



Apache Kafka

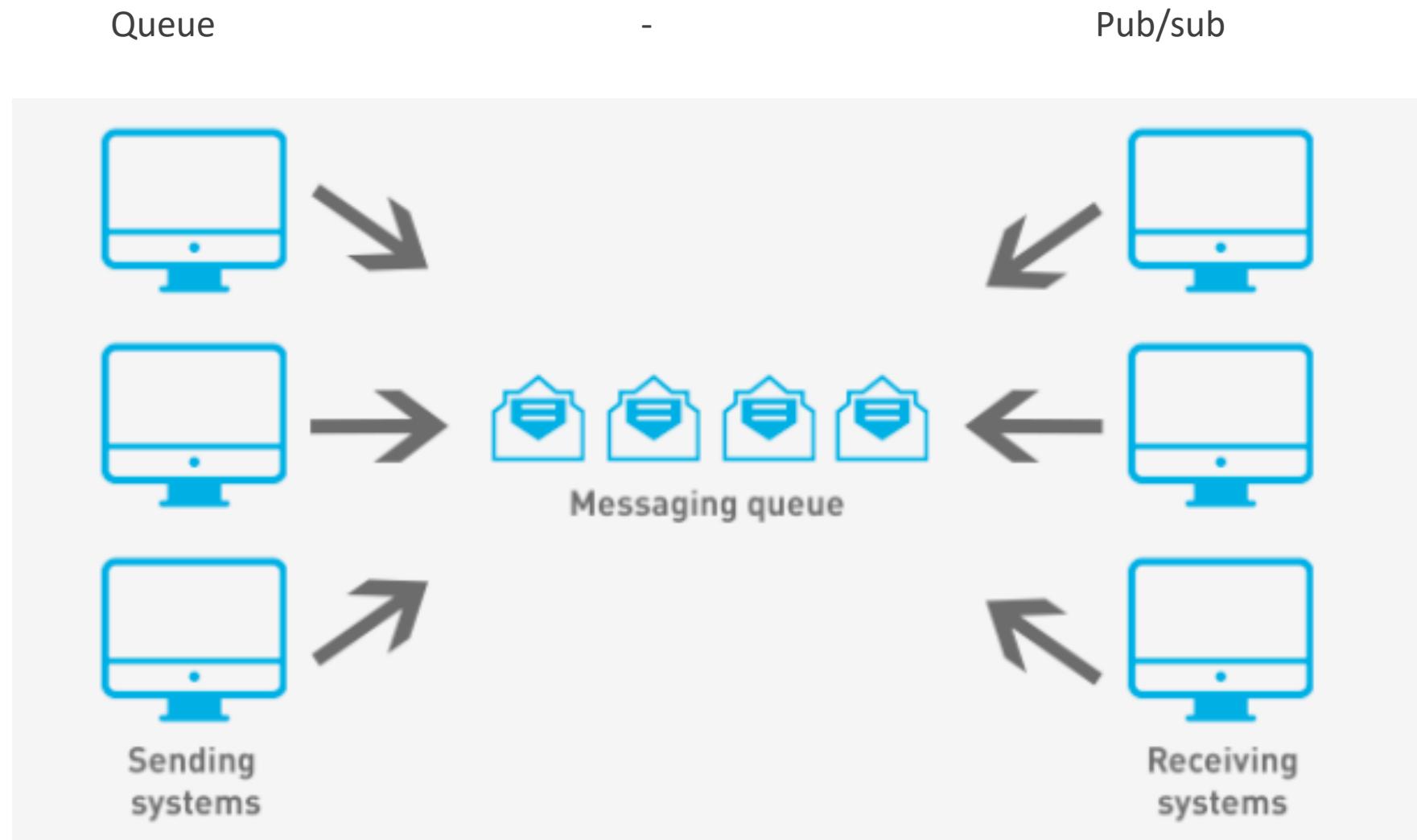
và hệ thống xử lý dữ liệu hướng sự kiện
Apache Kafka Series



Giới thiệu Queue Message

Apache Kafka Series

Giới thiệu Queue Message



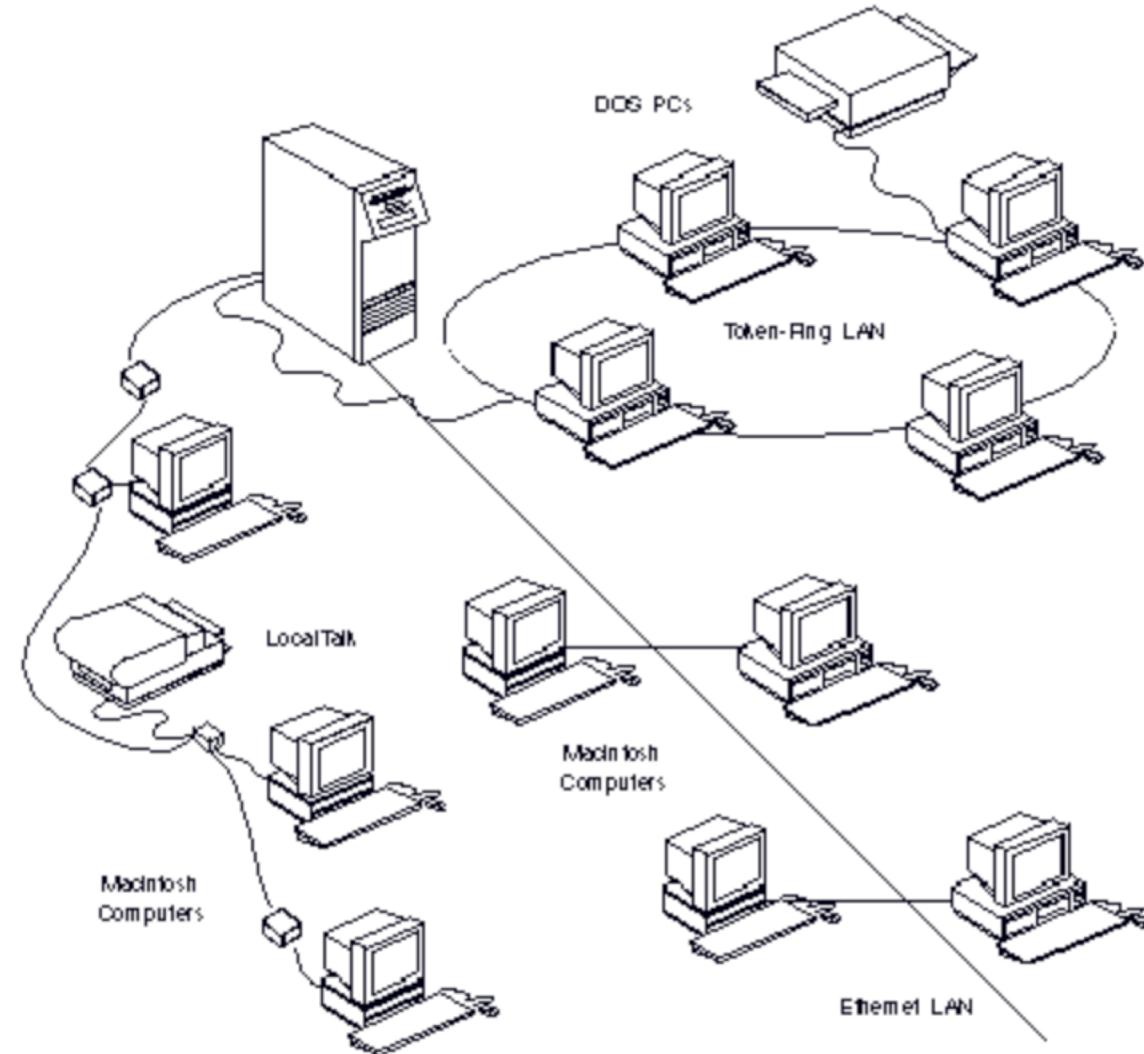
Ứng dụng mua hàng

USE-CASE



Web tracking

USE-CASE



Xây dựng hệ thống hiện đại

“Công nghệ phát triển
giúp cho con người được đáp ứng nhu cầu ngay lập tức...”

- Xử lý từng sự kiện
- Phản hồi tức thì
- Độ chính xác cao

From a static snapshot...



Occasional call to a friend

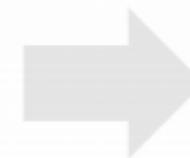
...to a continuous stream of events



A constant feed about the activities of all your friends



Daily news reports



Real time news feeds, accessible online anytime, anywhere

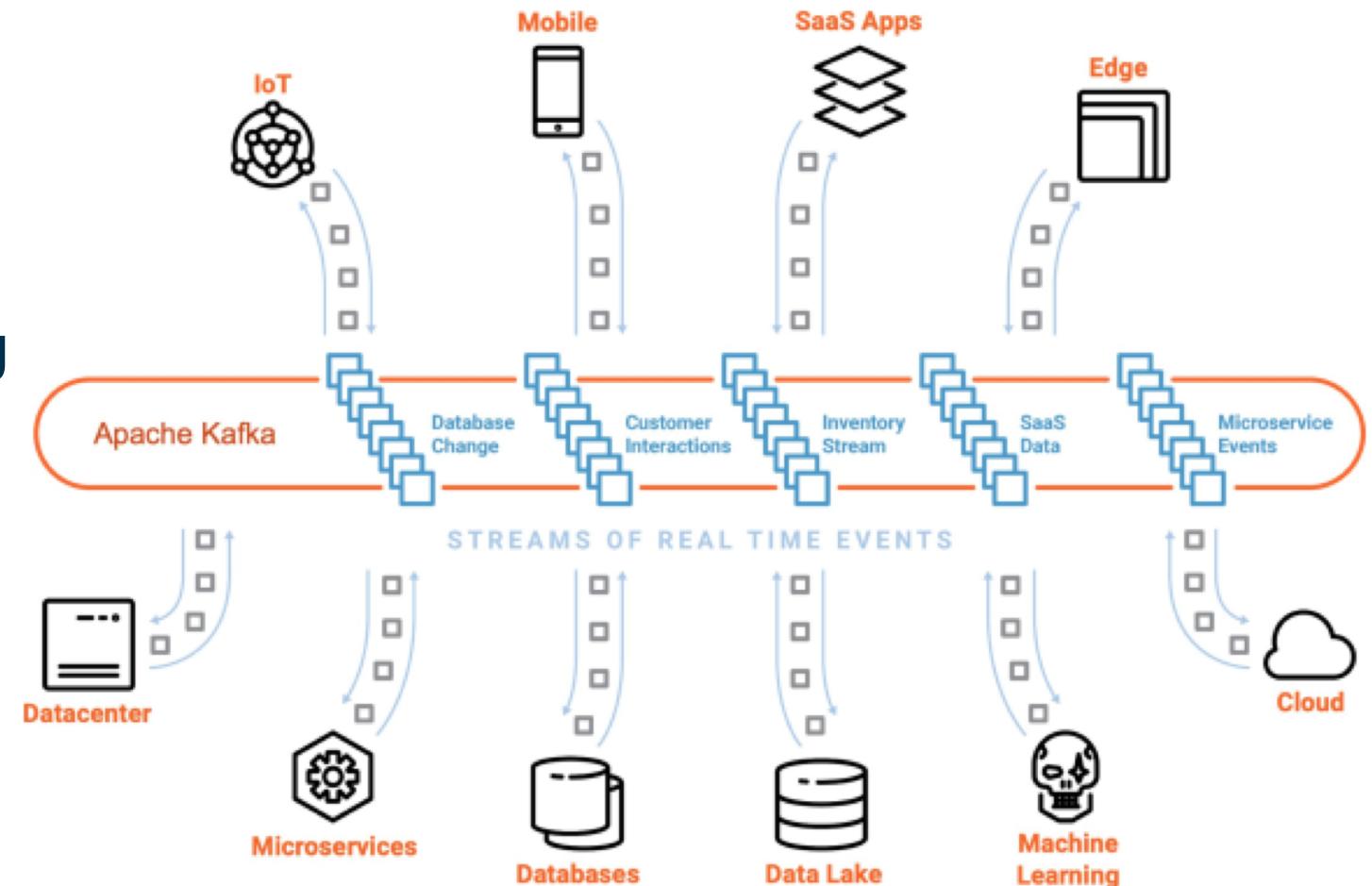
Xây dựng hệ thống hiện đại

*"Hàng ngàn doanh nghiệp lựa chọn Kafka
trong việc xây dựng hệ thống hướng sự kiện"*



Vị trí của Kafka

- Là message queue phân tán
- Trung gian giữa các hệ thống
- Publish và subscribe bản ghi theo luồng



Ưu điểm Kafka



Khả năng mở rộng

Hỗ trợ mở rộng cụm Kafka theo chiều ngang, tương tự cách mà Hadoop mở rộng

Chịu tải và sự ổn định

Nhờ cơ chế mở rộng và partitioning, chỉ 7-8 node xử lý tới hàng chục tỉ bản ghi mỗi ngày

Độ tin cậy dữ liệu

Dữ liệu được lưu trữ phân tán, có replication từng node cùng với cơ chế quản lý từng bản ghi thông qua offset

Zero-downtime

Một hệ thống Kafka được cấu hình tốt có thể vận hành với thời gian chết (downtime) = 0

Use-case thực tế

Phát hiện gian lận tín dụng

- Thông tin giao dịch của khách hàng
- Được xử lý thông qua các mô hình Machine – Learning
- Phát hiện bất thường hoặc gian lận, từ đó dừng các giao dịch nghi ngờ

Hệ thống xe tự lái

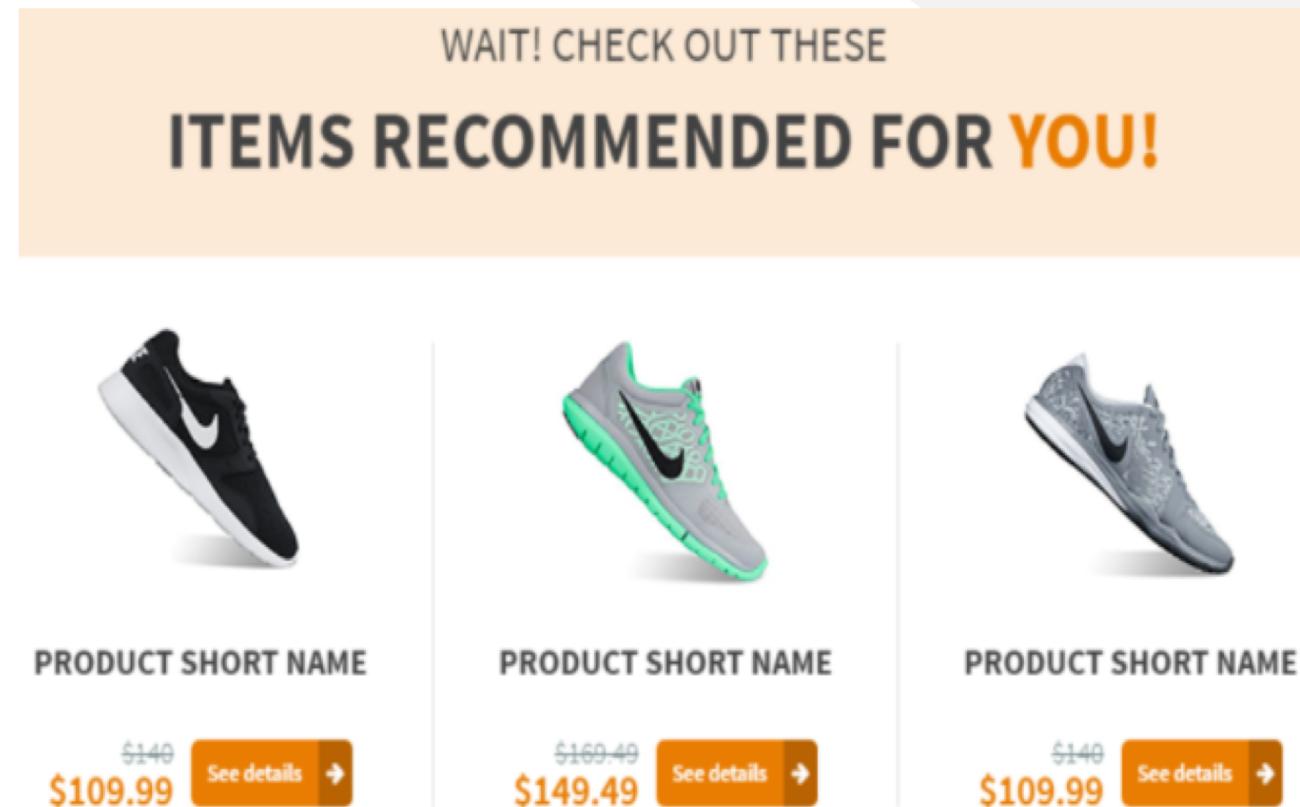
- Dữ liệu là các thông tin từ cảm biến
- Được xử lý thời gian thực để đưa ra hành vi lái xe.



Use-case thực tế

Gợi ý sản phẩm cho khách hàng

- Các thao tác của người dùng trên hệ thống được ghi nhận thành các sự kiện
- Xử lý realtime
- Sau đó phân tích và đưa ra các sản phẩm, dịch vụ tương ứng với hành vi của khách hàng



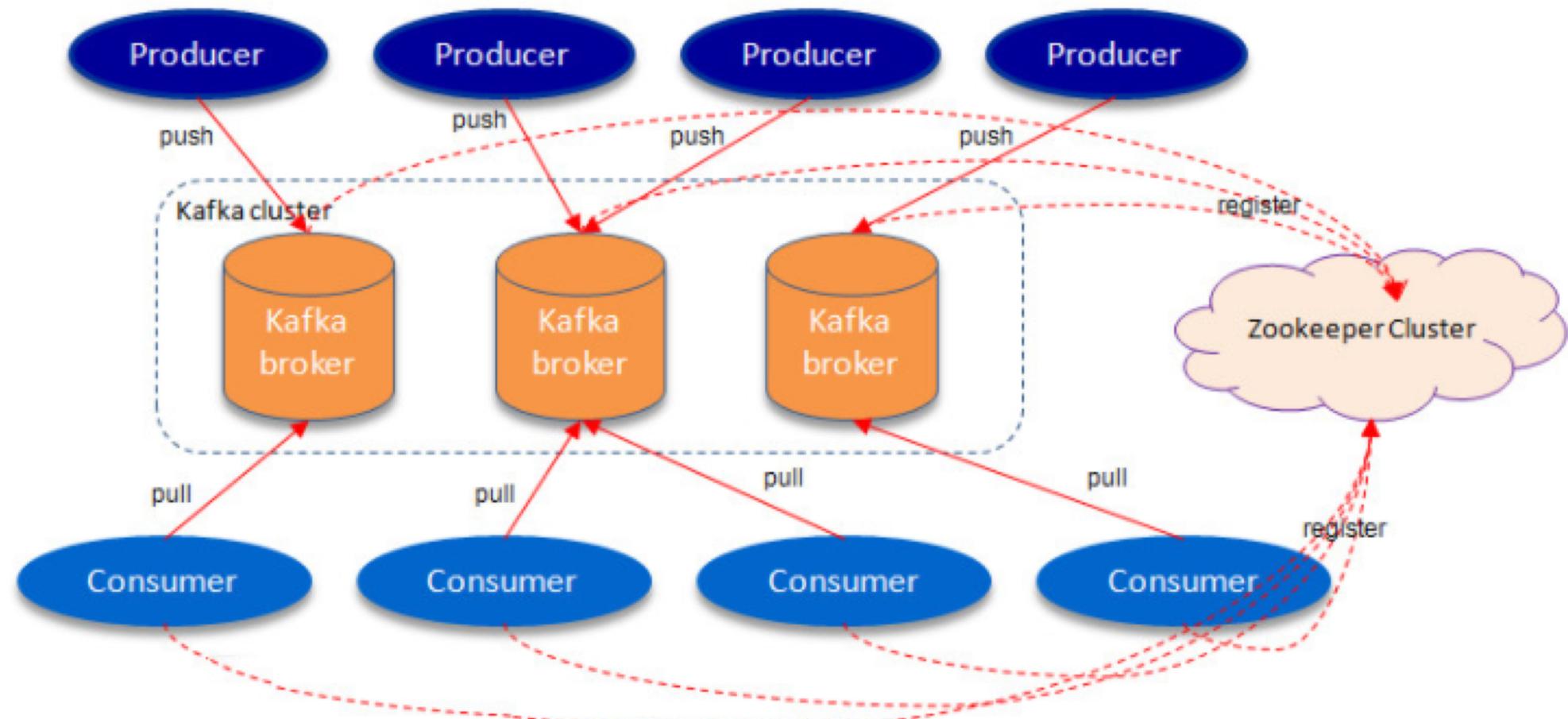


Kiến trúc của Kafka

Apache Kafka Series

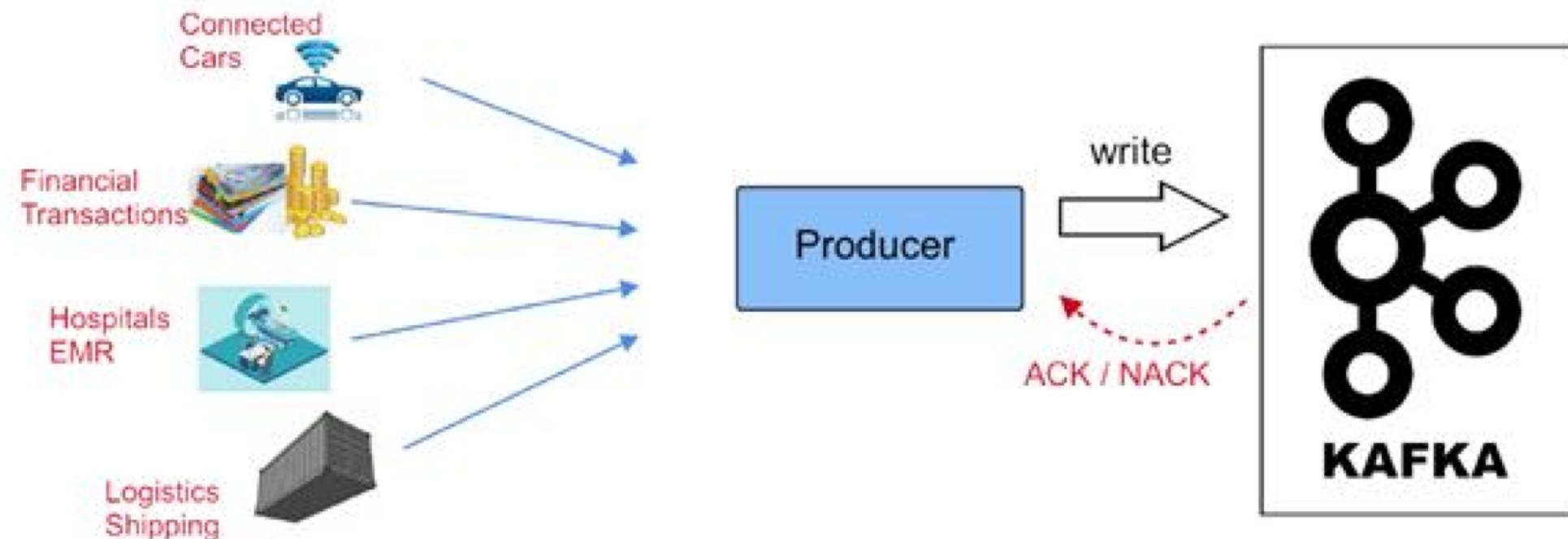
Tổng quan

KIẾN TRÚC KAFKA



Các thành phần

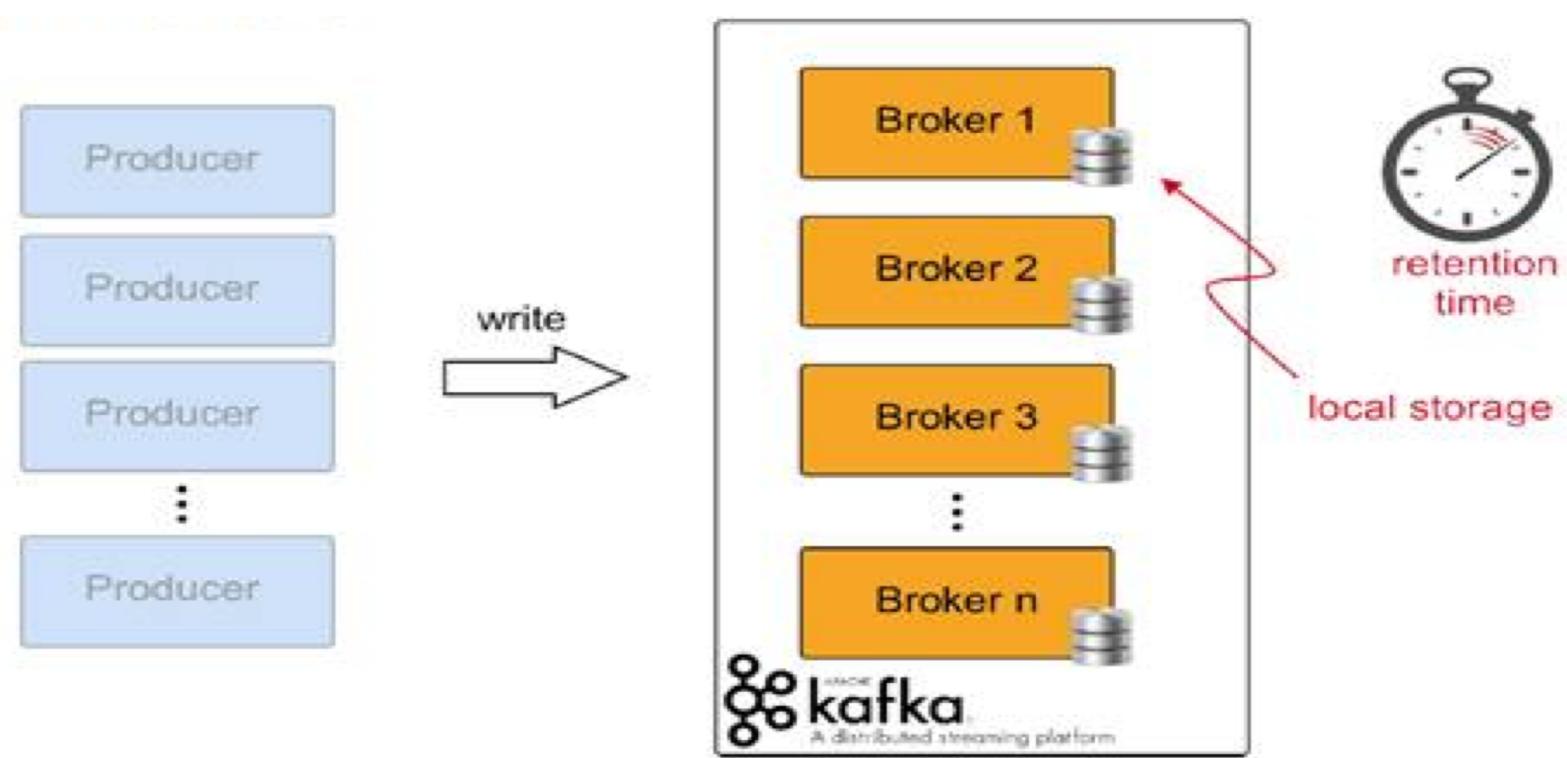
Kafka Producers



- Chịu trách nhiệm thu thập dữ liệu
- Đưa và hệ thống Kafka (publish message)

Các thành phần

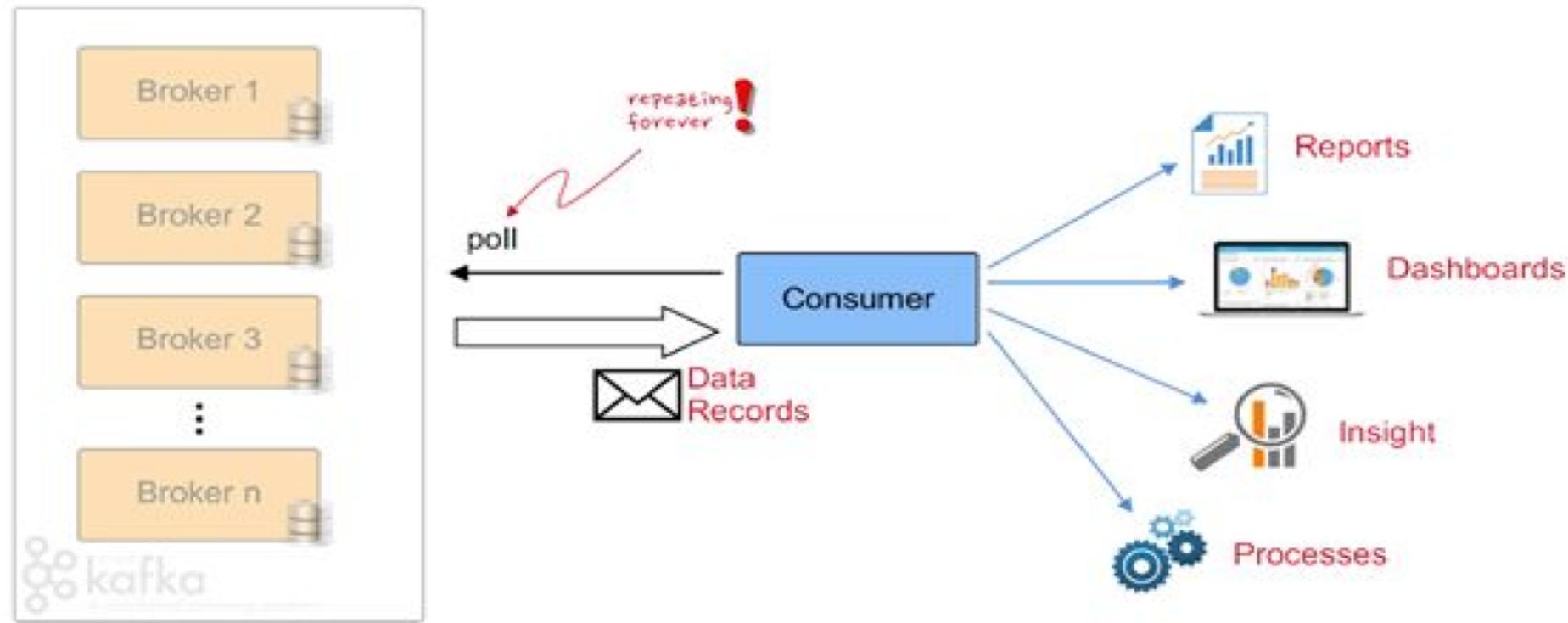
Kafka Brokers



- Các chương trình thực thi của Kafka
- Một cụm Kafka gồm nhiều Broker

Các thành phần

Consumers

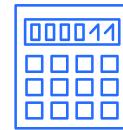


- Lấy dữ liệu từ Kafka ra
- Đưa dữ liệu tới các hệ thống xử lý

Producer vs Consumer

Producer và consumer có sự tách biệt rõ ràng

Tính chịu lỗi



Producer bị chậm hoặc lỗi sẽ
không ảnh hưởng tới
Consumer

Tính mở rộng



Việc thêm/bớt các consumer
không ảnh hưởng tới producer

An toàn dữ liệu



Nếu consumer lỗi, dữ liệu vẫn
được bảo an toàn, tránh
mất mát dữ liệu

Các thành phần

Zookeeper

- ◆ Quản lý cấu hình ứng dụng trong cụm
- ◆ Leader election
- ◆ Kiểm soát việc sử dụng tài nguyên chung
- ◆ Kiểm soát các thành viên trong cụm



Khái niệm quan trọng



TOPIC



PARTITION



SEGMENT



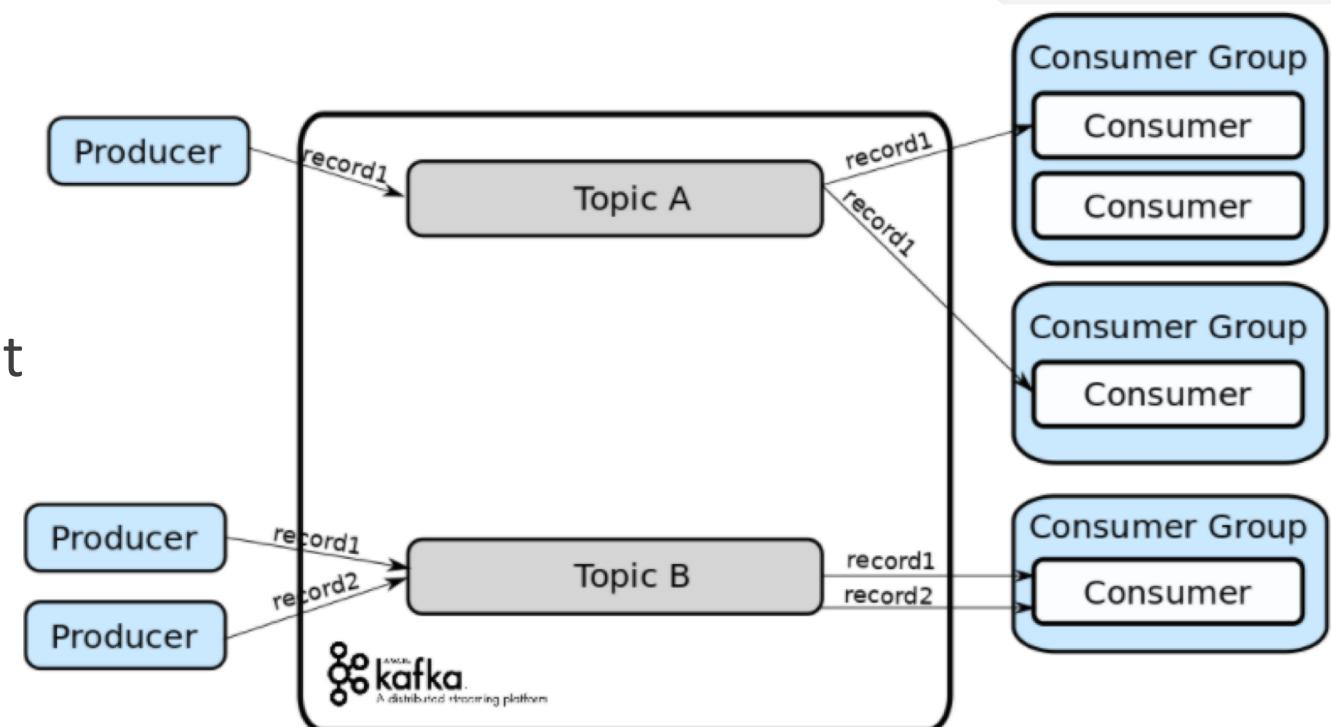
LEADER



TOPIC

Streams of “related” Messages in Kafka

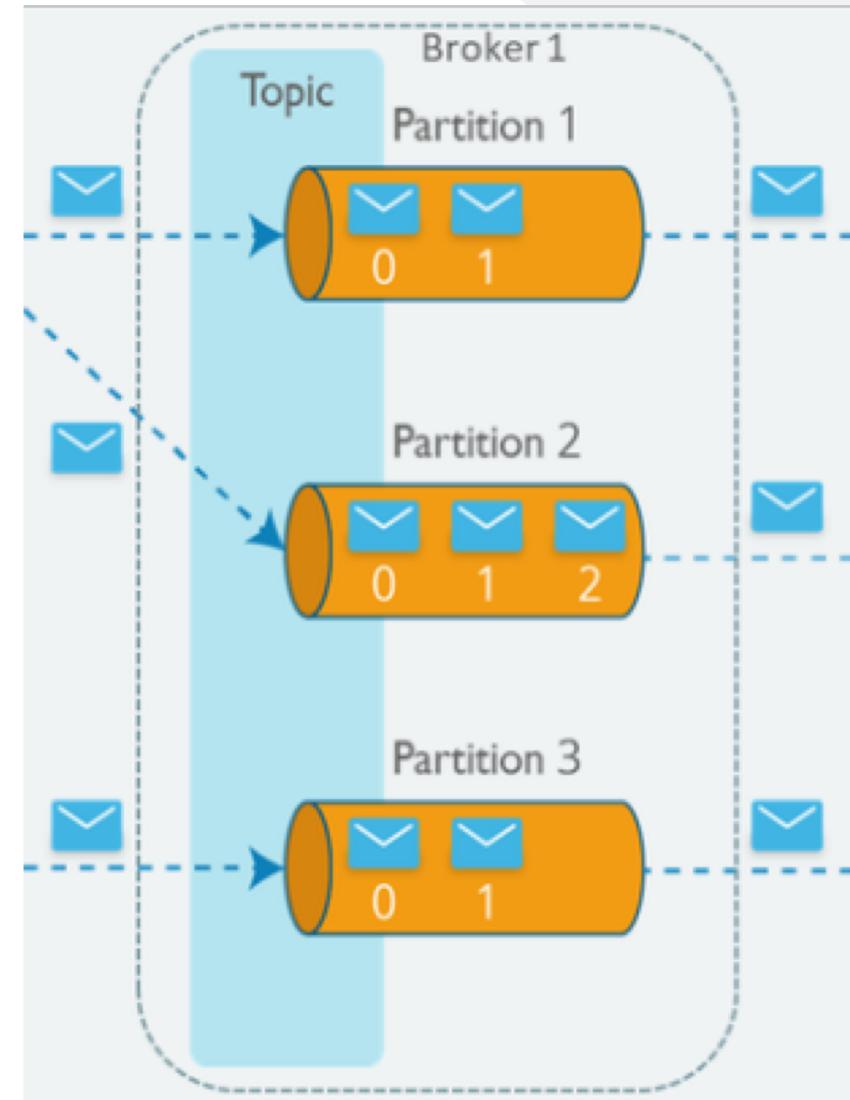
- ✓ Là cách chúng ta tổ chức lưu trữ dữ liệu về mặt logic
- ✓ Nếu coi Kafka như một database thì topic là một table
- ✓ Mỗi message lưu trong topic như một bản ghi
- ✓ Mỗi topic gồm nhiều partition



PARTITION

...

- ✓ Nơi lưu trữ message của topic
- ✓ Số lượng partition giúp khả năng xử lý song song hiệu quả
- ✓ Các partition được lưu theo thứ tự bất biến (offset)
- ✓ Mỗi partition có ít nhất 1 replica chống lỗi
- ✓ Số lượng partition luôn bé hơn số brokers

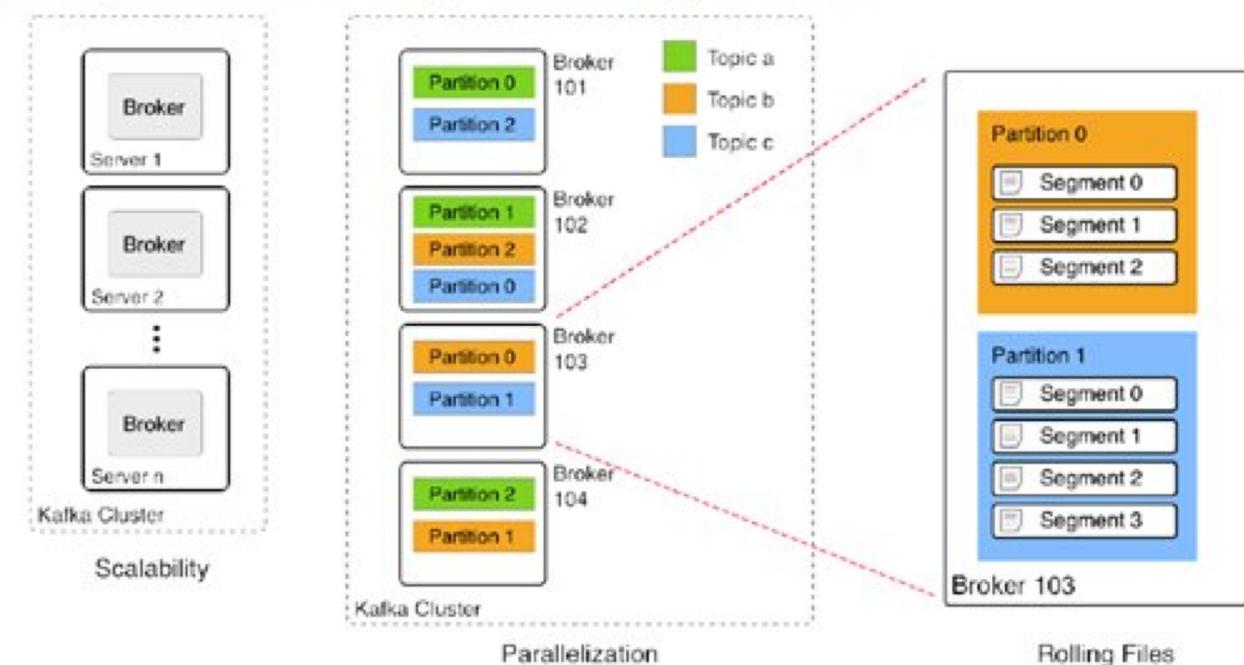


SEGMENT

...

- ✓ Là đơn vị lưu trữ nhỏ nhất
- ✓ Mỗi partition là một tập hợp các segment
- ✓ Mỗi segment là 1 file vật lý trên ổ cứng (mặc định 1GB)

Topics, Partitions, and Segments

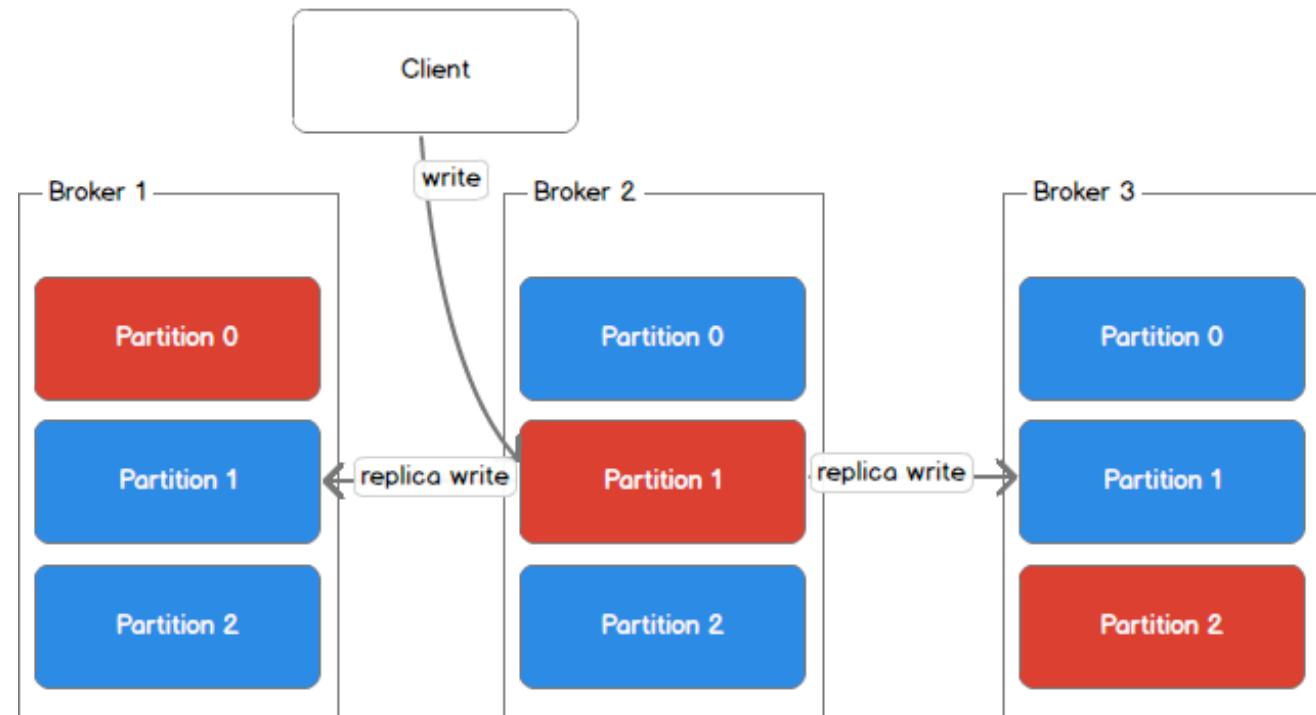


LEADER

...

- ✓ Chịu trách nhiệm cho tất cả tác vụ đọc ghi vào topic
- ✓ Chỉ partition leader mới nhận yêu cầu và xử lý dữ liệu
- ✓ Mỗi partition có 1 leader partition, một broker có thể có nhiều leader

Leader (red) and replicas (blue)

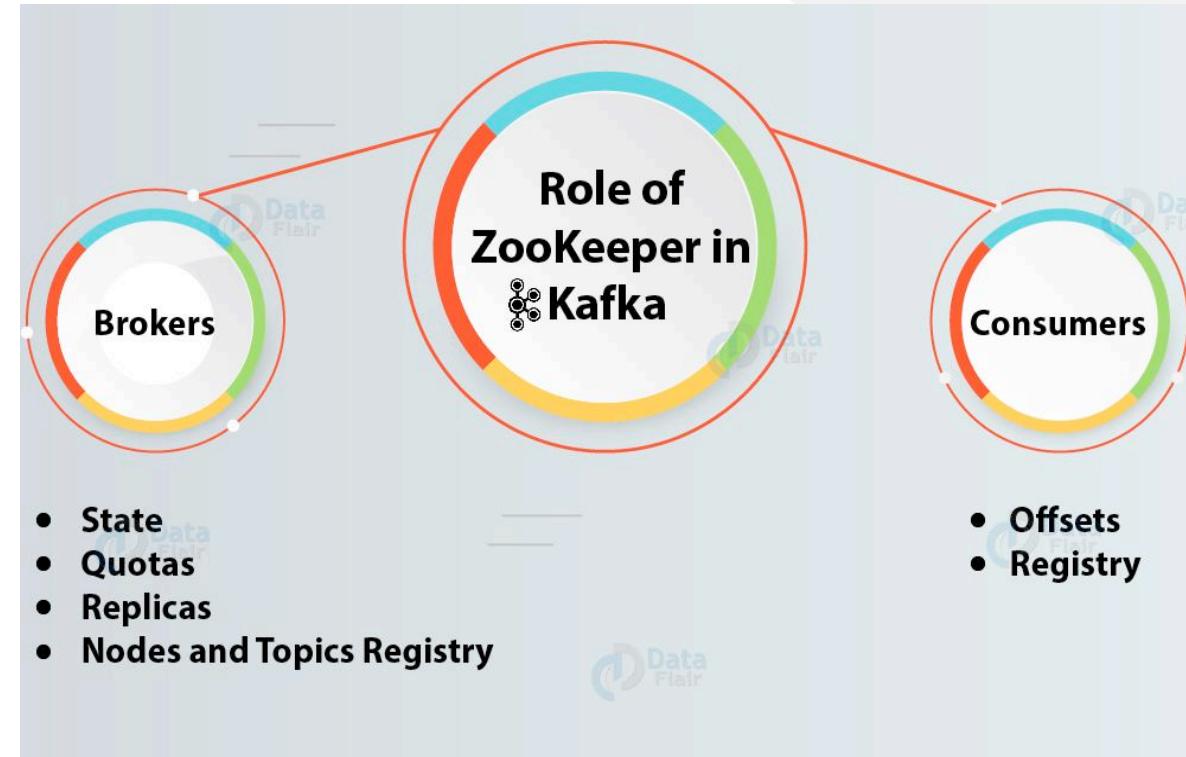


Cơ chế xử lý lỗi của Kafka



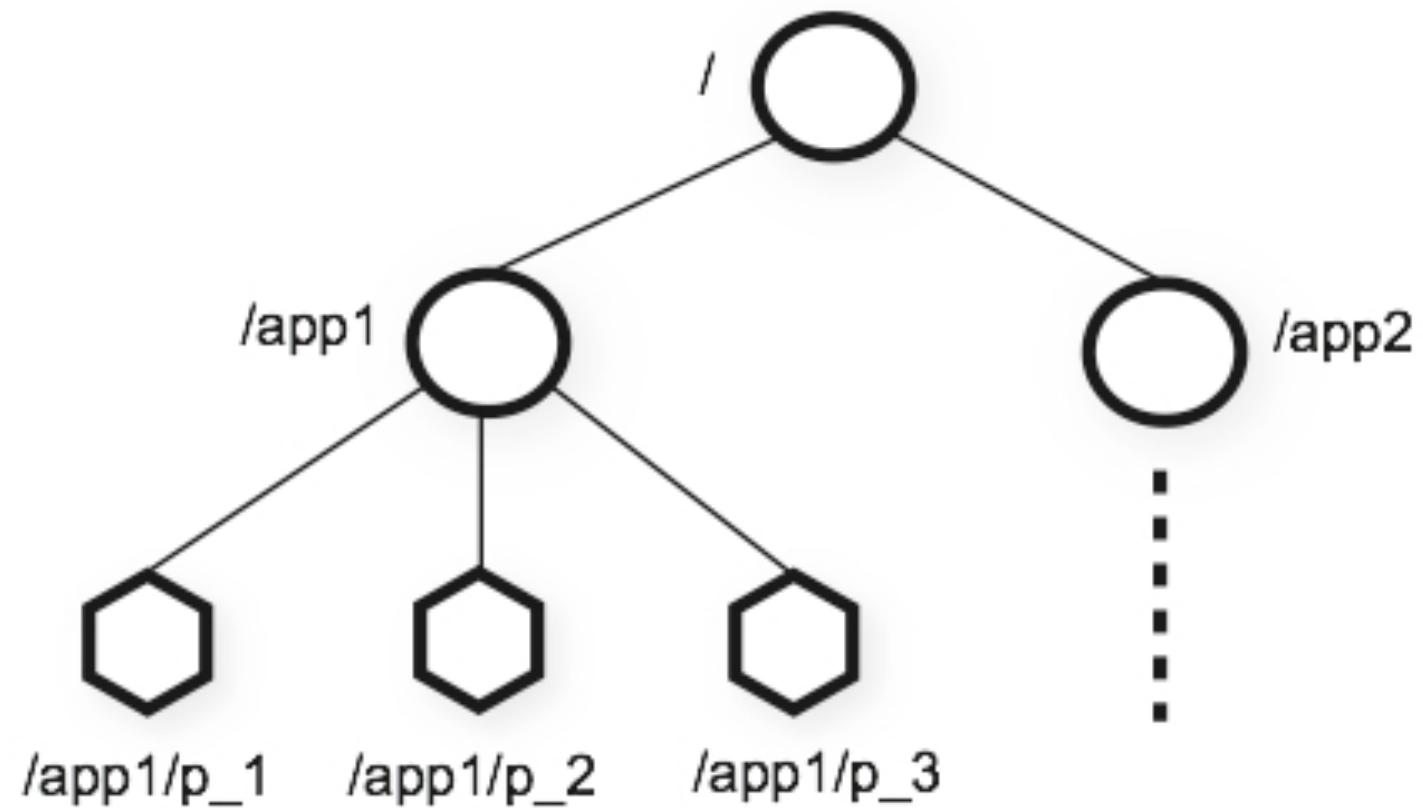
Zookeeper và vai trò

- ✓ Là coordinator trong cụm
- ✓ Bầu chọn leader mới cho partition khi một broker gặp sự cố
- ✓ Đồng bộ thông tin cụm kafka



Zookeeper và vai trò

Giao tiếp với Kafka



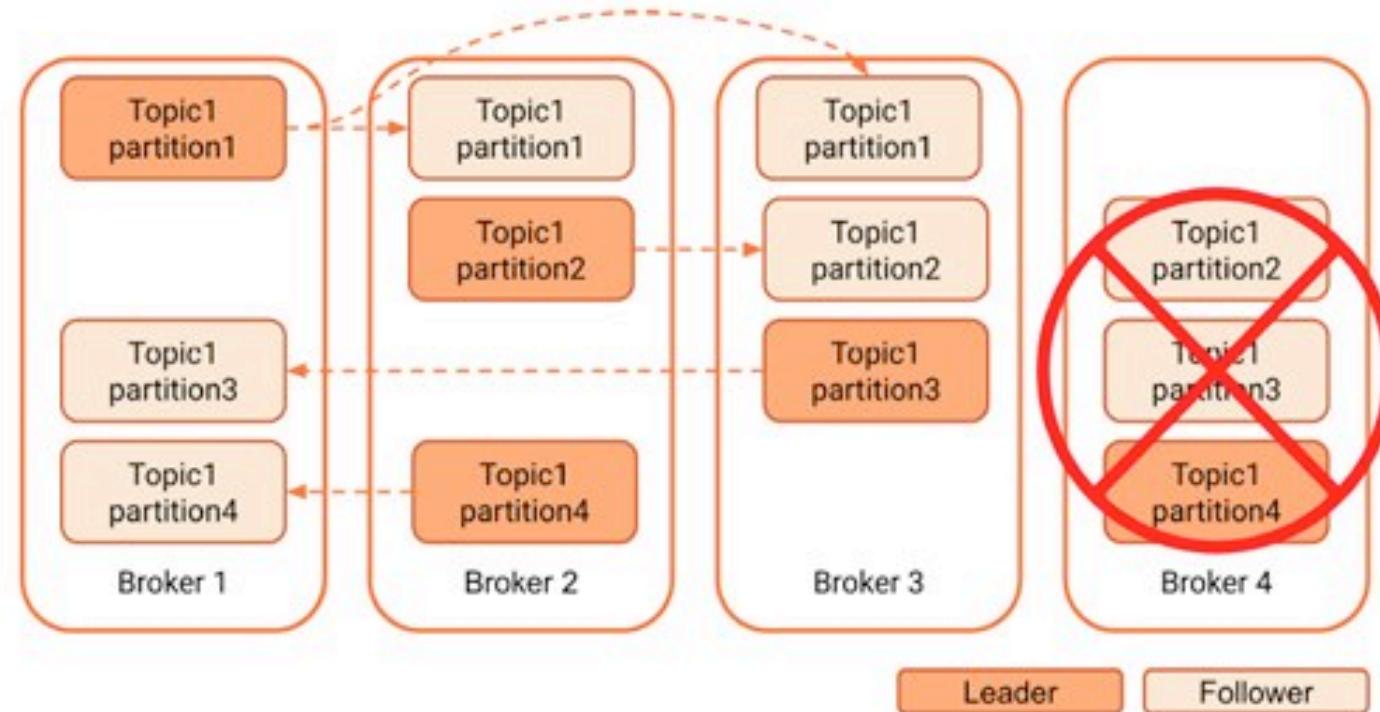
Zookeeper và vai trò

Broker down



7

Partition Leadership & Replication

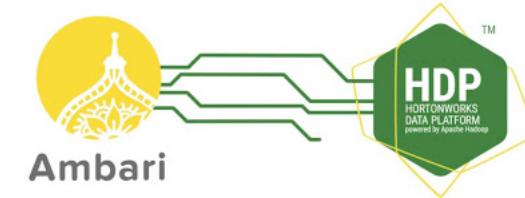
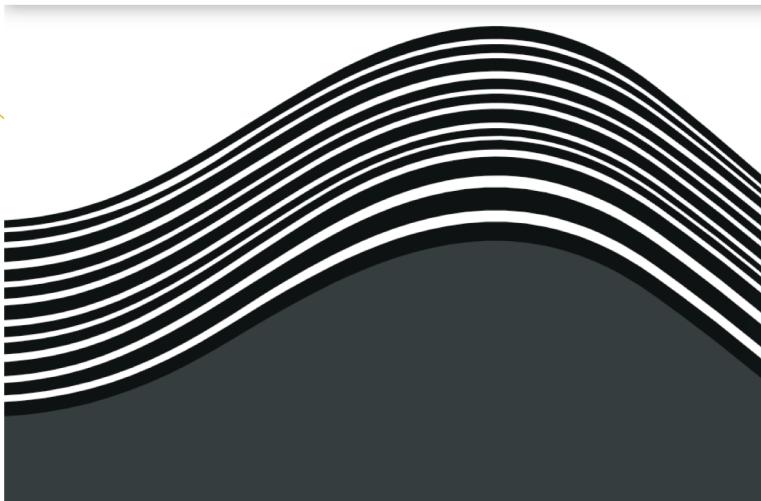


Thực hành triển khai Kafka Cluster

Apache Kafka Series

Cách tiếp cận

- Triển khai thủ công
- Công cụ hỗ trợ triển khai



Chuẩn bị

Zookeeper Installation

```
zoo.cfg x
1 # The number of milliseconds of each tick
2 tickTime=2000
3 # The number of ticks that the initial
4 # synchronization phase can take
5 initLimit=10
6 # The number of ticks that can pass between
7 # sending a request and getting an acknowledgement
8 syncLimit=5
9 # the directory where the snapshot is stored.
10 # do not use /tmp for storage, /tmp here is just
11 # example sakes.
12 dataDir=E:\\Kafka\\\\data\\\\zookeeper
13 # the port at which the clients will connect
14 clientPort=2181
```



Apache ZooKeeper™

Chuẩn bị

Zookeeper Installation

- ✓ Download zookeeper tại <https://zookeeper.apache.org/releases.html#download>:

<https://www.apache.org/dyn/closer.lua/zookeeper/zookeeper-3.6.2/apache-zookeeper-3.6.2-bin.tar.gz>

(E:) > Kafka > apache-zookeeper-3.6.1-bin	
Name	Date modified
bin	8/19/2020 3:26 PM
conf	8/19/2020 4:03 PM
docs	8/19/2020 3:26 PM
lib	8/19/2020 3:26 PM
logs	8/19/2020 4:06 PM
LICENSE.txt	4/21/2020 9:59 PM
NOTICE.txt	4/21/2020 9:59 PM
README.md	4/21/2020 9:59 PM
README_pack...	4/21/2020 9:59 PM

E:) > Kafka > apache-zookeeper-3.6.1-bin > bin	
Name	Date modified
README.txt	4/21/2020 9:59 PM
zkCleanup.sh	4/21/2020 9:59 PM
zkCli.cmd	4/21/2020 9:59 PM
zkCli.sh	4/21/2020 9:59 PM
zkEnv.cmd	4/21/2020 9:59 PM
zkEnv.sh	4/21/2020 9:59 PM
zkServer.cmd	4/21/2020 9:59 PM
zkServer.sh	4/21/2020 9:59 PM
zkServer-initialize.sh	4/21/2020 9:59 PM
zkSnapShotToolkit.cmd	4/21/2020 9:59 PM



Apache ZooKeeper™

Chuẩn bị

Zookeeper Installation

Start zookeeper

```
apache-zookeeper-3.6.1-bin % ./bin/zkServer.sh start  
/usr/bin/java  
ZooKeeper JMX enabled by default  
Using config: /Users/nguyentien/Programs/apache-zookeeper-3.6.1-  
bin/bin/../conf/zoo.cfg  
Starting zookeeper ... STARTED
```

```
Connecting to localhost:2181  
.....  
.....  
.....  
Welcome to ZooKeeper!  
.....  
.....  
WATCHER:::  
WatchedEvent state:SyncConnected type: None path:null  
[zk: localhost:2181(CONNECTED) 0]
```



Apache ZooKeeper™

Chuẩn bị

Kafka Installation

- ✓ Download kafka tại <https://kafka.apache.org/downloads>:

https://www.apache.org/dyn/closer.cgi?path=/kafka/2.6.0/kafka_2.12-2.6.0.tgz

(E:) > Kafka > kafka_2.12-2.6.0 >		
Name	Date modified	Type
bin	8/19/2020 4:16 PM	File folder
config	8/19/2020 4:16 PM	File folder
libs	8/19/2020 4:16 PM	File folder
logs	8/20/2020 1:01 PM	File folder
site-docs	8/19/2020 4:16 PM	File folder
LICENSE	7/29/2020 1:16 AM	File
NOTICE	7/29/2020 1:16 AM	File

(E:) > Kafka > kafka_2.12-2.6.0 > bin >		
Name	Date modified	
windows	8/19/2020 4:16 PM	
connect-distributed.sh	7/29/2020 1:16 AM	
connect-mirror-maker.sh	7/29/2020 1:16 AM	
connect-standalone.sh	7/29/2020 1:16 AM	
kafka-acls.sh	7/29/2020 1:16 AM	
kafka-broker-api-versions.sh	7/29/2020 1:16 AM	
kafka-configs.sh	7/29/2020 1:16 AM	
kafka-console-consumer.sh	7/29/2020 1:16 AM	
kafka-console-producer.sh	7/29/2020 1:16 AM	



Chuẩn bị

Kafka Installation

- ✓ Sửa file config/server.properties

```
server.properties

1 ##### Server Basics #####
2 # The id of the broker. This must be set to a unique integer for each broker.
3 broker.id=15
4 listeners=PLAINTEXT://:9092
5 # A comma separated list of directories under which to store log files
6 log.dirs=/u01/vbi_app/kafka-logs
7
8 # The minimum age of a log file to be eligible for deletion due to age
9 log.retention.hours=168
10 # Zookeeper connection string (see zookeeper docs for details).
11 zookeeper.connect=testlab-64:8775,datanode03:8775,datanode02:8775/k25
```



Practice

- ✓ Start Kafka

```
./bin/kafka-server-start.sh config/server.properties
```

- ✓ Stop Kafka

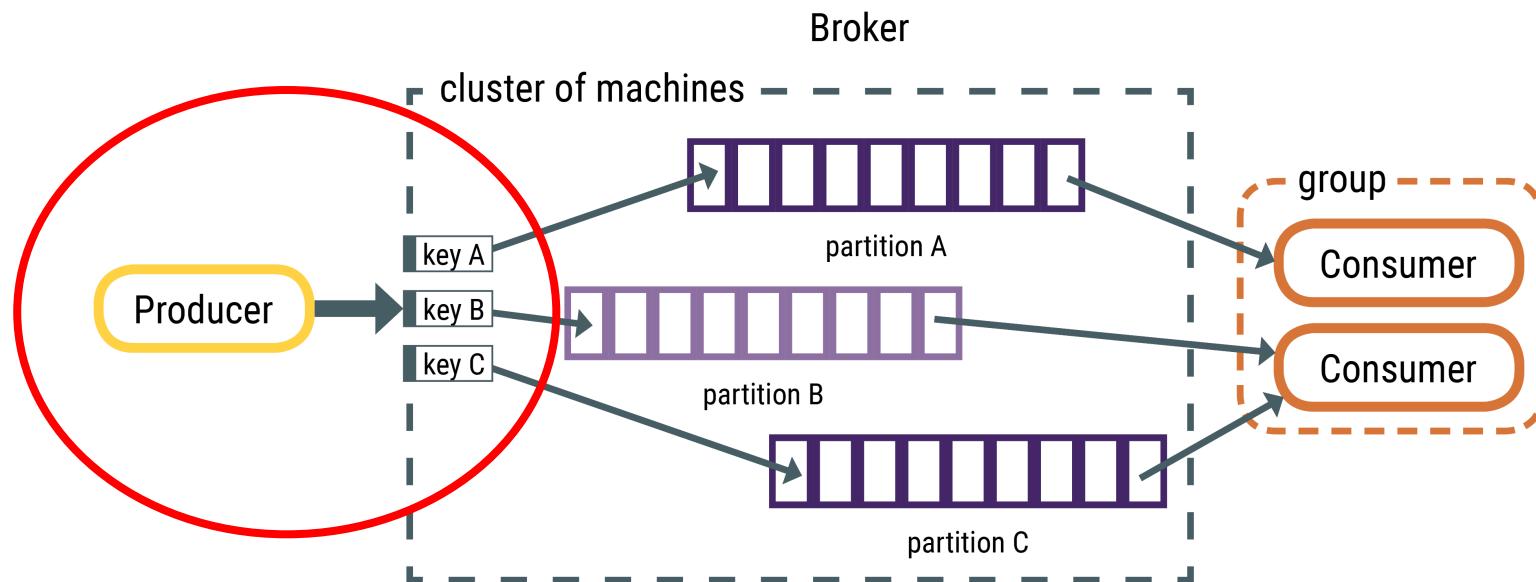
```
./bin/kafka-server-stop.sh
```



Practice

Kafka-console-producer

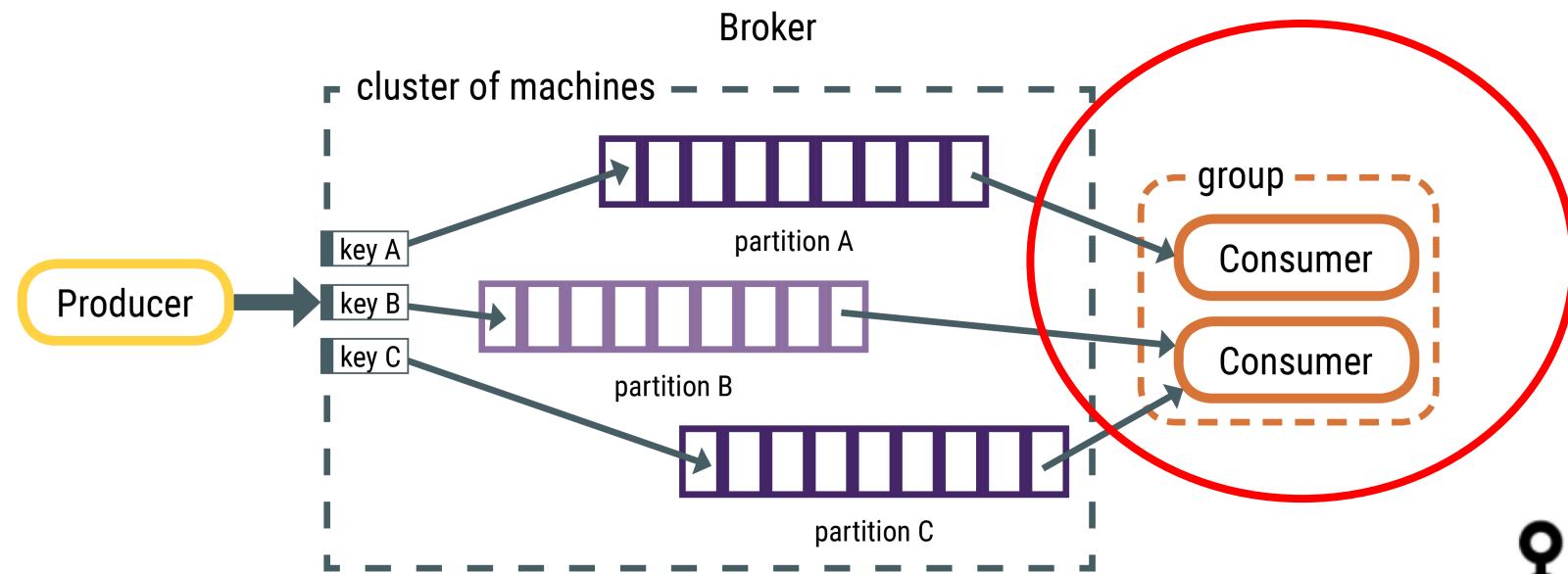
```
./bin/kafka-console-producer.sh --bootstrap-server localhost:9092 --topic topic-one
```



Practice

Kafka-console-consumer

```
./bin/kafka-console-consumer.sh --bootstrap-server localhost:9092\  
--topic topic-one --from-beginning
```





Thank You.

 Nguyen Duc Tien

 +84 36 9589 622

 tiennd24@viettel.com.vn

 Data Governance Department -
Viettel Group