Apache Kafka

Chu-Pan Wong

Aug 30, 2019

Stream Processing

- The ability to work with an infinite stream of data with continuous computation as it flows.
- Stream processing v.s. batch processing

Use Cases

- Credit card fraud detection
- Intrusion detection
- Financial industry?


The
New York
Times


adidas


LINE


airbnb


box


coursera


LinkedIn

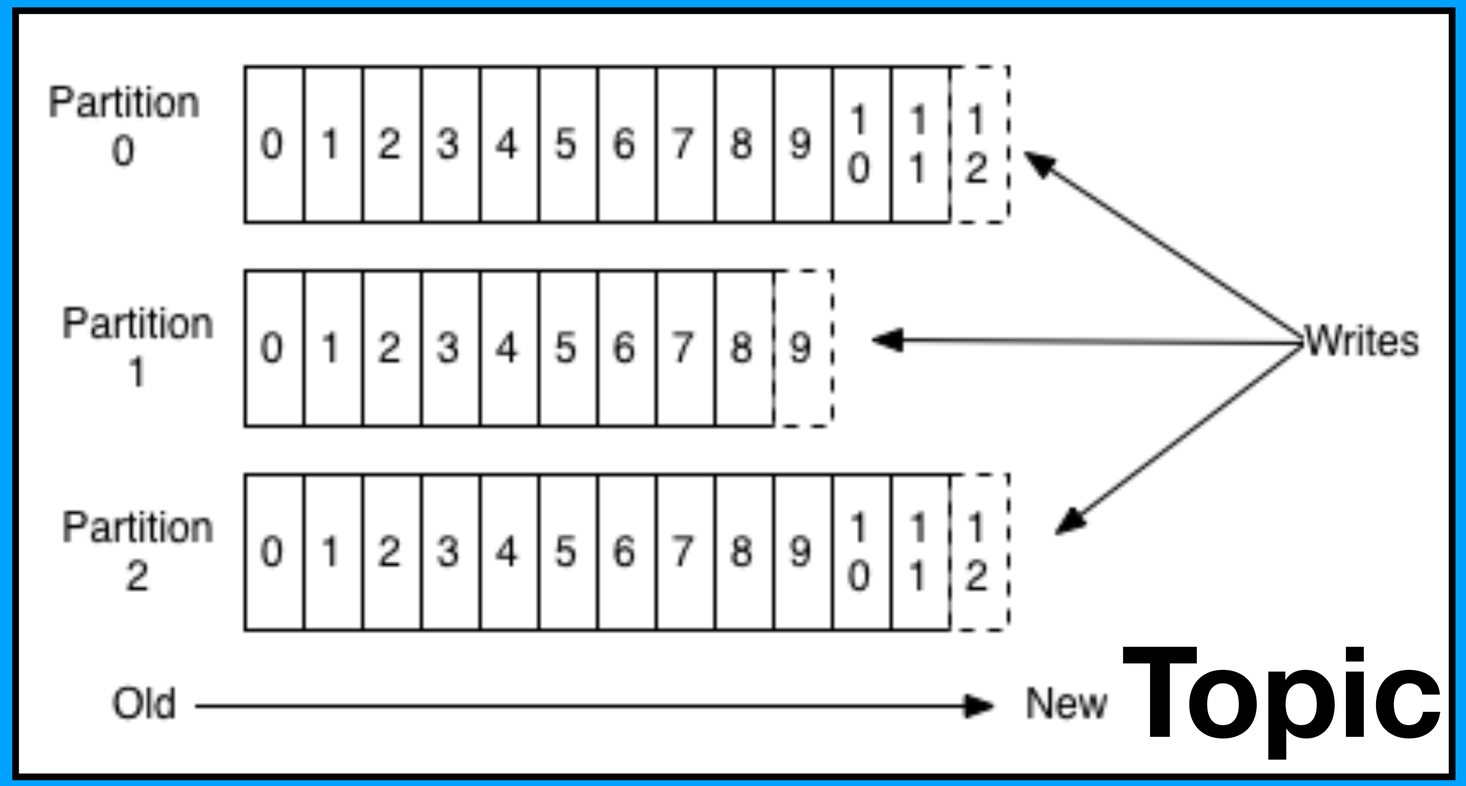

CISCO


NETFLIX


ORACLE

Kafka Stream
in next assignment…

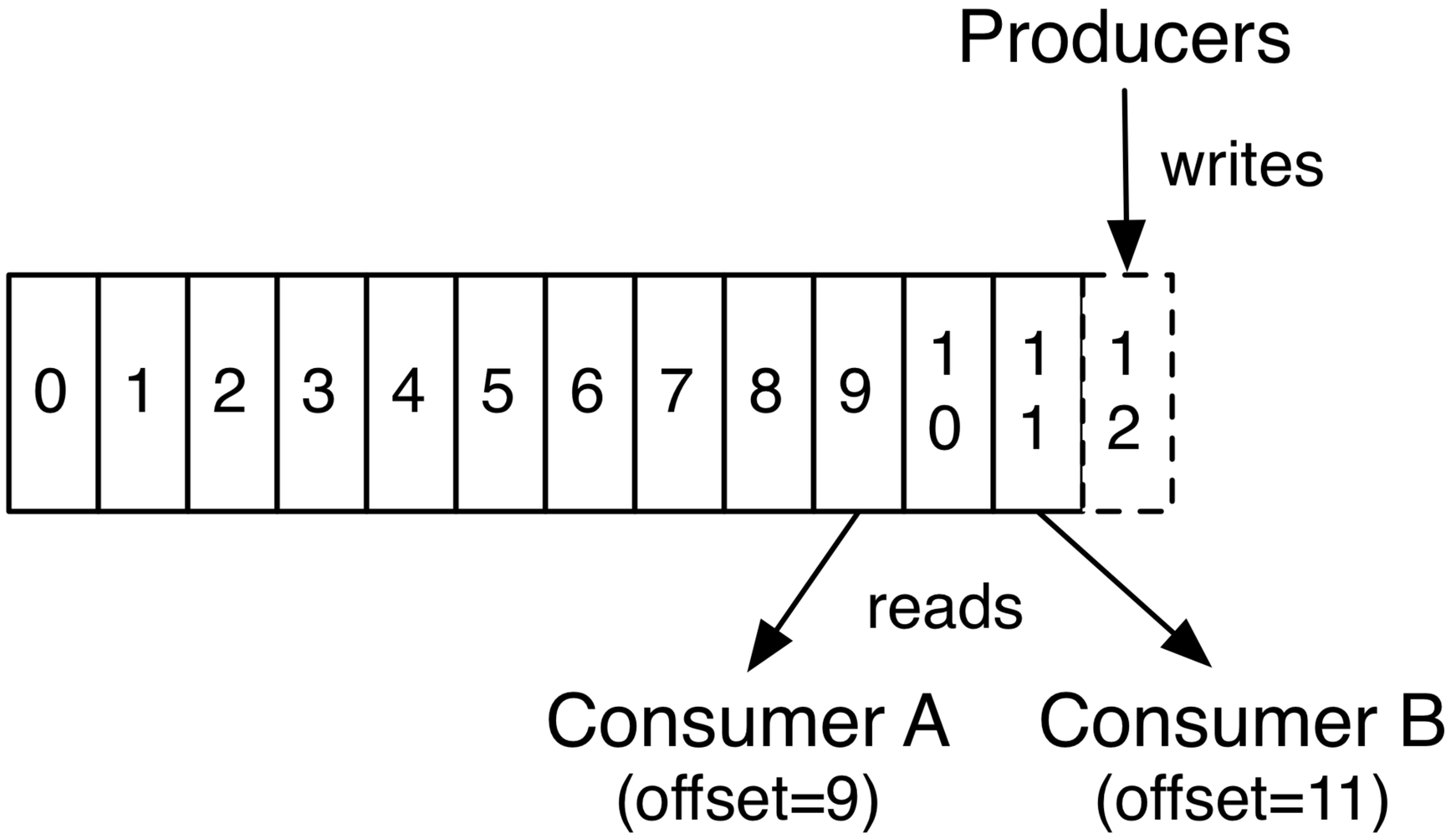
Cluster, Topic, Partition



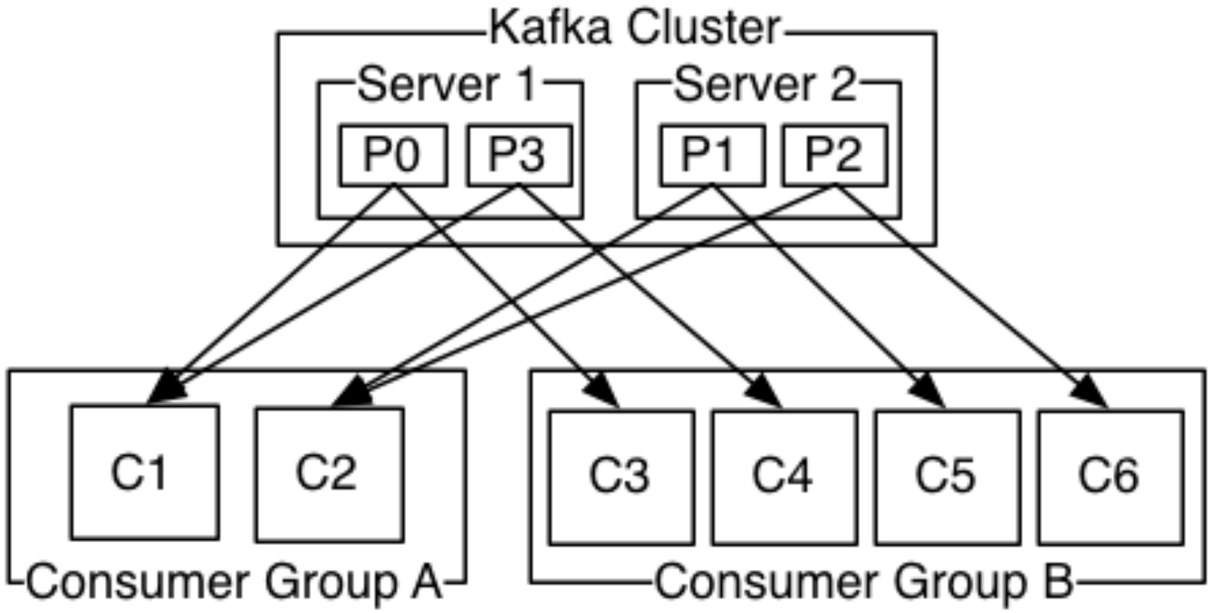
Cluster

- **Clusters** and **topics** are designed to store data across machines
- **Partitions** are for load balancing and scalability. Decided by hashing or simple round robin.

Producer and Consumer



- Logically related **Consumers** can be grouped into a **Consumer Group**
- Each message in a **topic** is delivered to each **consumer group**, but only one **consumer** can get it.



Guarantees

- Messages sent by a producer to a particular topic partition will be appended in the order they are sent. That is, if a record M1 is sent by the same producer as a record M2, and M1 is sent first, then M1 will have a lower offset than M2 and appear earlier in the log.
- A consumer instance sees records in the order they are stored in the log.
- For a topic with replication factor N, we will tolerate up to N-1 server failures without losing any records committed to the log.

Tasks

T1: Basic Commands

- Download Kafka 2.3.0
  - https://kafka.apache.org/downloads
- `bin/kafka-console-consumer.sh --bootstrap-server jenkins-vbc.isri.cmu.edu:9092 --topic commands —from-beginning`
- (keep this terminal open)

kafkacat

- Install Kafkacat (e.g., `brew install kafkacat`)
- `kafkacat -b jenkins-vbc.isri.cmu.edu:9092 -L`

Chatroom Log

- Go to jenkins-vbc.isri.cmu.edu:9000 in your favorite browser
- log-stream

T2: Split Line

- Clone https://github.com/chupanw/kafka-stream.git
- Open/Import the project with your favorite IDE/editors
- Checkout src/java/myapps/LineSplit.java and finish the TODOs
- To run
  - `mvn clean package`
  - `mvn exec:java -Dexec.mainClass=myapps.LineSplit`

T3: Content Control

- Create a new file ContentControl.java
- Find lines that contain the word ‘dirty’
- Append your Andrew ID to those lines, and pipe to the topic ‘dirty-filter’
- To run
  - `mvn clean package`
  - `mvn exec:java -Dexec.mainClass=myapps.ContentControl`

More on Kafka

- Messaging
- Website Activity Tracking
- Metrics
- Log Aggregation
- Commit Log