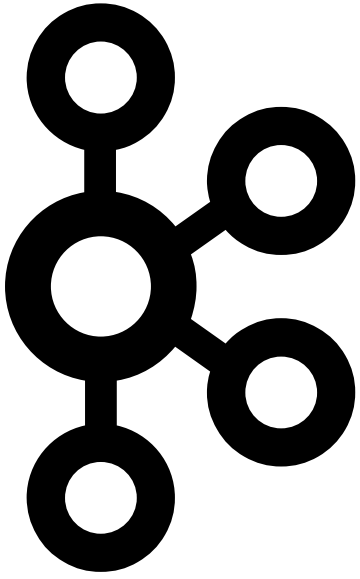


Apache Kafka 101

Samuel Costa - 43552

ISEL
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Challenges

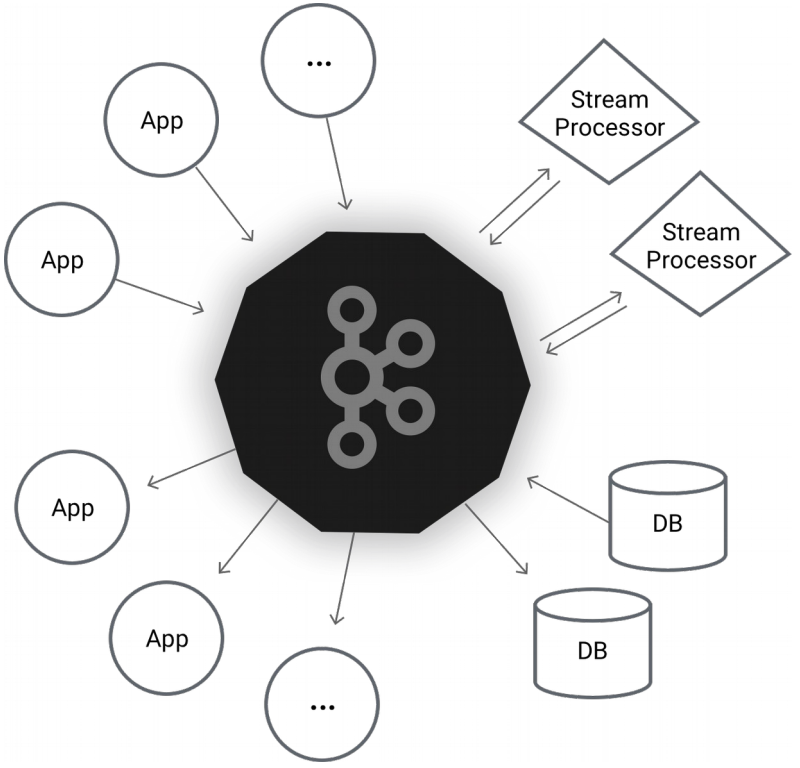
- Integrate complex systems that interact in complicated ways
- Assembling data that has multiple sources
- Synchronization between production and consumption
- Horizontal scaling
- Current patterns include:
  - Event-driven architectures
  - Save event history for re-processing
  - Scaling not done by adding faster hardware

Features

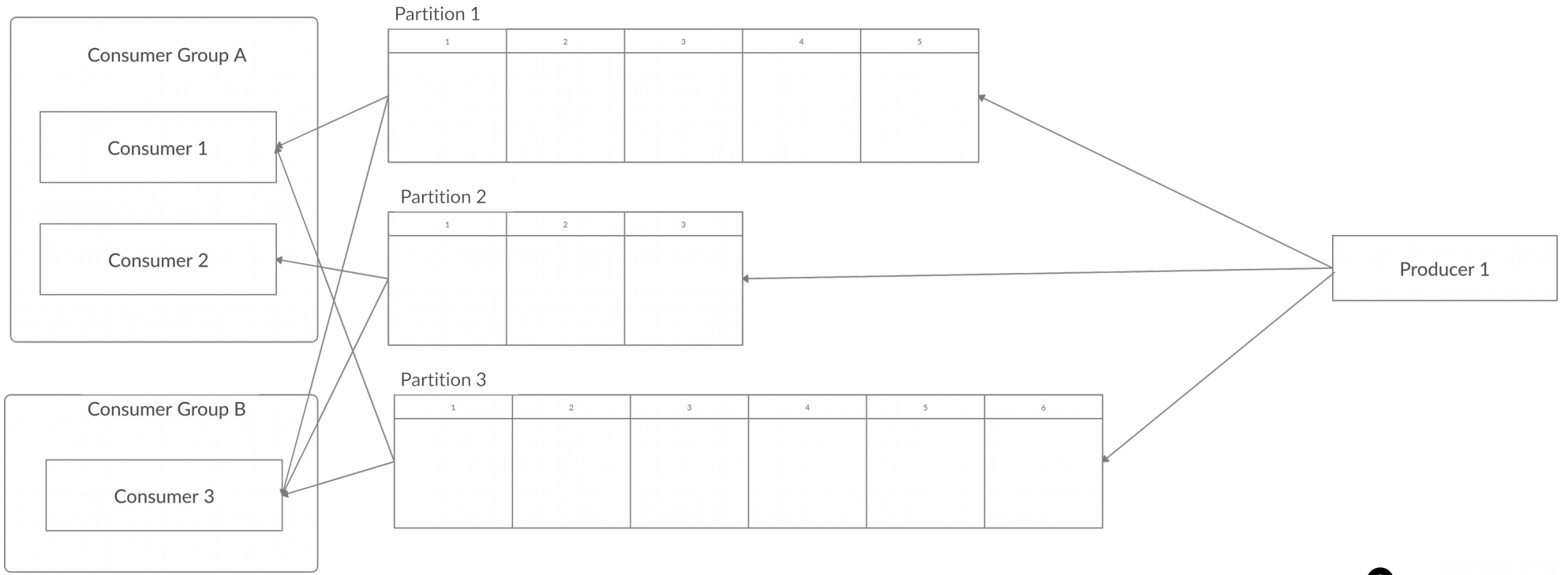
- Pub & Sub
- Process
- Store
- Ordering
- Distributed
- Redundant

Architecture

- Producers
- Consumers
- Connectors
- Stream Processors
- Topics
- Partitions
- Consumer Groups



Anatomy of a Topic



Samuel Costa - 43552

Takeaways

- Records are composed of key (sequential number), value (byte array) and timestamp
- Replication factor is not the same as number of partitions
- Attention to ordering guarantees!

Who uses it

NETFLIX

The New York Times

airbnb

ISEL
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Demo

Commands used in this presentation are available at

sscosta/kafka-demo

Further References

- Narkhede, Neha, et al. Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale. First edition, O'Reilly Media, 2017
- "Benchmarking Apache Kafka: 2 Million Writes Per Second (On Three Cheap Machines)" - https://engineering.linkedin.com/kafka/benchmarking-apache-kafka-2-million-writes-second-three-cheap-machines
- Netflix Tech Blog - "Kafka Inside Keystone Pipeline" - https://netflixtechblog.com/kafka-inside-keystone-pipeline-dd5aeabaf6bb