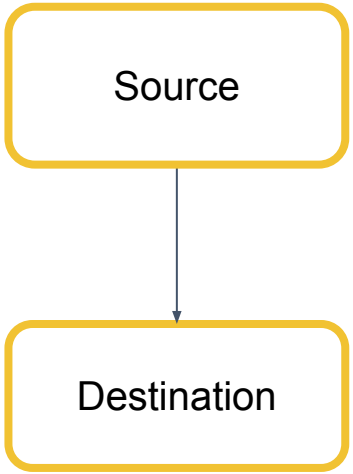
02 Data coming in real time is collected and processed.

03 Examples: Bank ATMs, fraud detection, Twitter

BATCH VS REAL-TIME PROCESSING

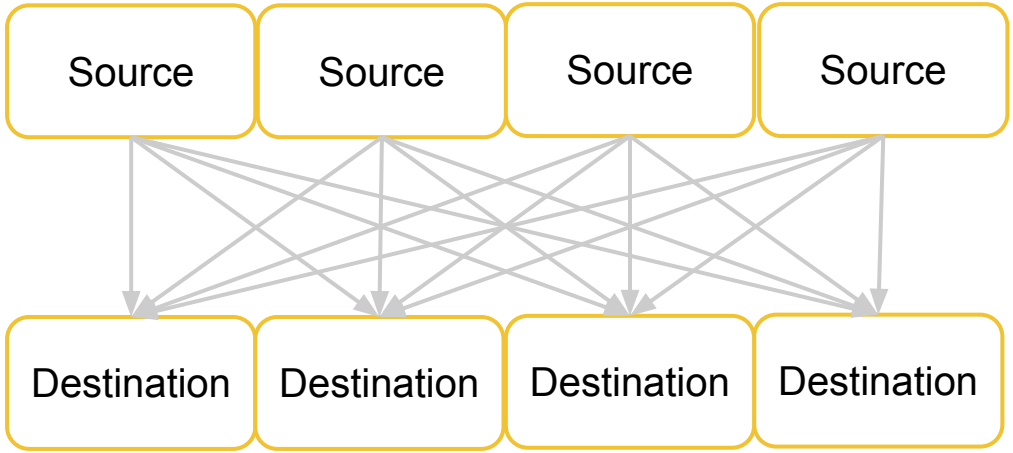


TRADITIONAL MESSAGING SYSTEMS



1. Use message queues
2. Code in source for writing data to queue:
   - Encoding in the right format
   - Compression
   - Right protocol to send data
3. Code in destination to read from queue:
   - Decode it into the right format
   - Decompress data

CHALLENGES



- 4 source and 4 destinations

CHALLENGES

01 Write code in each source and destination

02 Need to consider what protocol to use, for example, Rest, FTP, SMTP

03 Lot of replicated codes

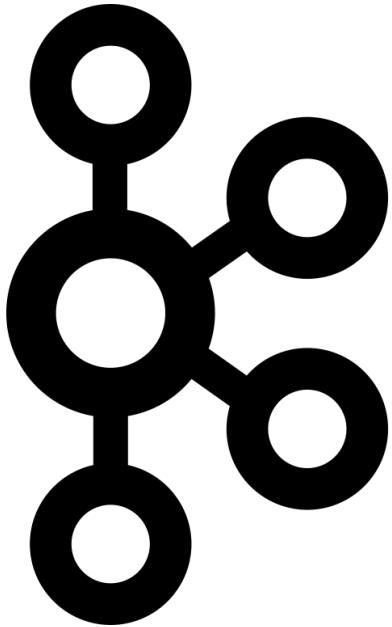
04 What if we have to change anything in the code?

- Same deployment in multiple servers
- Maintain all source and destination servers

05 What happens if a server goes down?

- Bring in a new server
- Deploy the same code in the new server
- Make sure it is integrated properly

WHAT IS KAFKA?



- Apache Kafka is a distributed streaming platform
- It is used by many companies for:
  - High-performance data pipelines
  - Streaming analytics
  - Data integration

KEY FEATURES OF STREAMING PLATFORM

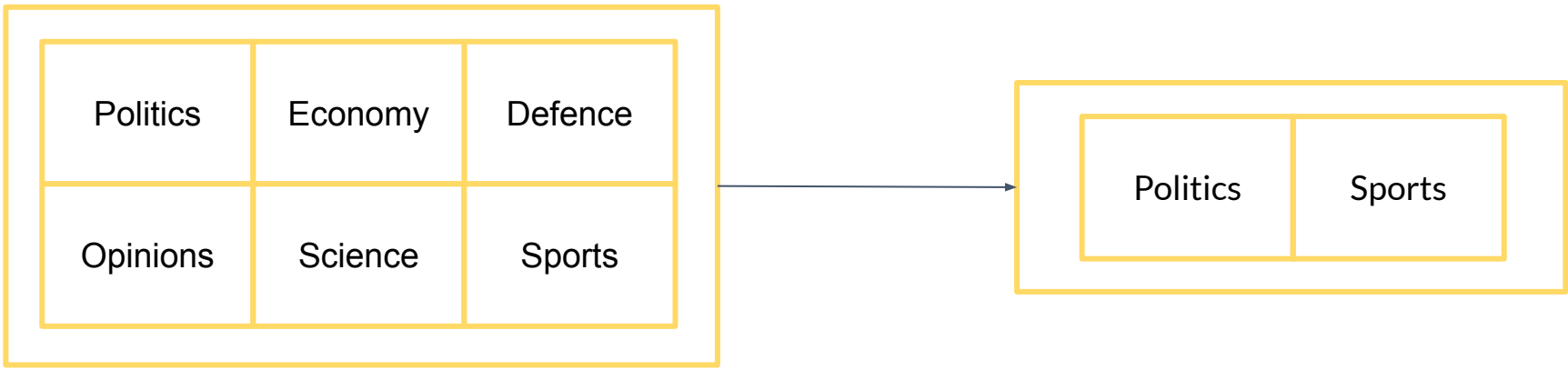
01 Publish and subscribe to streams of records

02 Store streams of records in a fault-tolerant manner

03 Process streams of records as they occur

PUB-SUB MODEL

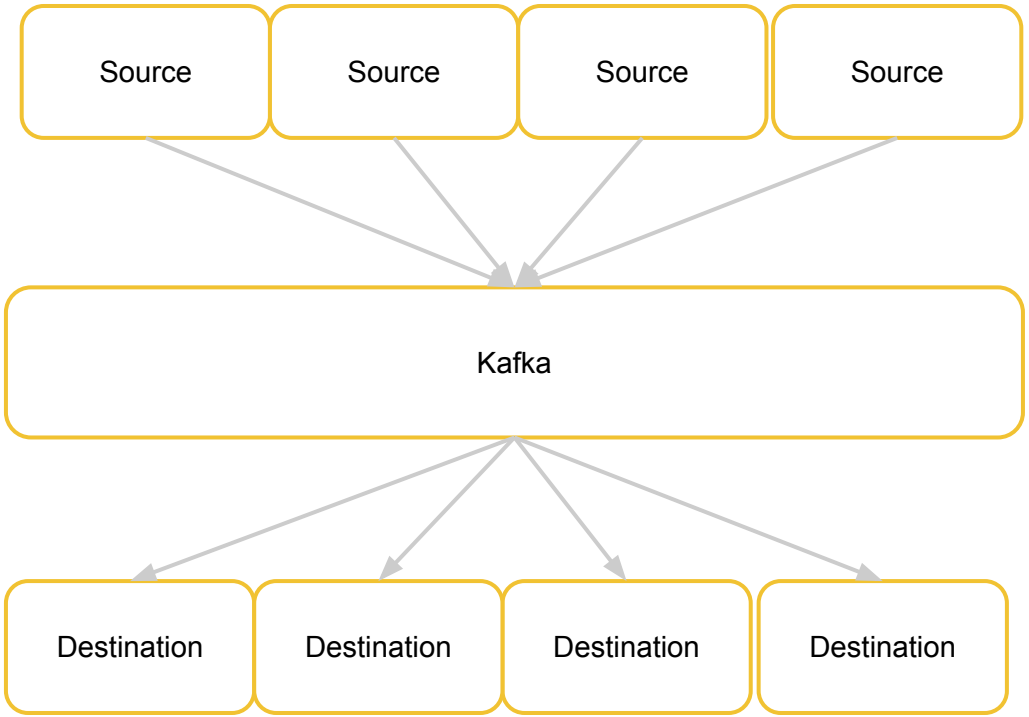
You can think of the Pub-Sub model as a newspaper website.



Newspaper Website (Publisher)

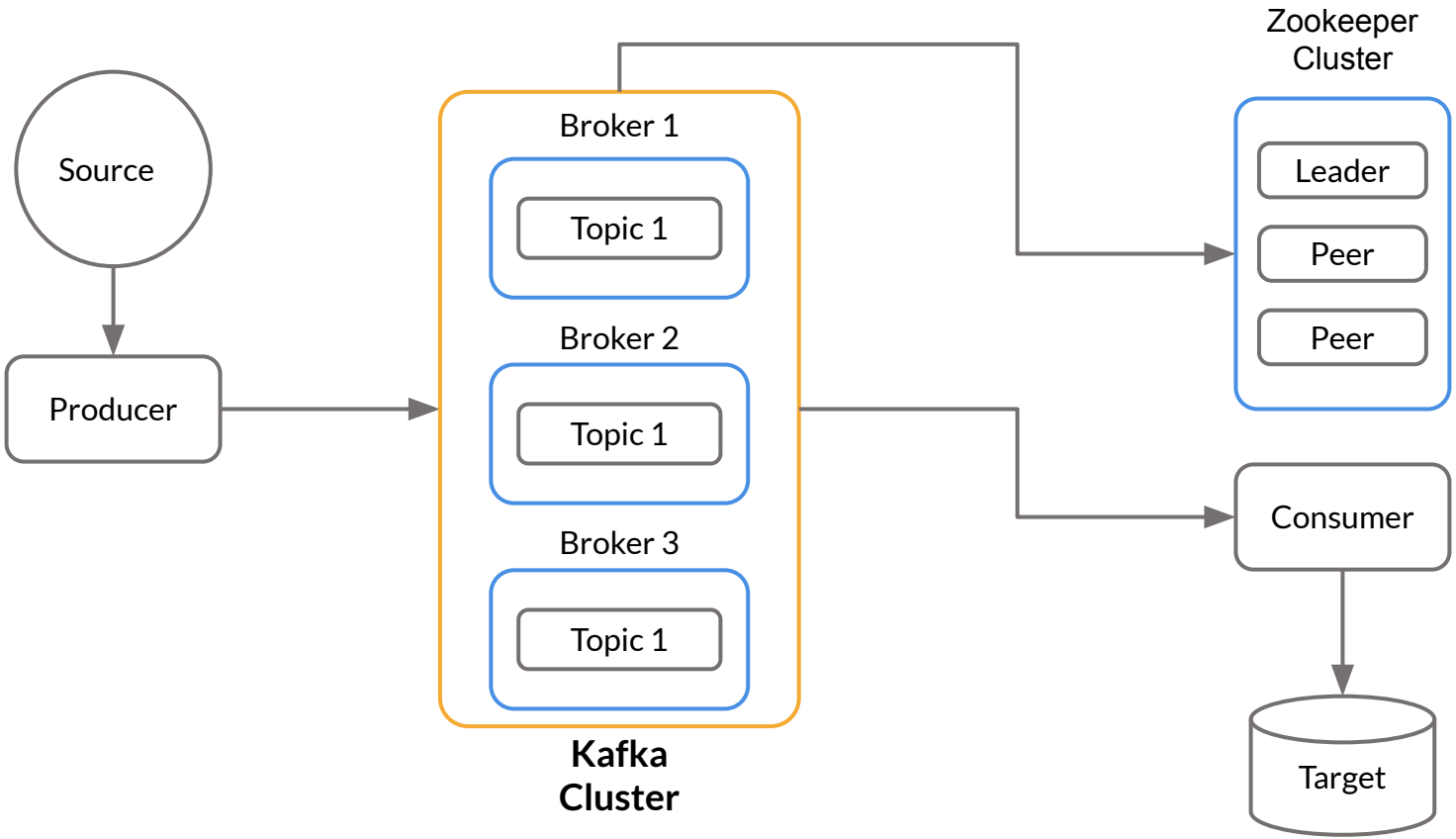
Common Man (Subscriber)

SALIENT FEATURES OF KAFKA



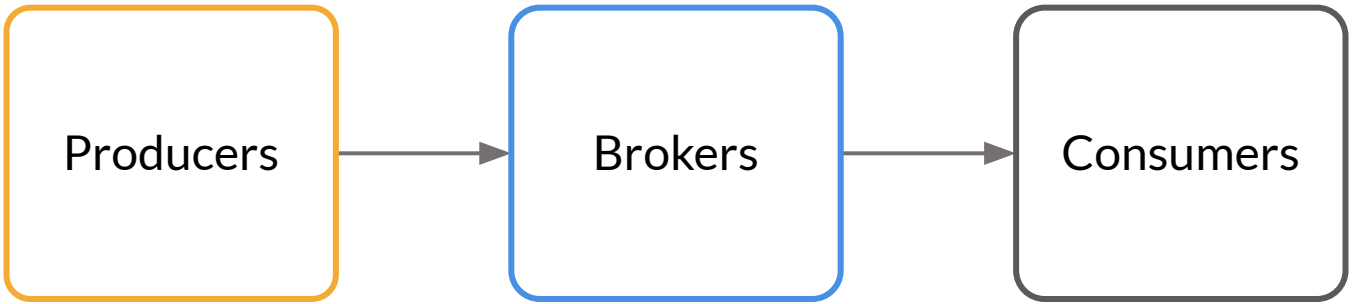
- It is a distributed system
- All sources have to write to Kafka
- All destination servers have to read from Kafka
- It solves all the problems of traditional messaging systems
- Brings best of both models of messaging system - Queueing and Publish - Subscribe
- Data can be stored in Kafka

KAFKA ARCHITECTURE



HOW DOES KAFKA FOLLOW THE PUB-SUB MODEL?

- Producers write data
- Brokers store data
- Consumers consume data as per their requirements



upGrad

Thank You