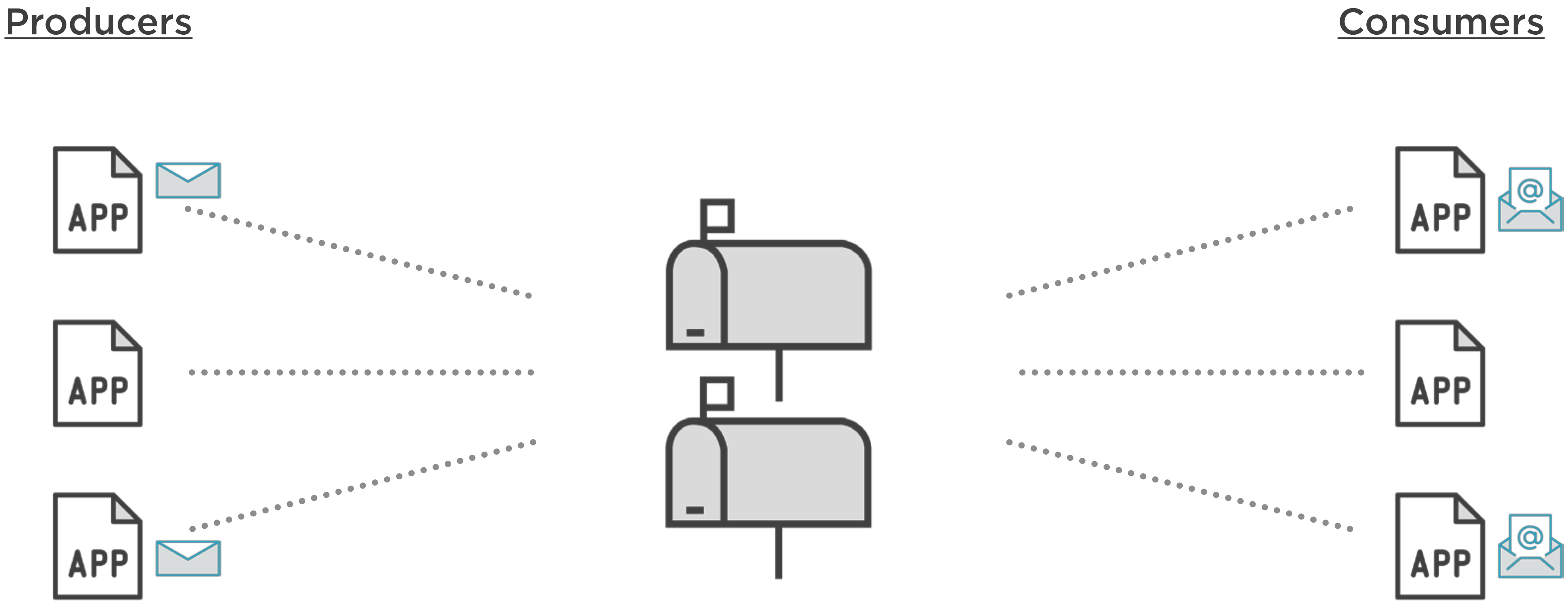
Getting to Know Apache Kafka’s Architecture

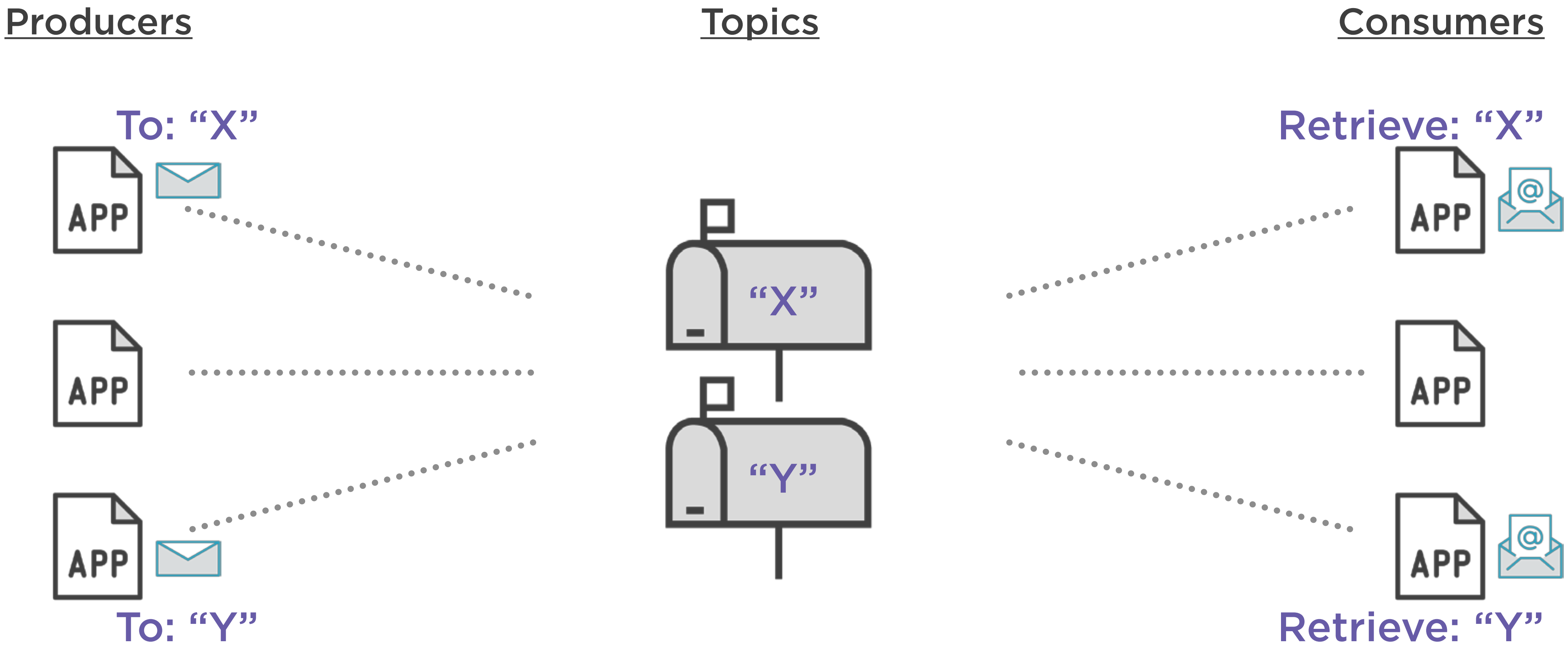
Ryan Plant
COURSE AUTHOR

@ryan_plant blog.ryanplant.com

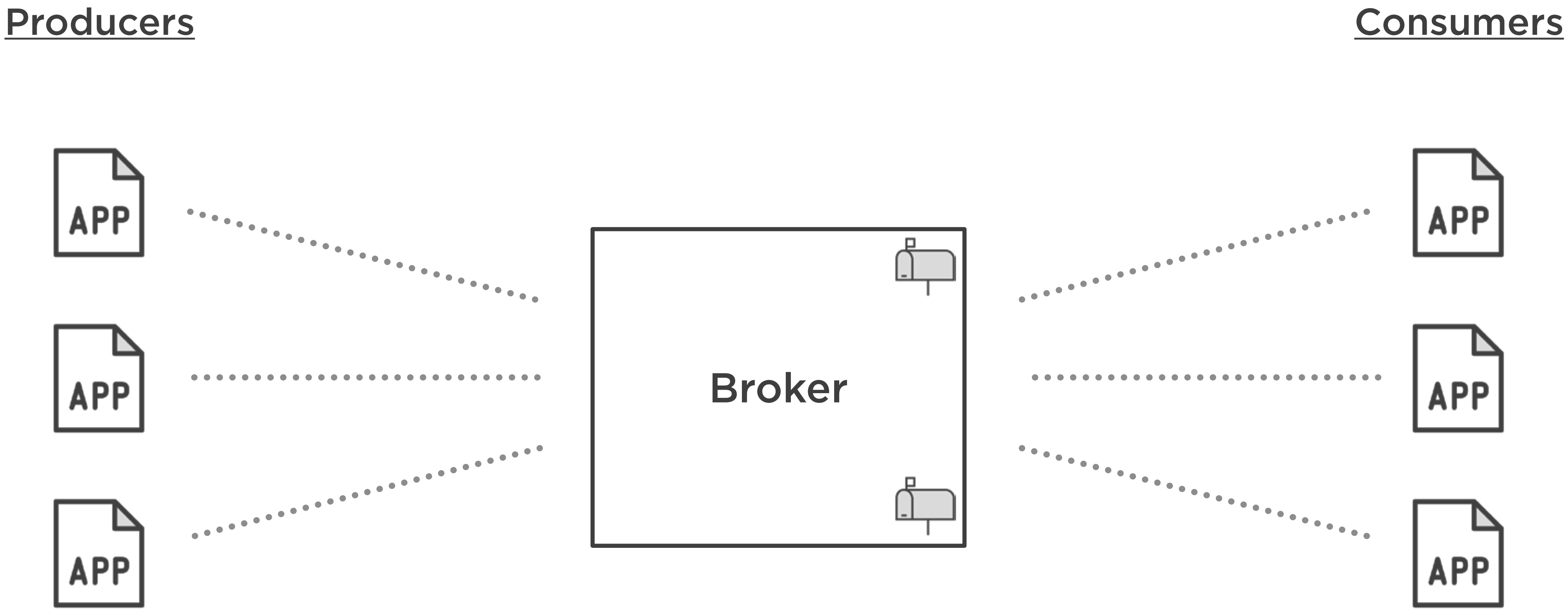
Apache Kafka as a Messaging System



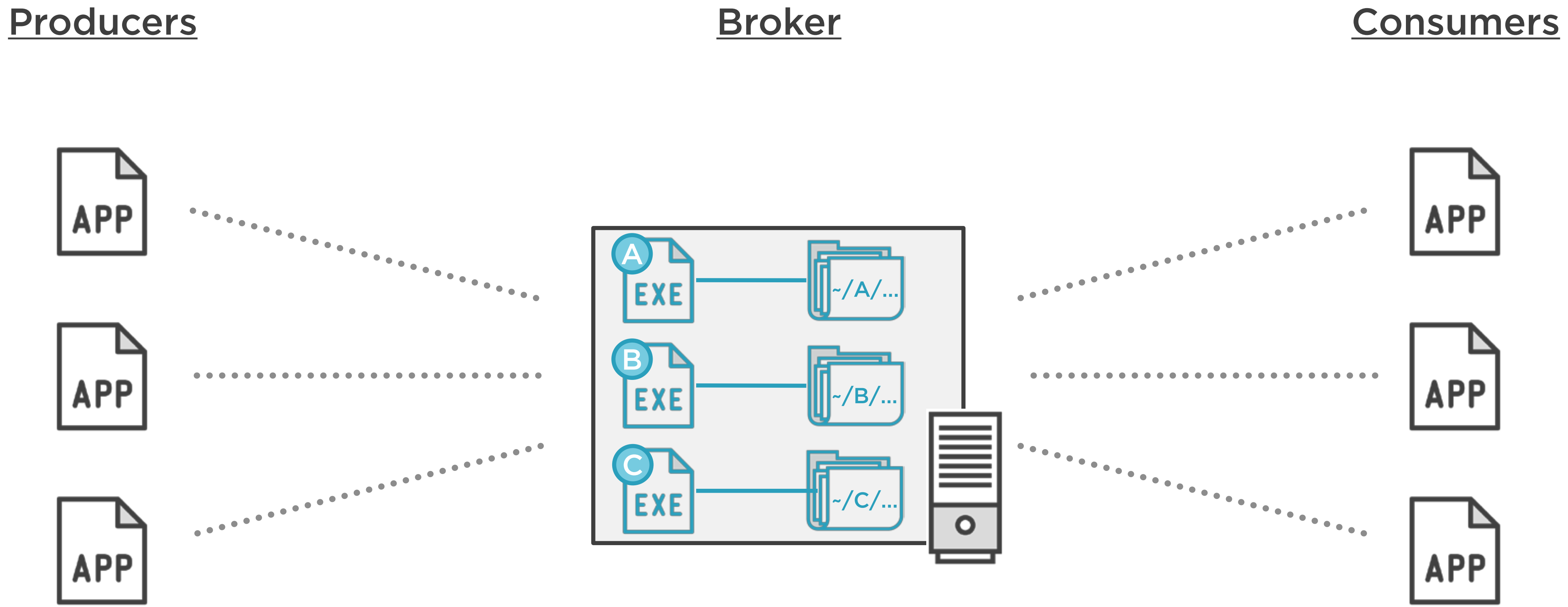
Apache Kafka as a Messaging System



Apache Kafka as a Messaging System



Apache Kafka as a Messaging System

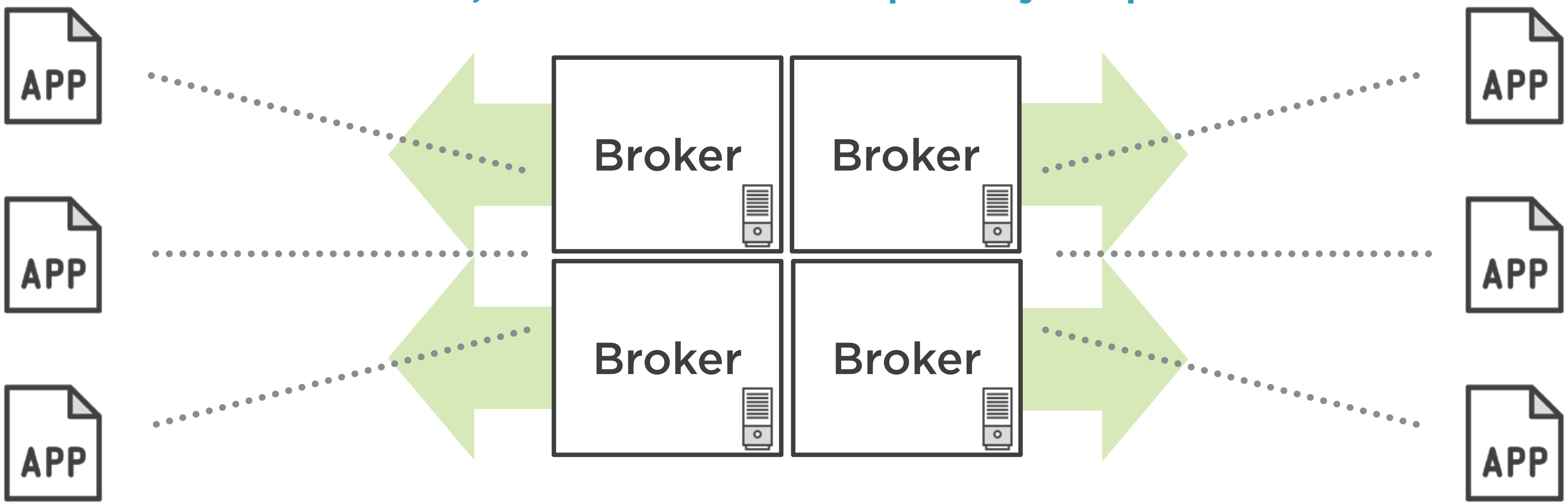


How Apache Kafka Starts to Differentiate

Producers

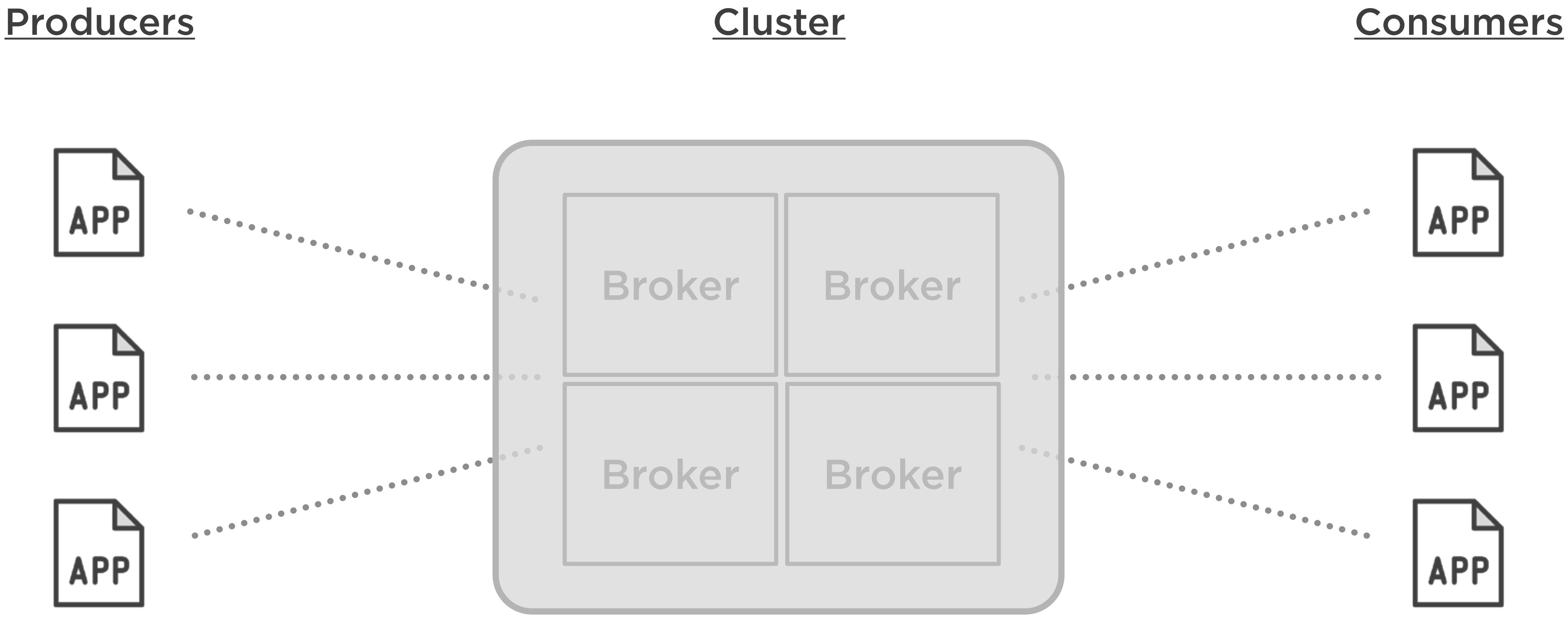
Consumers

LinkedIn: 1,400 brokers => 2 petabytes per week

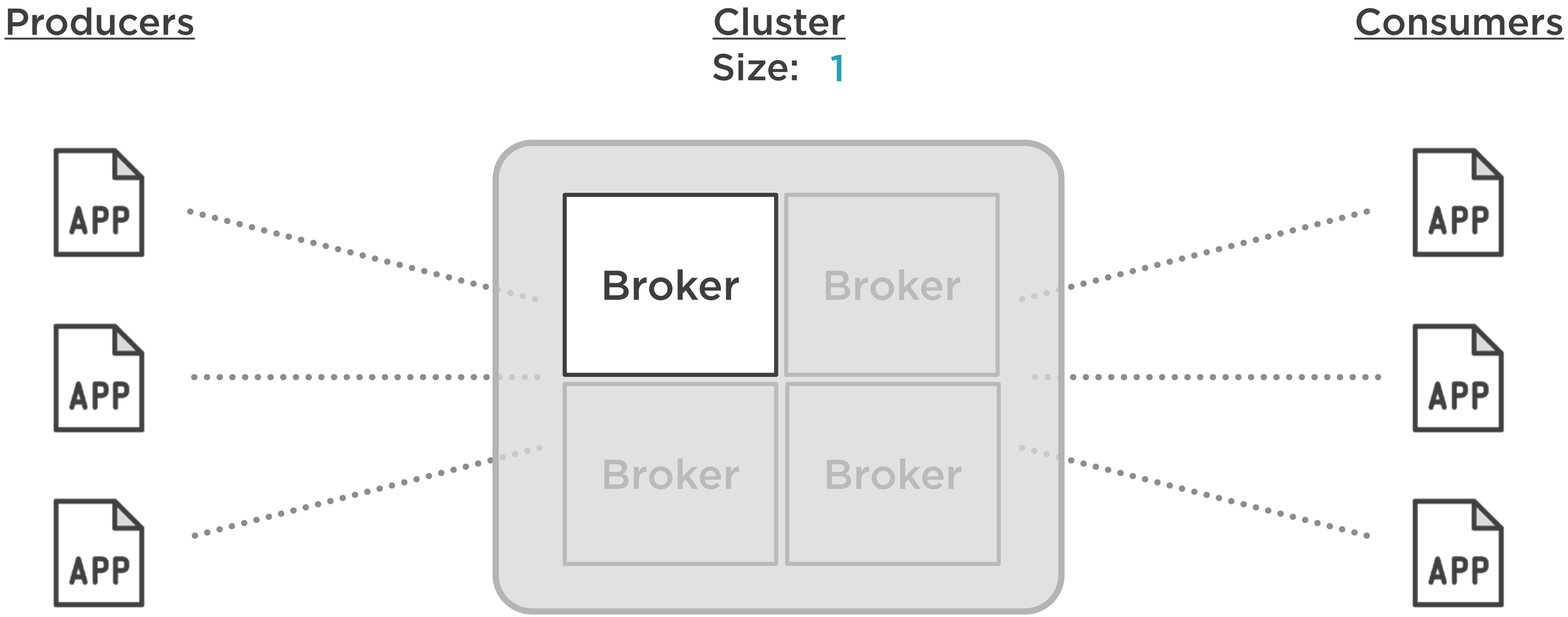


“A high-throughput distributed messaging system.”

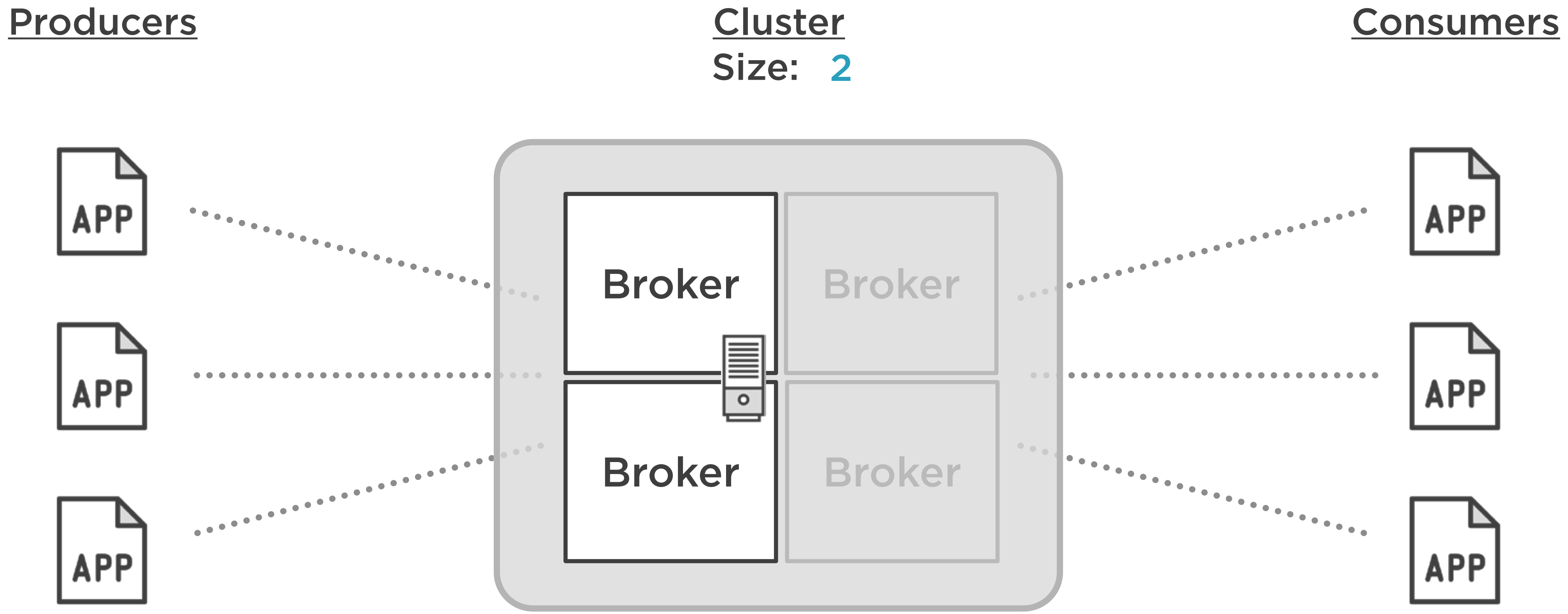
The Apache Kafka Cluster



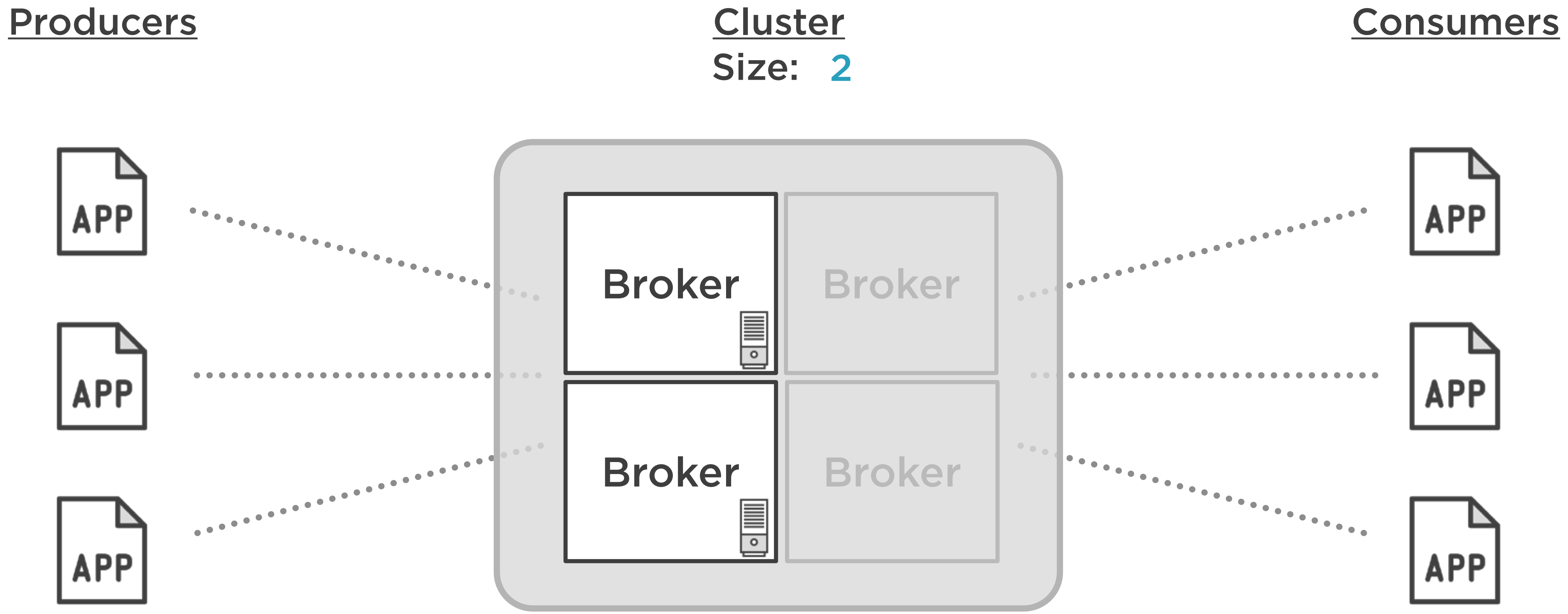
The Apache Kafka Cluster



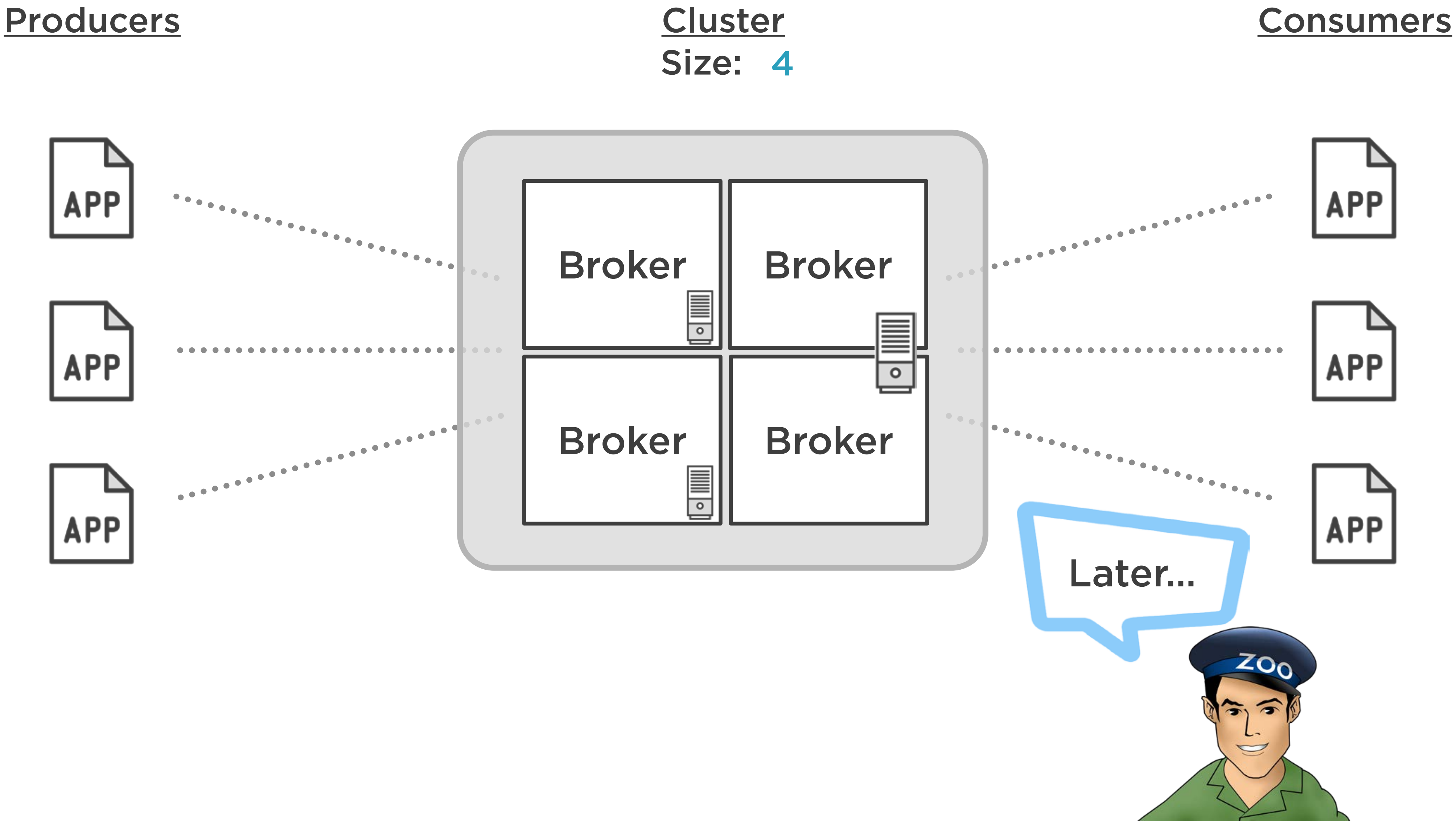
The Apache Kafka Cluster



The Apache Kafka Cluster



The Apache Kafka Cluster



Distributed Systems

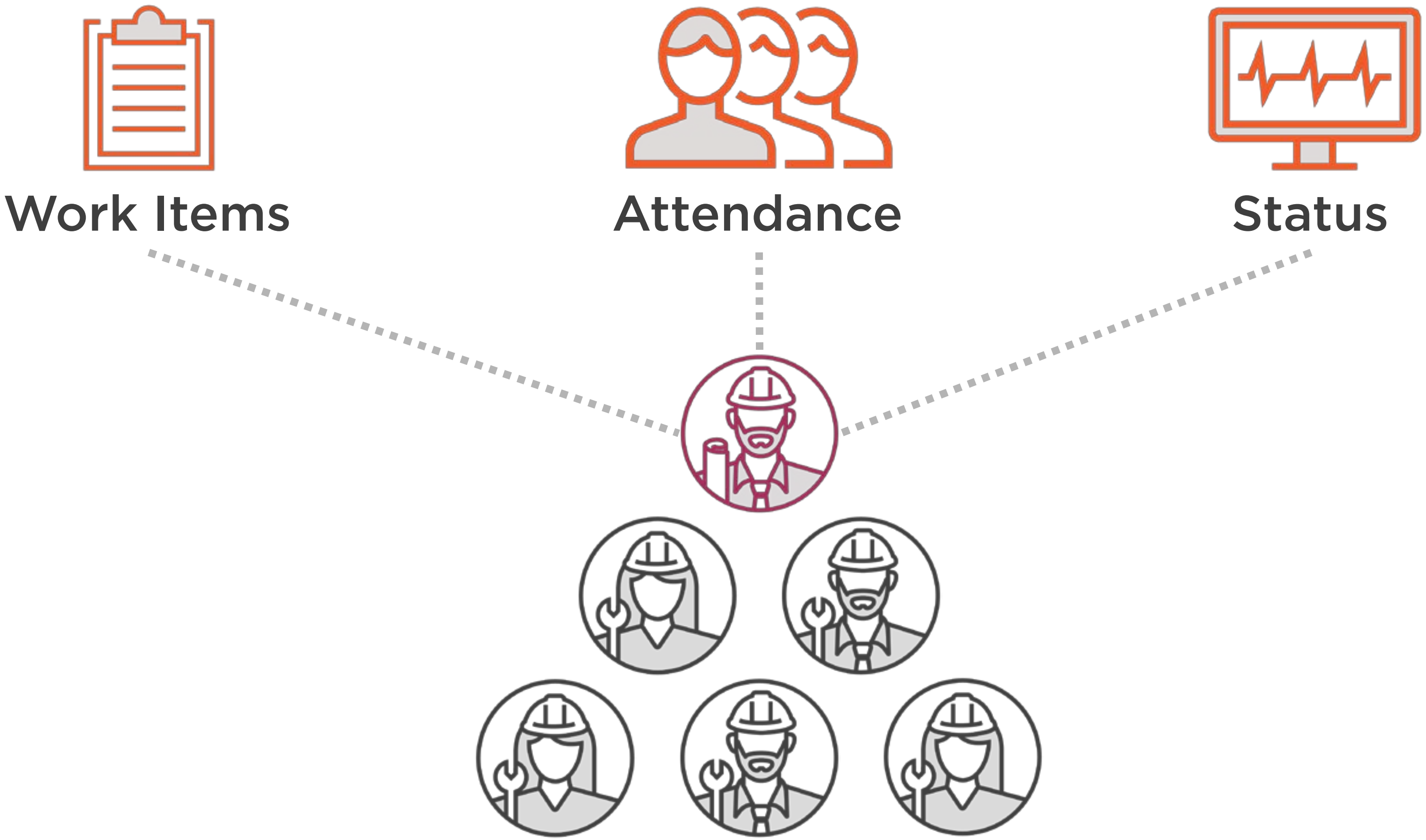
Collection of resources that are instructed to achieve a specific goal or function

Consist of multiple workers or nodes

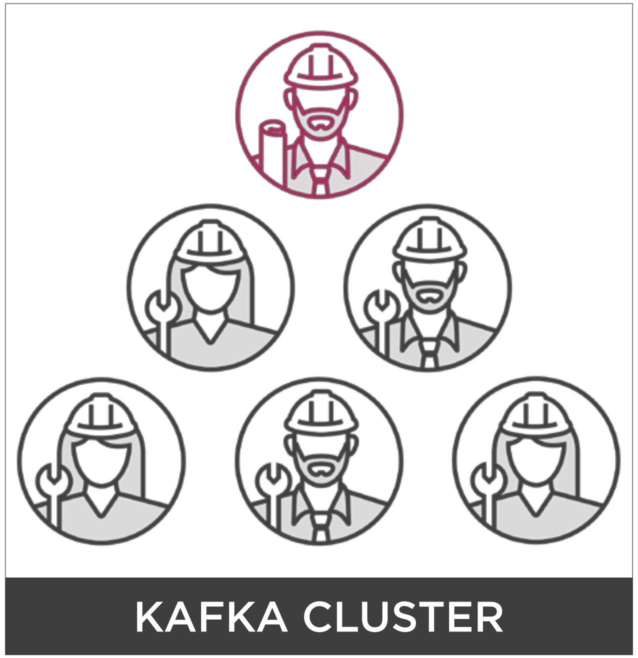
The system of nodes require coordination to ensure consistency and progress towards a common goal

Each node communicates with each other though messages

Distributed Systems: Controller Election



Distributed Systems: The Cluster



KAFKA CLUSTER

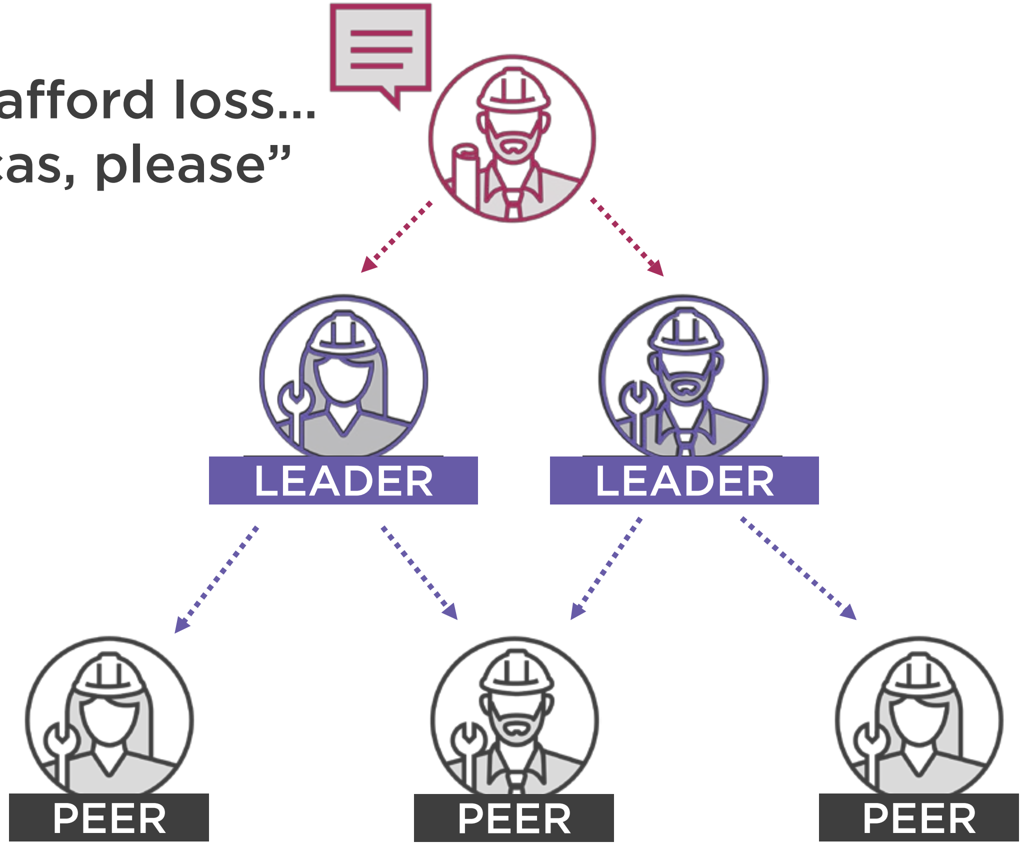
Distributed Systems: Getting Work Done

Worker availability and health

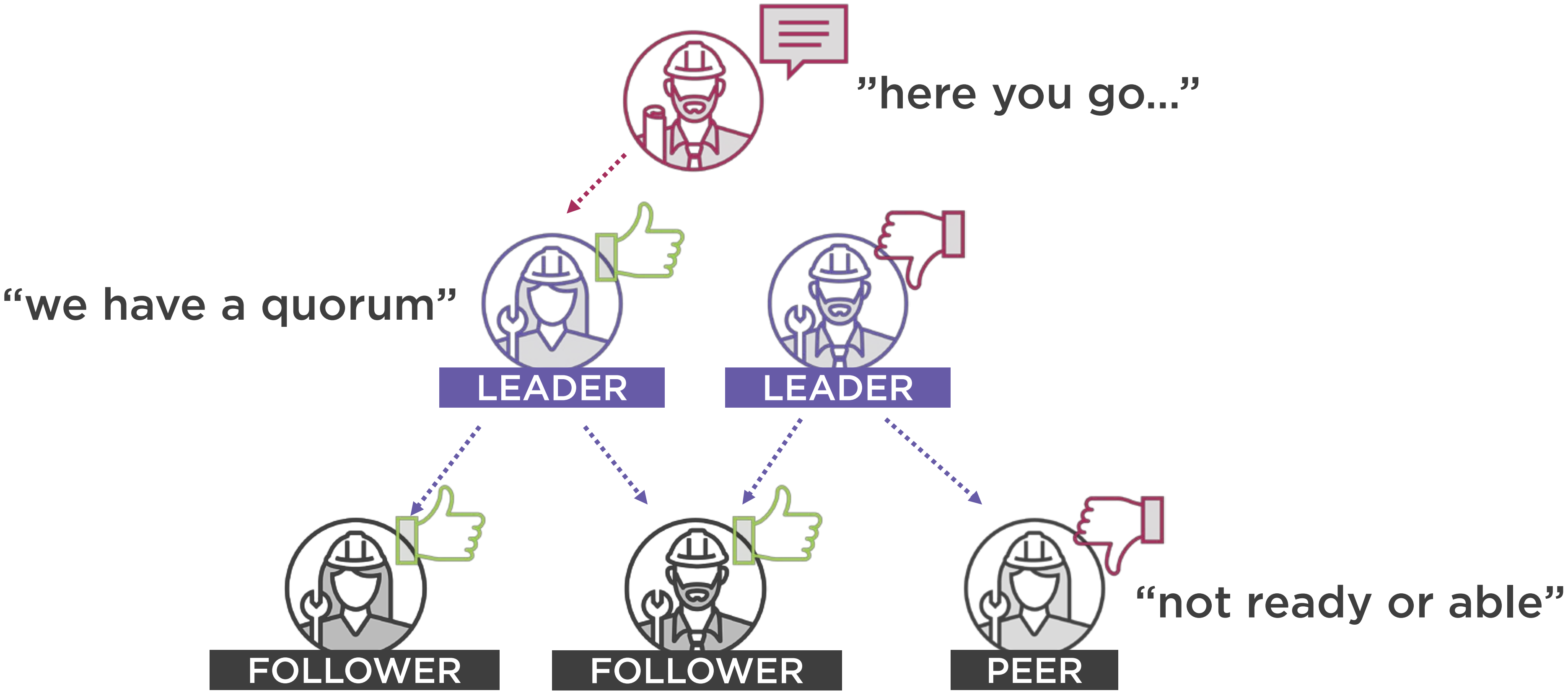
Task redundancy

Distributed Systems: Getting Work Done (Reliably)

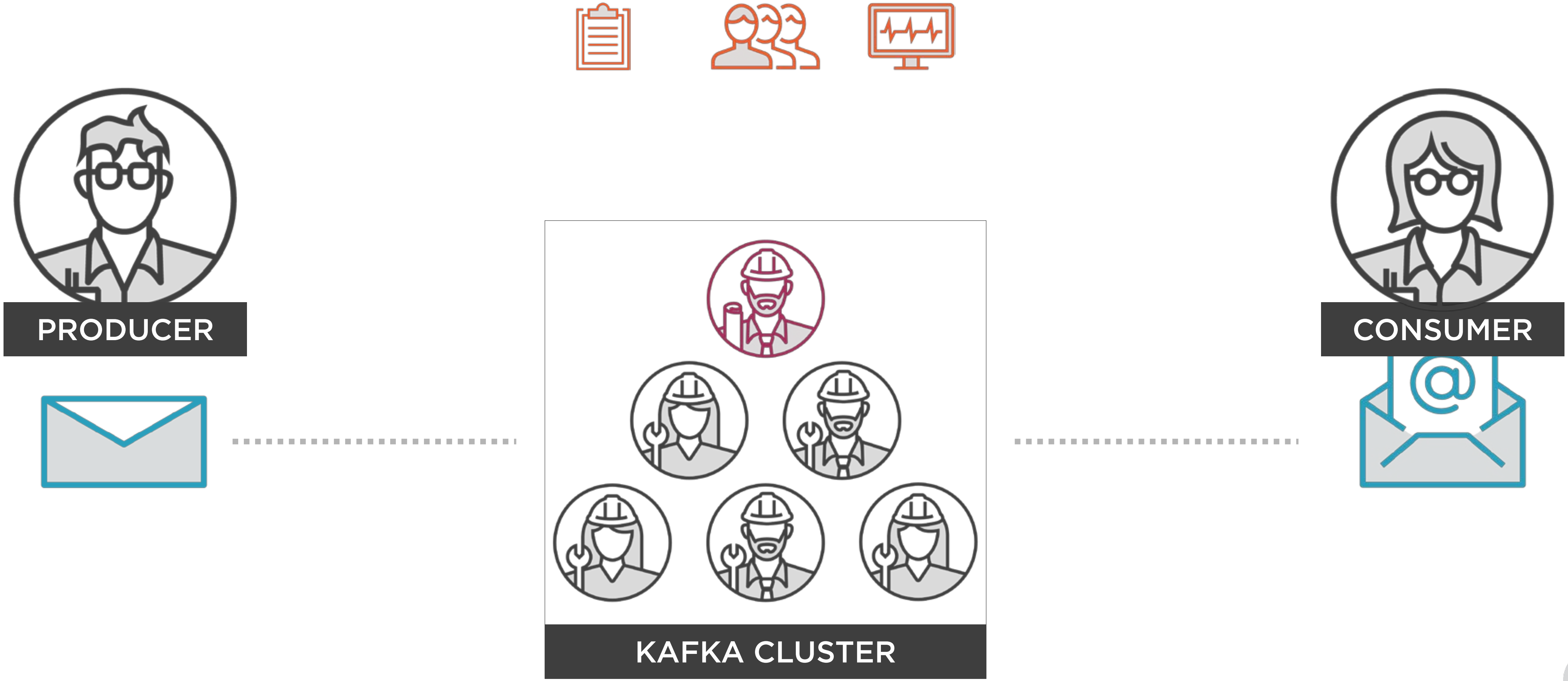
“we cannot afford loss…
three replicas, please”



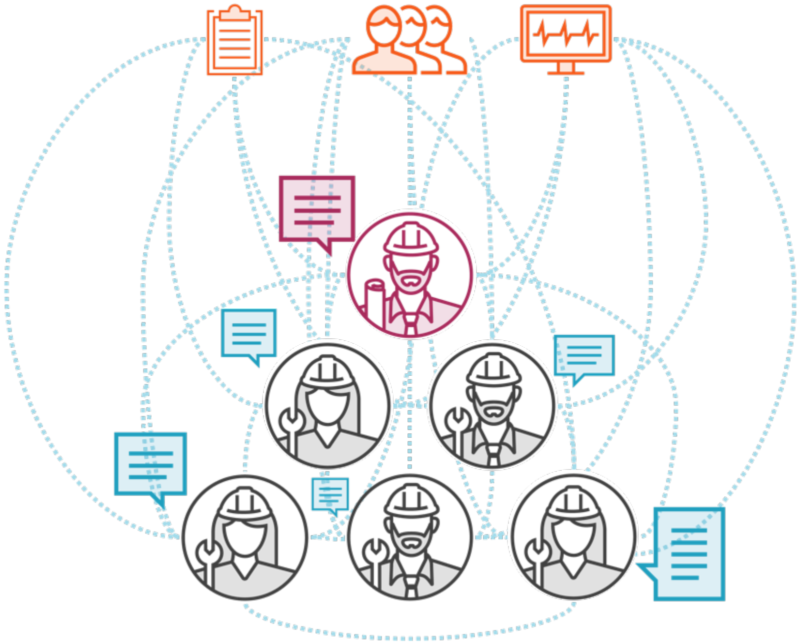
Distributed Systems: Getting Work Done (Reliably)



Sources of Work in Apache Kafka



Distributed Systems: Communication and Consensus



Worker node membership and naming

Configuration management

Leader election

Health status

Apache Zookeeper

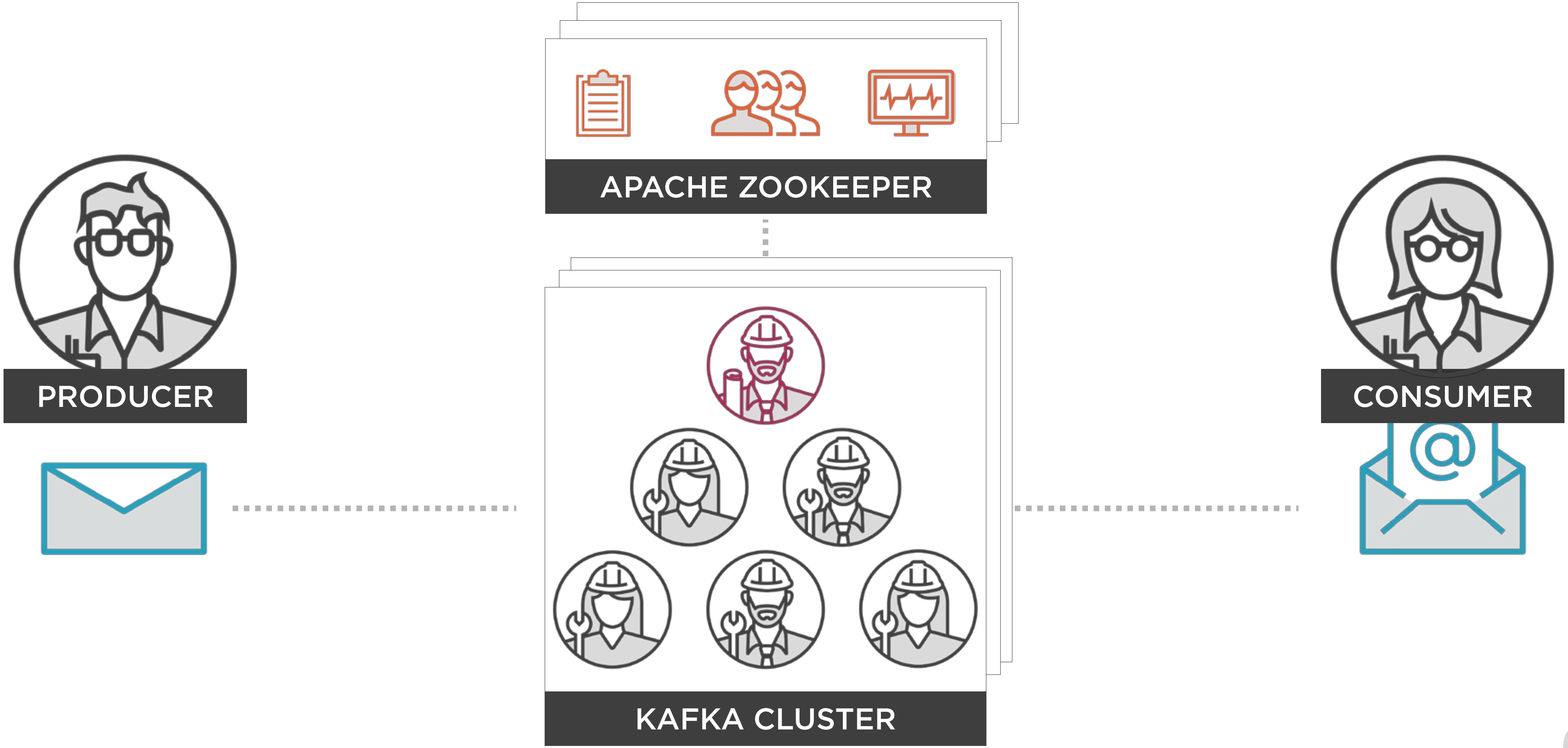
Centralized service for maintaining metadata about a cluster of distributed nodes

- Configuration information
- Heath status
- Group membership

Hadoop, HBase, Mesos, Solr, Redis, and Neo4j

Distributed system consisting of multiple nodes in an “ensemble”

Apache Kafka’s Distributed Architecture



Summary

Apache Kafka is a Pub-Sub messaging system, consisting of:

- Producers and Consumers
- Brokers within a Cluster

Characteristics of distributed systems

- Worker node roles: Controllers, Leaders, and Followers
- Reliability through replication
- Consensus-based communication

Role of Apache Zookeeper