Leveraging Apache Kafka for Web Crawling and Data Processing

OpenStack Cork, 2017-11-21

Image: https://softwareengineeringdaily.com/wp-content/uploads/2015/08/kafka-logo-wide.png

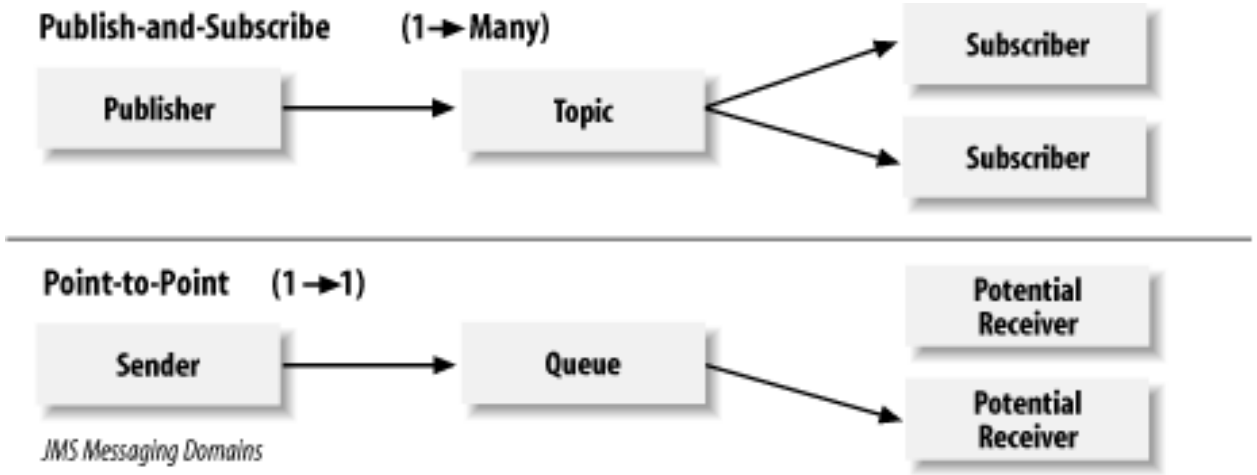
About Me

- Johannes Ahlmann
- fluquid.com
  - Sales & Client Intelligence
  - Intelligent Lead Generation
  - Large-scale web crawls
  - Gathering and Enriching Web Data
- webdata.org
  - Share Libraries and Best Practices
  - Bring Data Scientists and SME Companies together
  - ForDevelopers
  - AwesomeAvailableDatasets
- Contact:
  johannes@fluquid.com

Background: Queues & PubSub

- PubSub
  - multiple subscribers
  - no scaling

- Queues
  - multiple consumers
  - single-subscriber
  - scaling, load balancing



fluquid

A high-throughput distributed messaging system

- Decouples Data Pipelines
- Scalable & Fault-Tolerant
- Kafka Functionalities
  - Messaging
  - Processing
  - Storing
- Performance (>100k/s)
  - Batching
  - Zero Copy I/O
  - Leverages OS Cache
- Durability

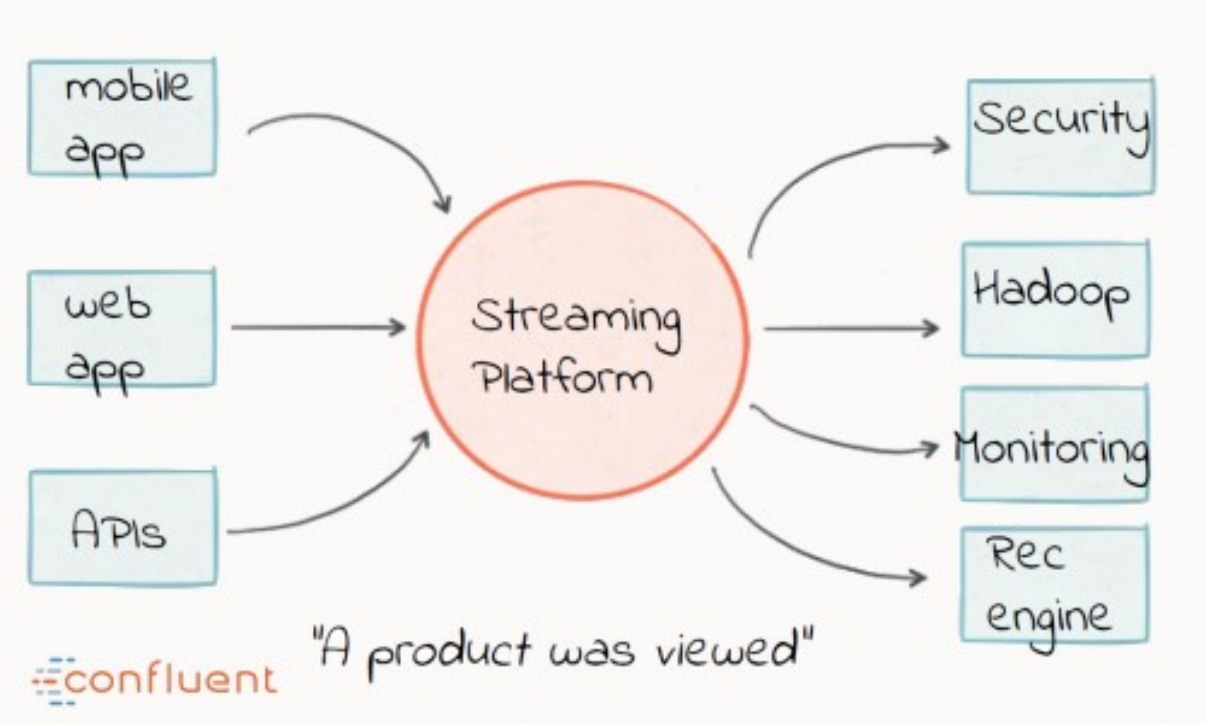


Image Credit: Confluent

fluquid

Core Concepts

- Producers
- Consumers
- Brokers
- Topics
- Zookeeper
  - Offsets
  - Broker Addresses

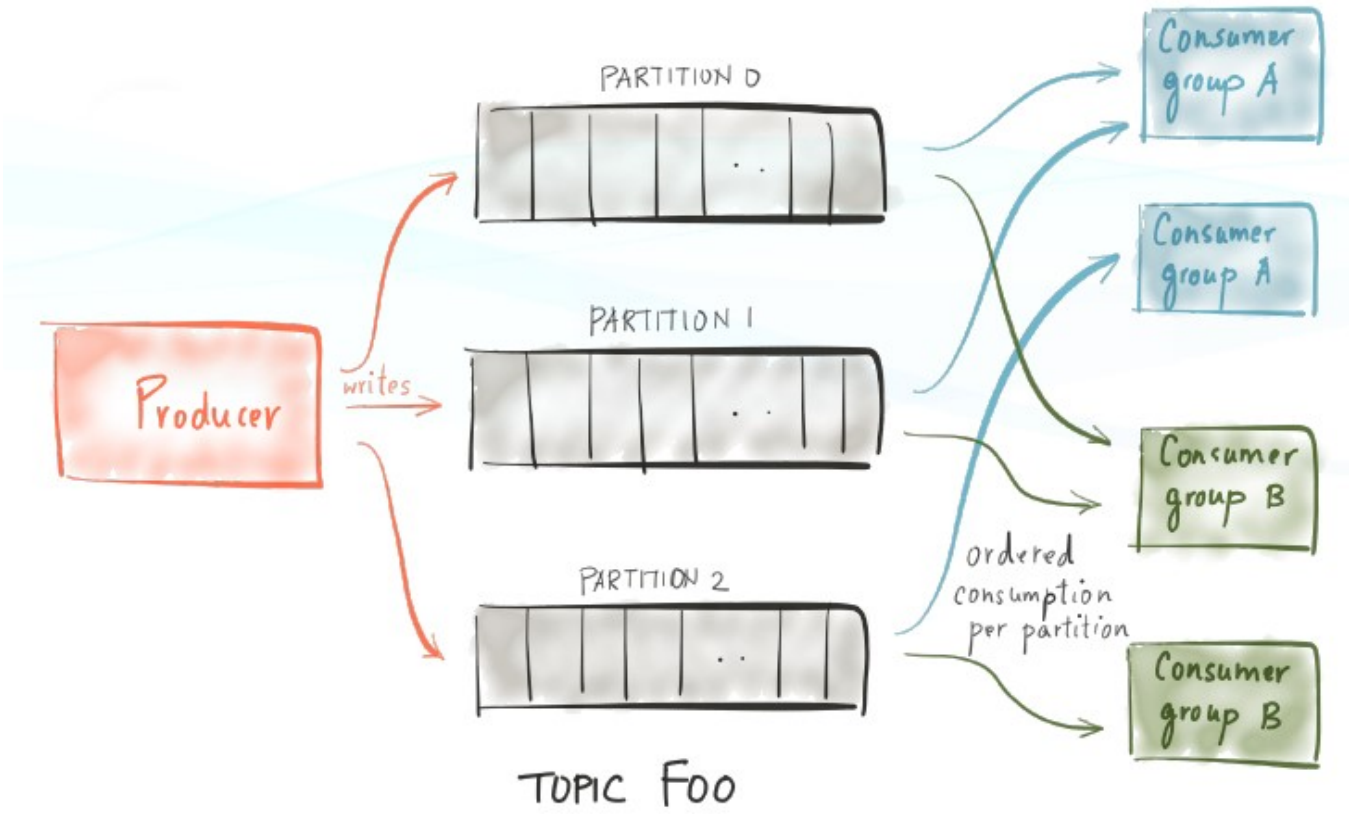
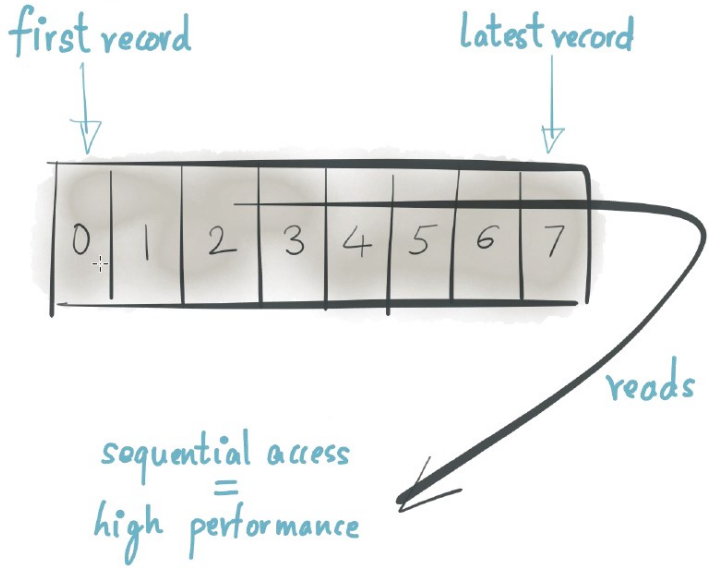


Image Credit: Confluent

Key Idea: Partitioned Log

- Very fast, due to zero copy I/O and batching
- Uses sendfile and OS buffer cache
- Sequential writes to FS
- Order guaranteed within partition
- Scaling



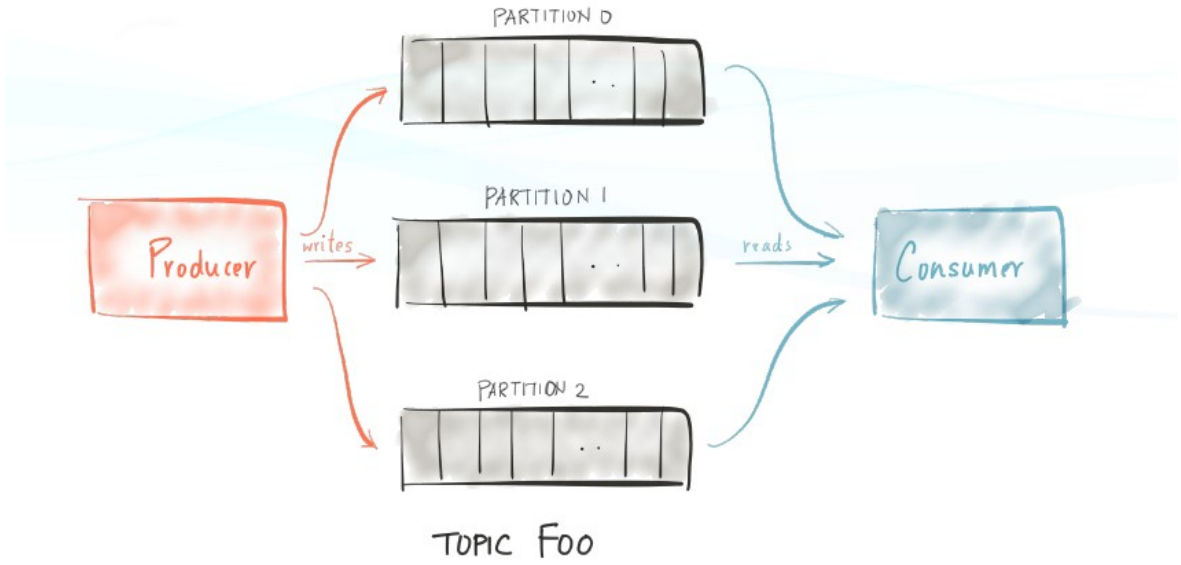
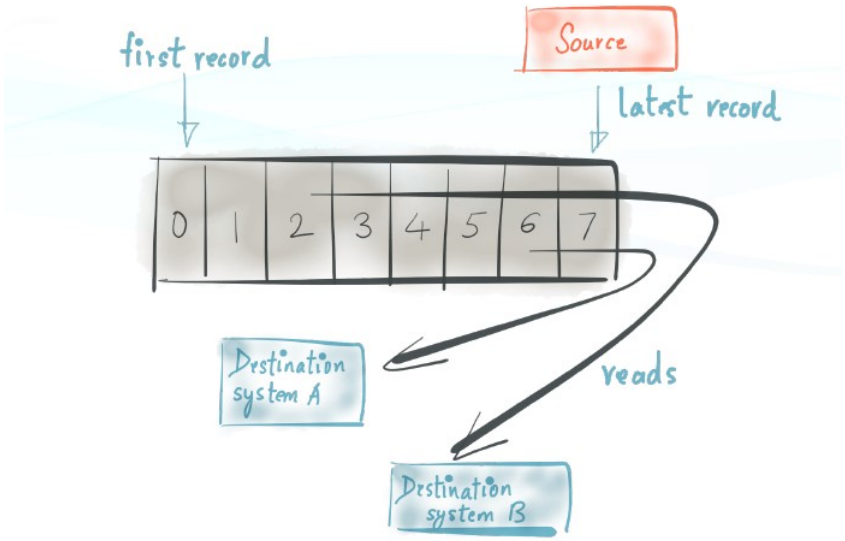


Image Credit: Confluent

fluquid

Logs & PubSub

- Consumers can be transient
- Consumer Groups
- Delivery Semantics
  - at least once (default)
  - at most once
  - exactly once
- Retention Policy
- Reprocessing



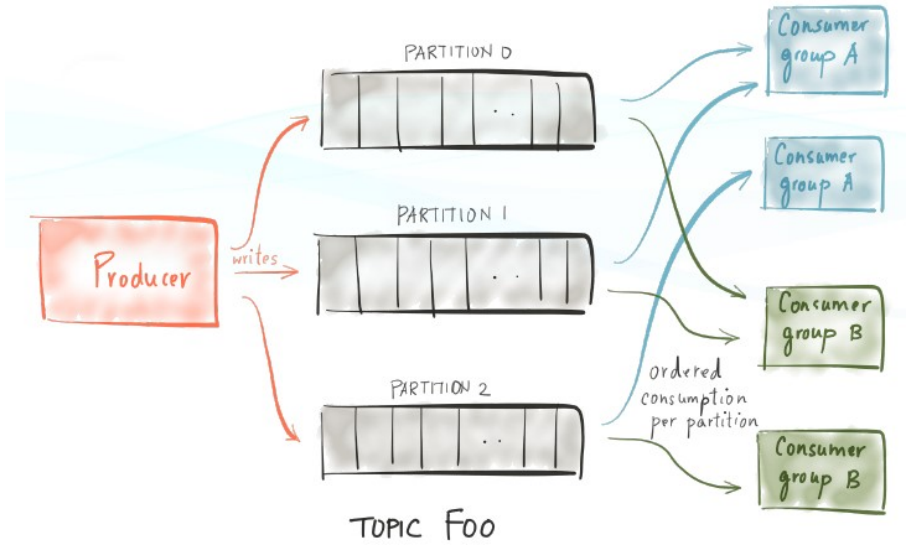


Image Credit: Confluent

fluquid

Partitions & Replication

- Partitions configurable
- Partition allocation
  - round-robin
  - semantic partition by key
- Replication
  - optional
  - 1 leader, 0 or more followers
  - sync or async
  - flush delay configurable

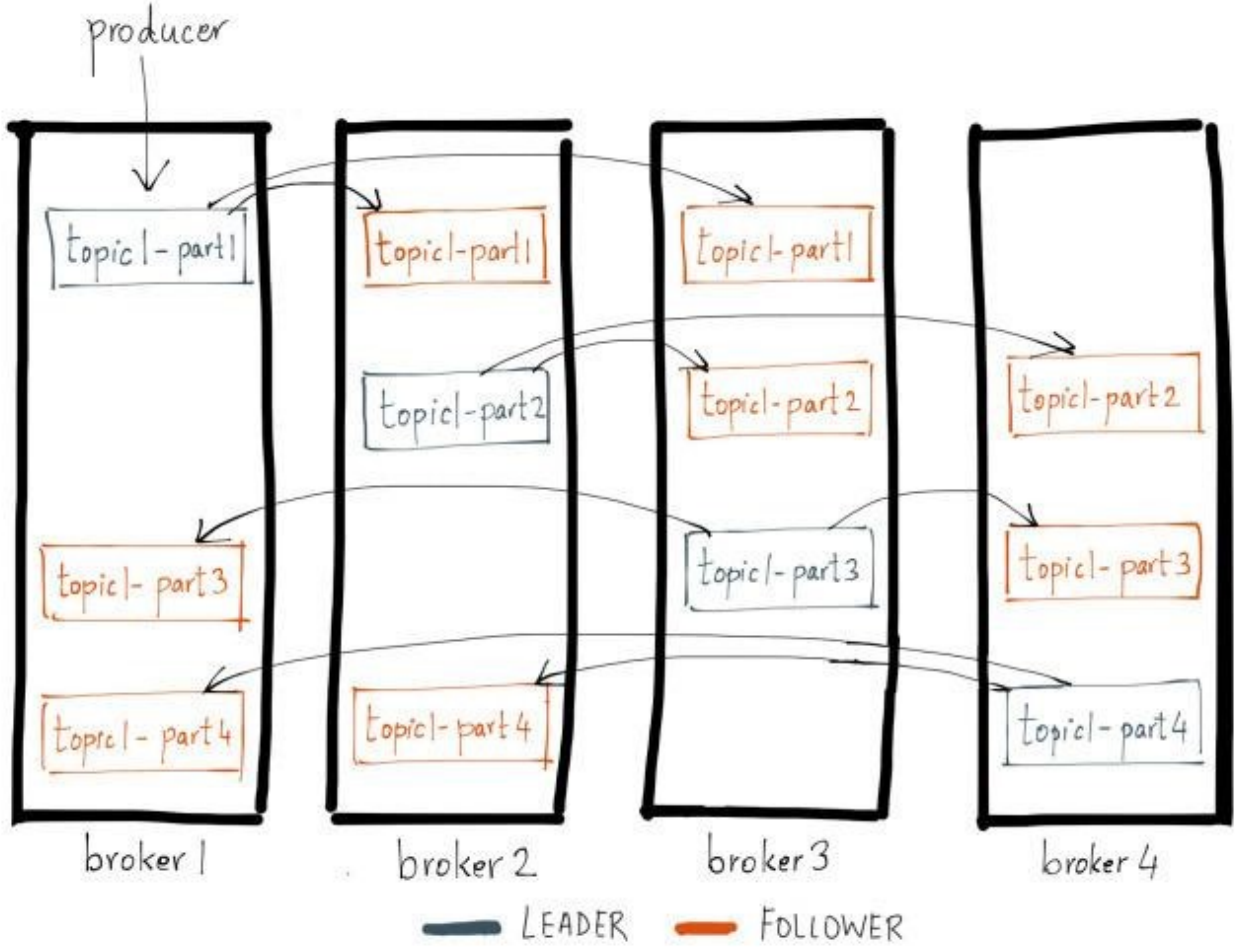


Image Credit: Confluent

Kafka Connect

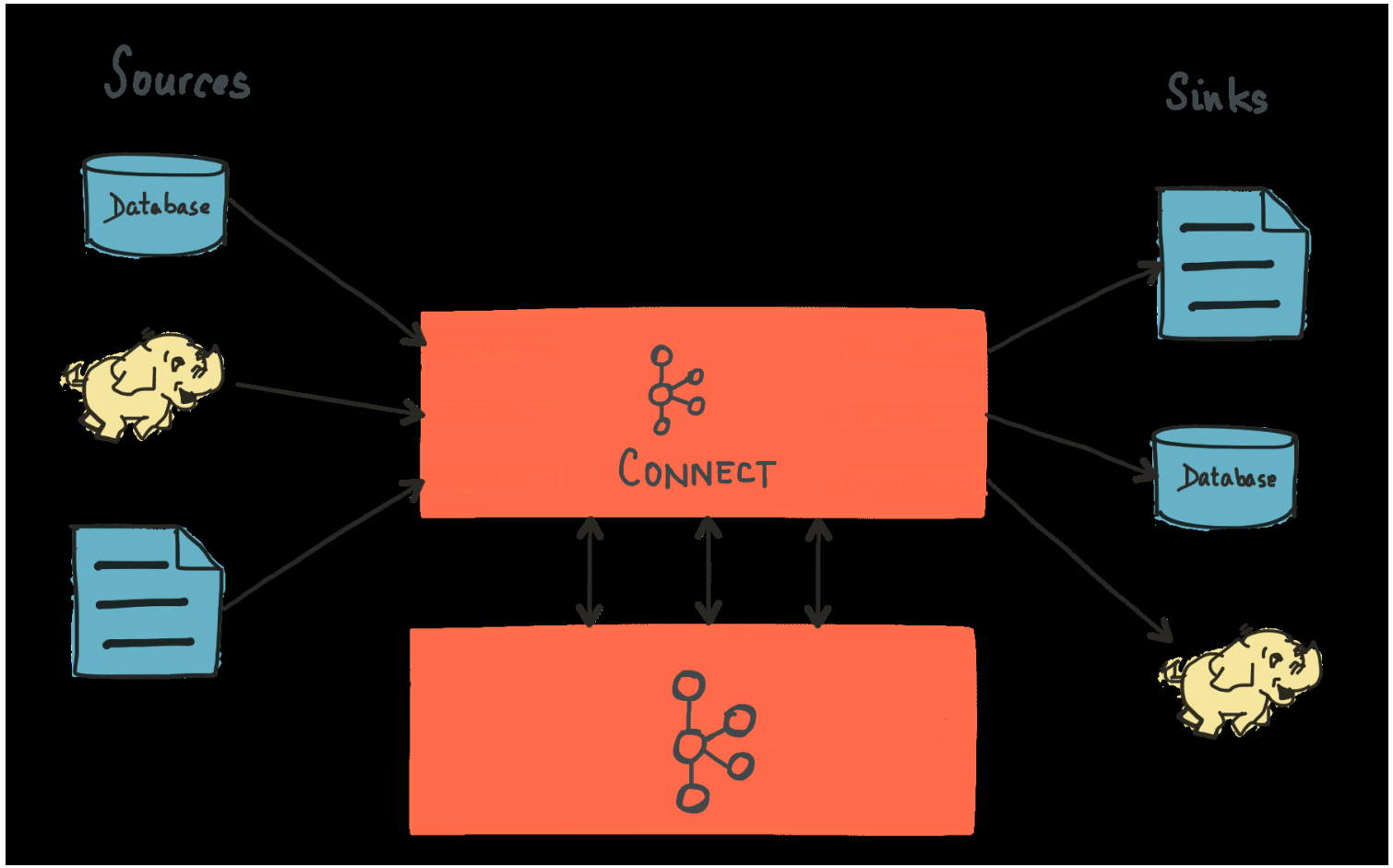


Image Credit: Confluent

Kafka Streams

- Operations
  - filter
  - map
  - join
  - aggregate

  } stateful
- KStream
- KTable
  - manages local state
- Windows

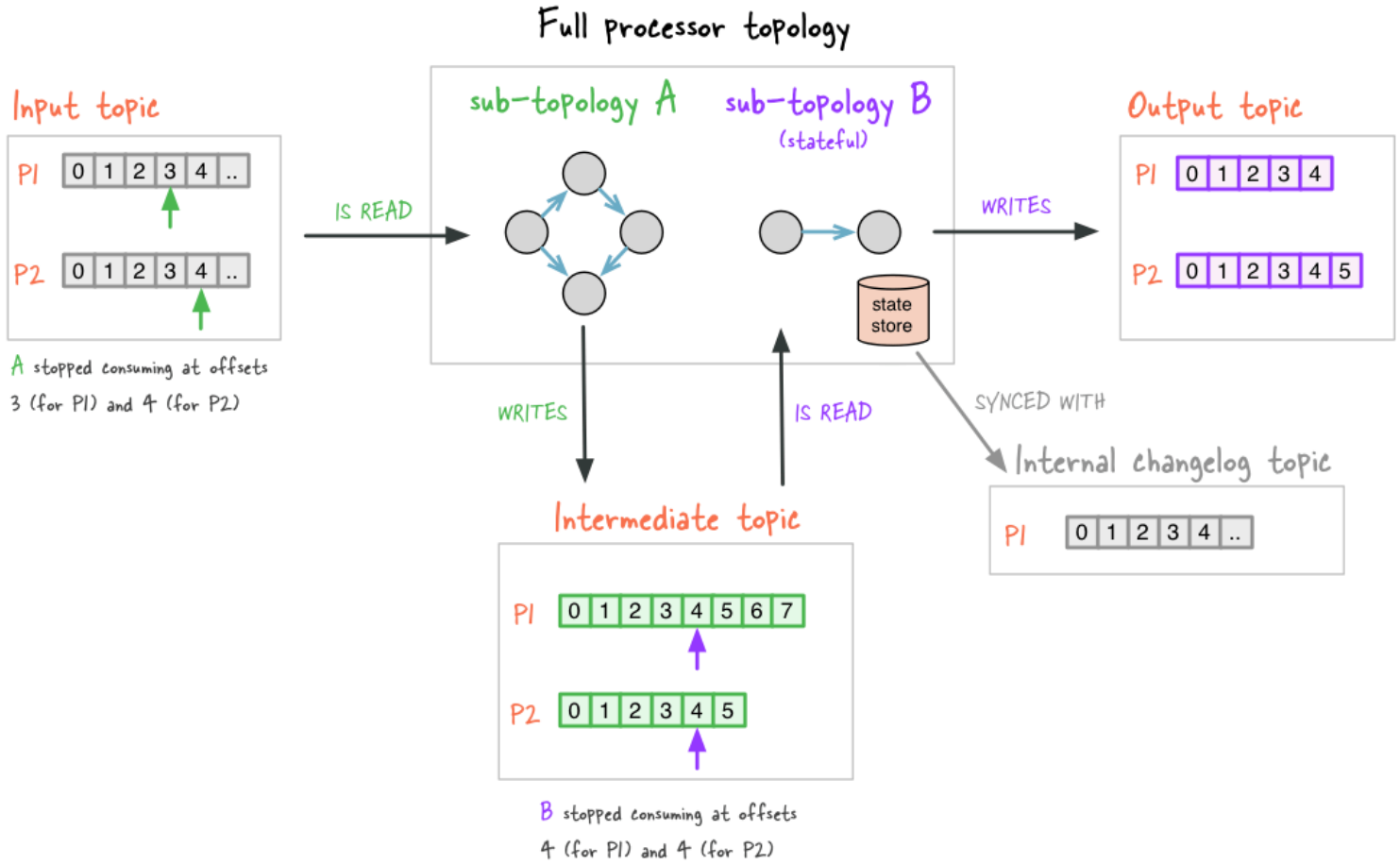


Image Credit: Confluent

fluquid

Summary

scalability of a **filesystem**

- hundreds of MB/s
- many TBs per server
- commodity hardware

guarantees of a **database**

- persistence
- ordering

distributed by **design**

- replication
- partitioning
- horizontal scalability
- fault tolerance

Image Credit: Confluent

fluquid