kafka

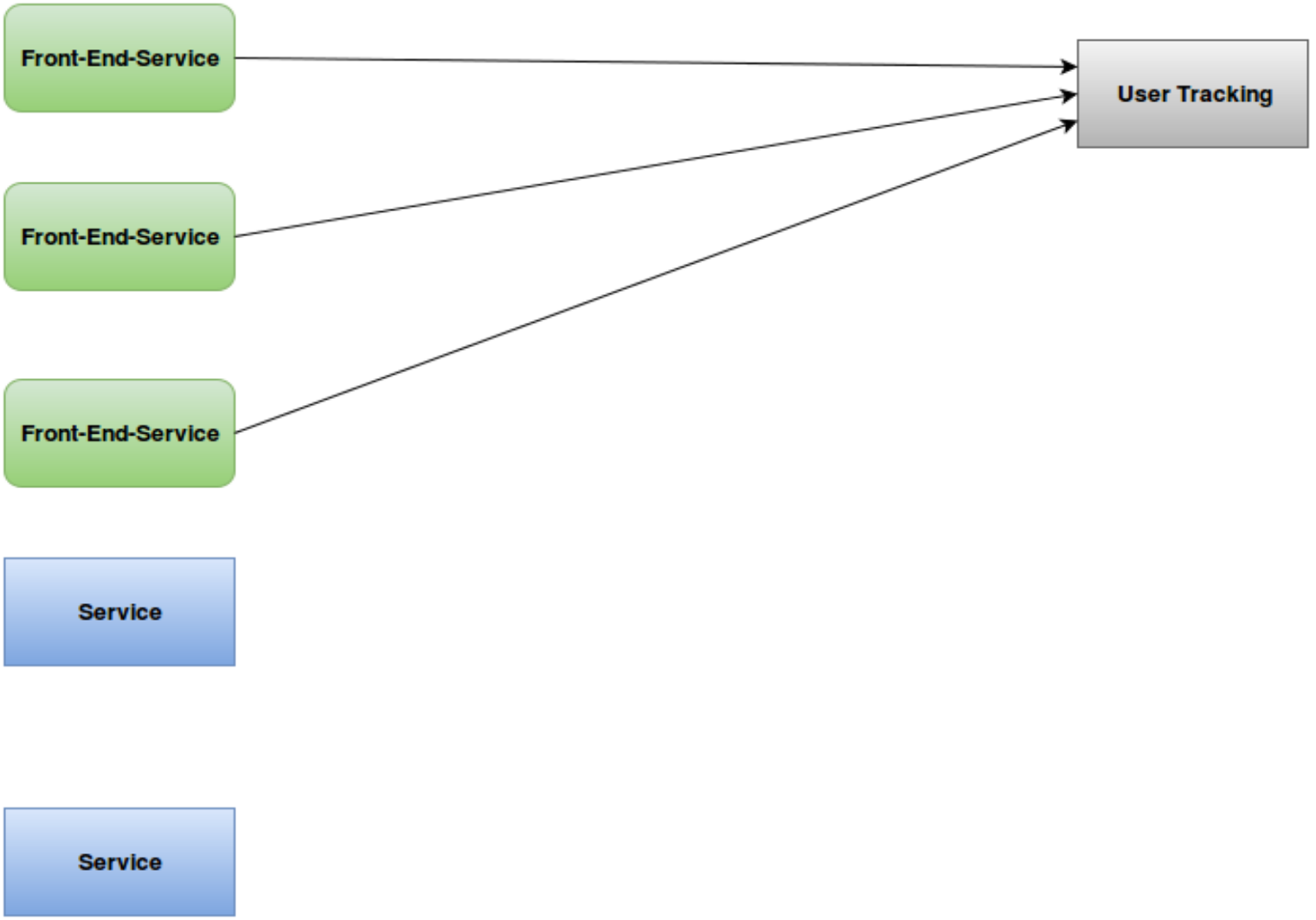
Topics Covered

- What is Kafka
- Why Kafka
- High level overview
- Use cases
- Key terminology
- Partitions distribution over brokers
- Replication protocol

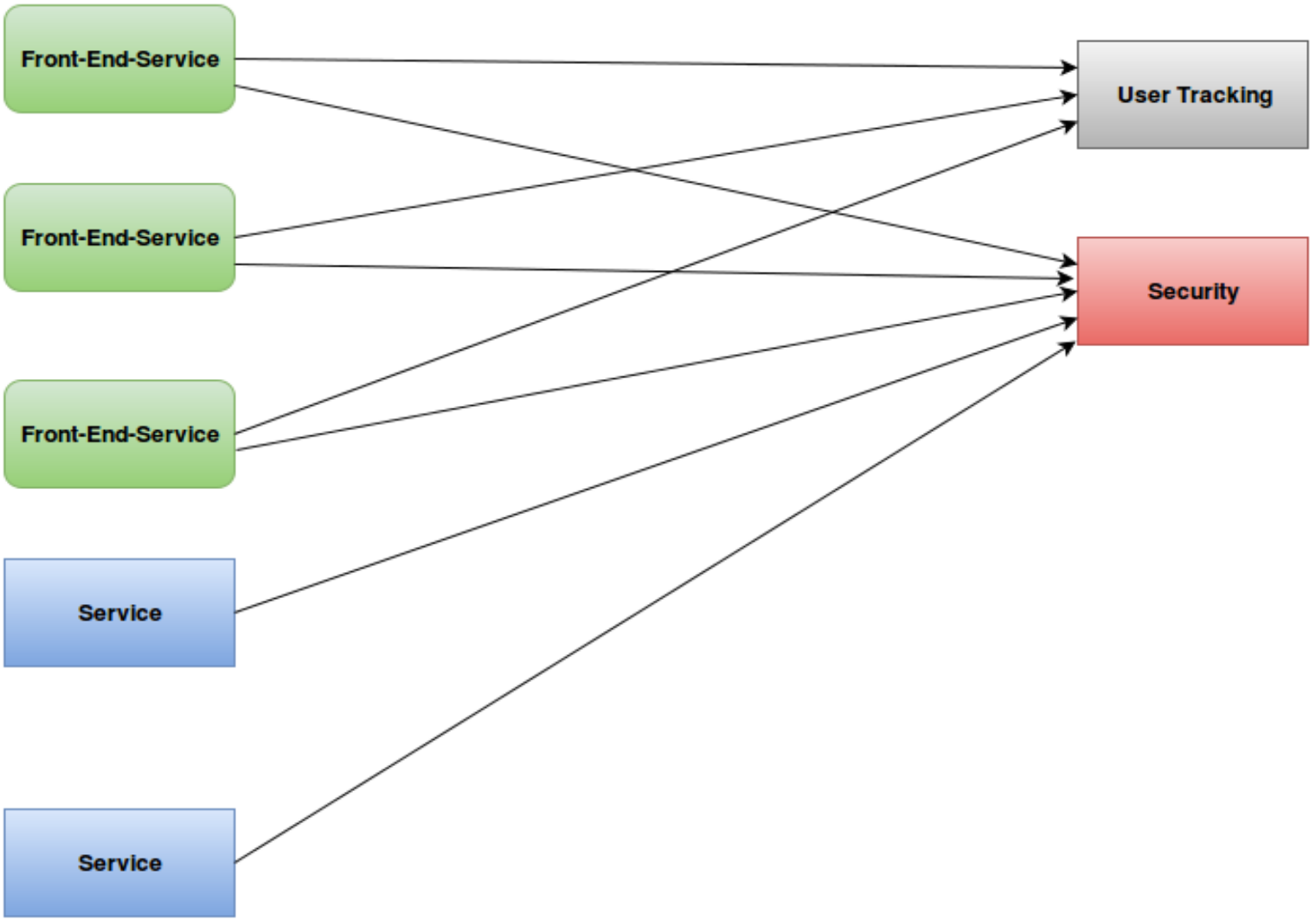
• What is Kafka

- publish-subscribe messaging system
- fast
- distributed by Design
- fault tolerant
- scalable
- durable
- written in **Scala**
- free and open source

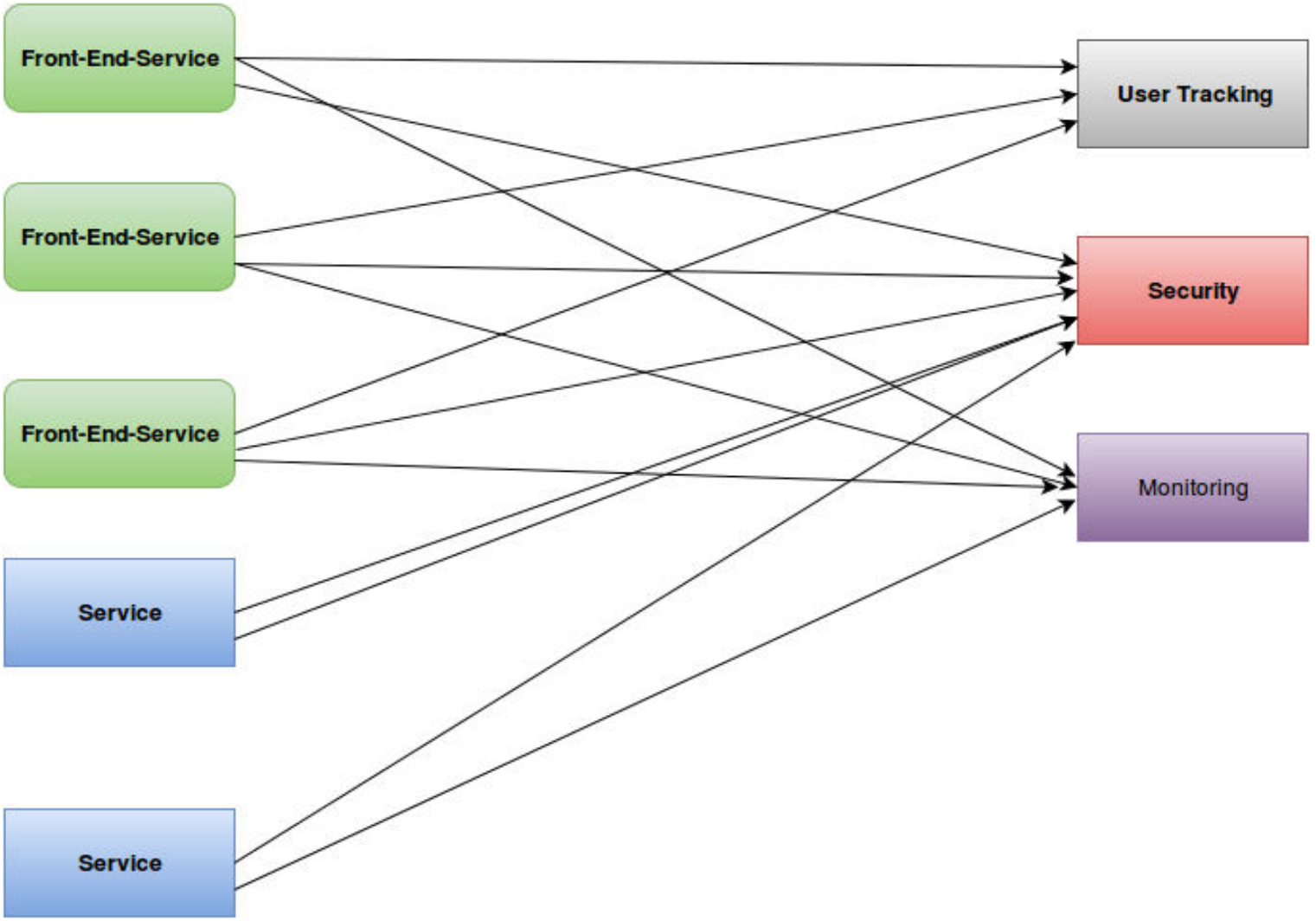
- Building Data Pipelines



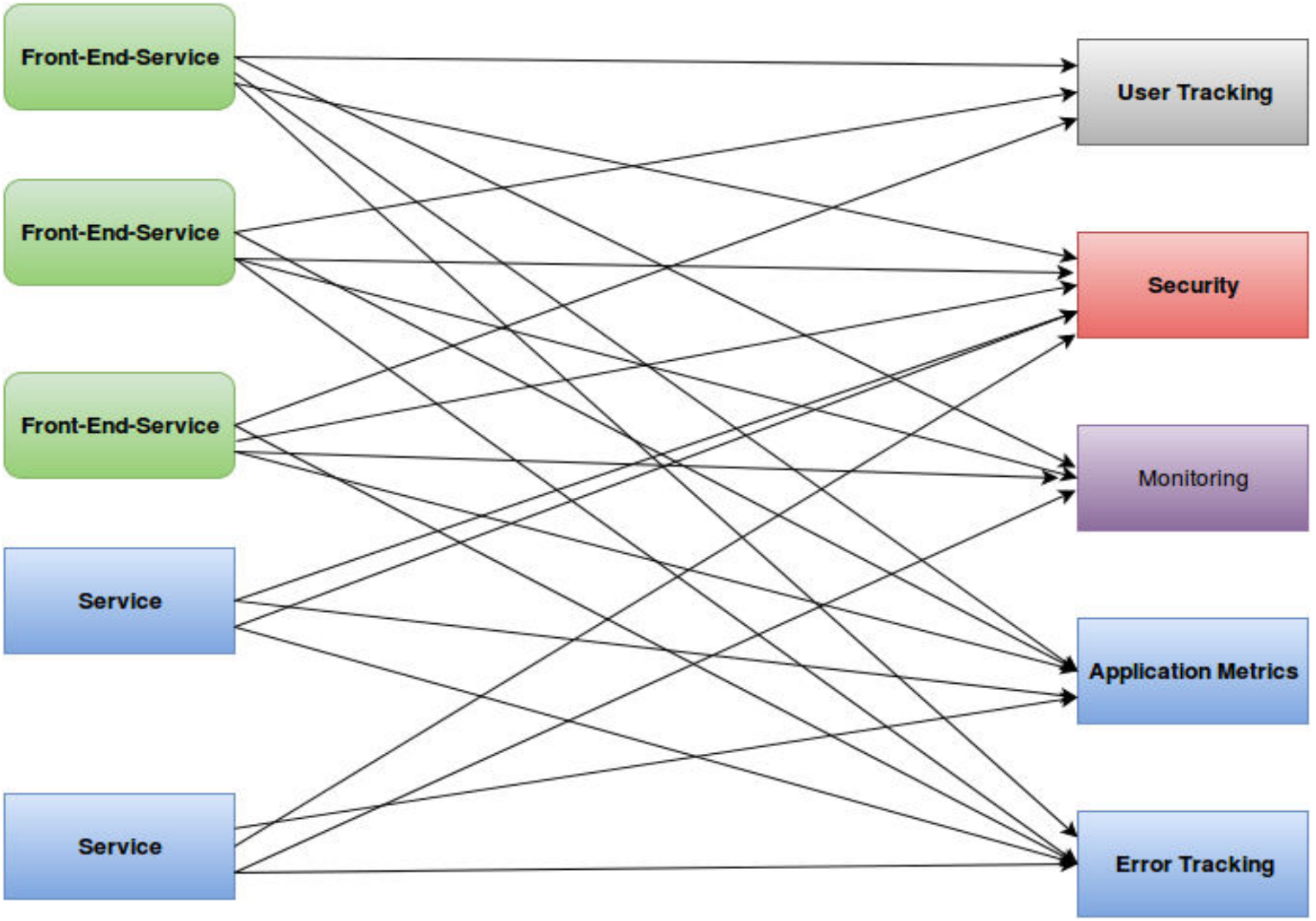
• Building Data Pipelines



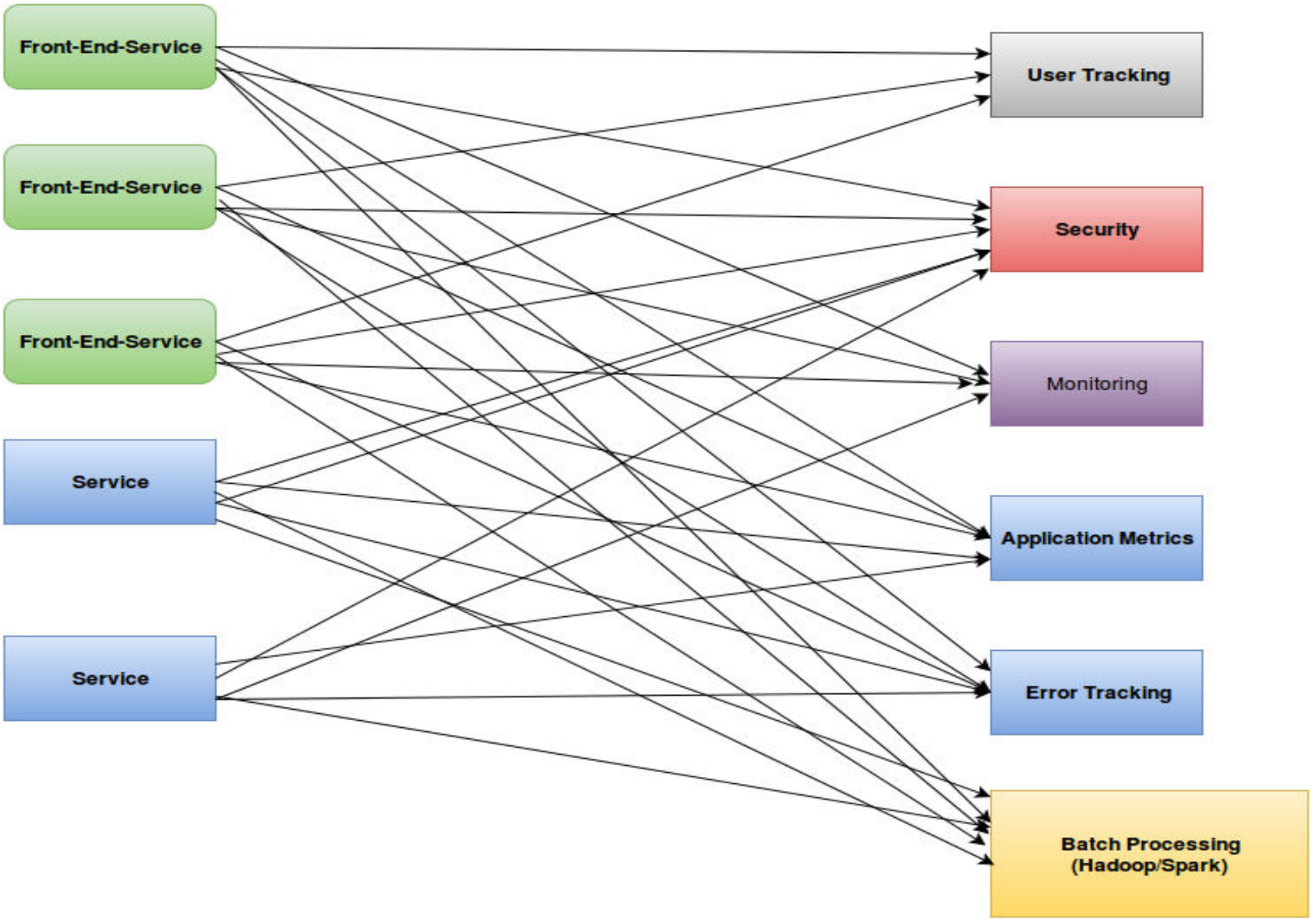
• Building Data Pipelines



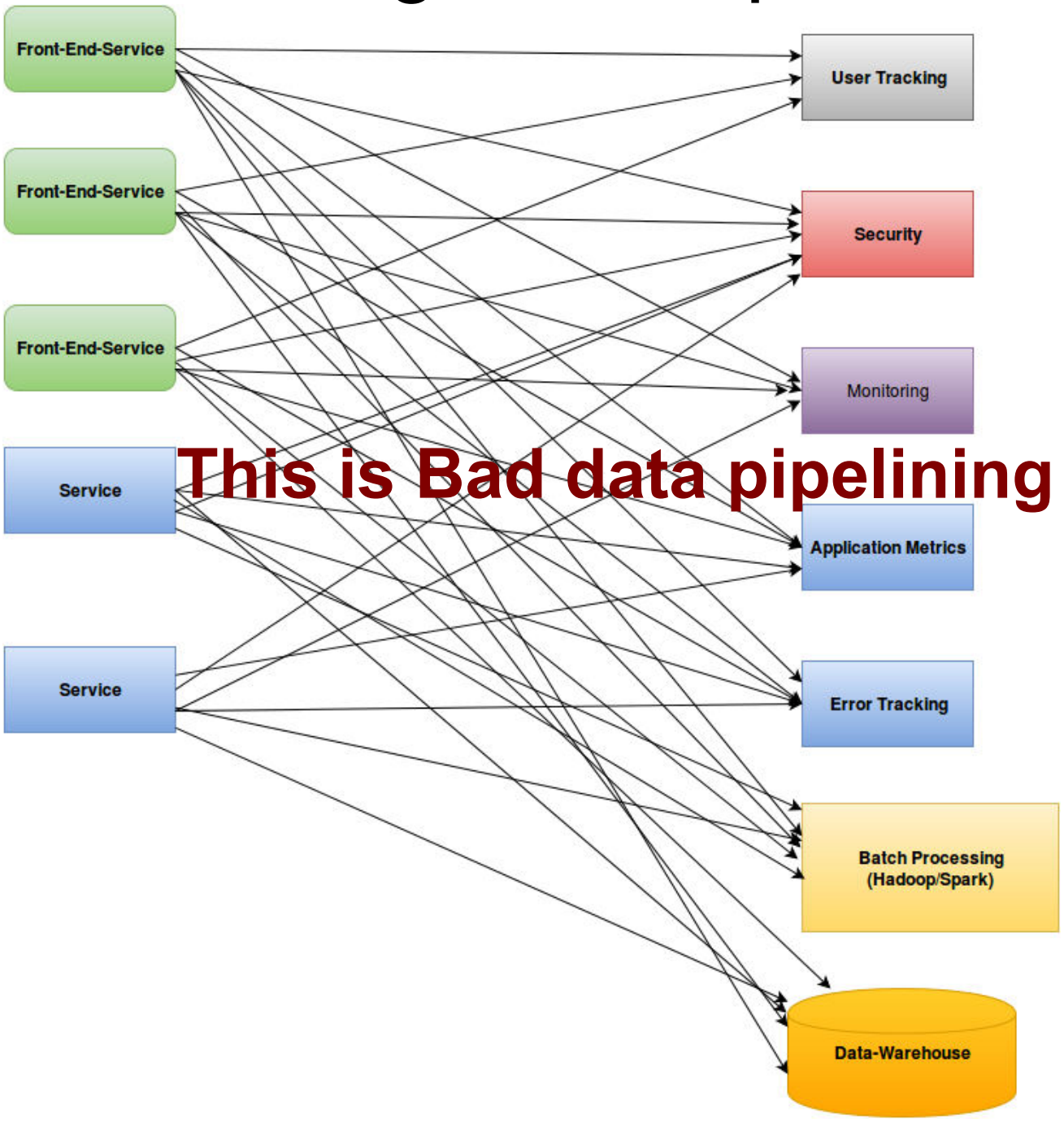
• Building Data Pipelines



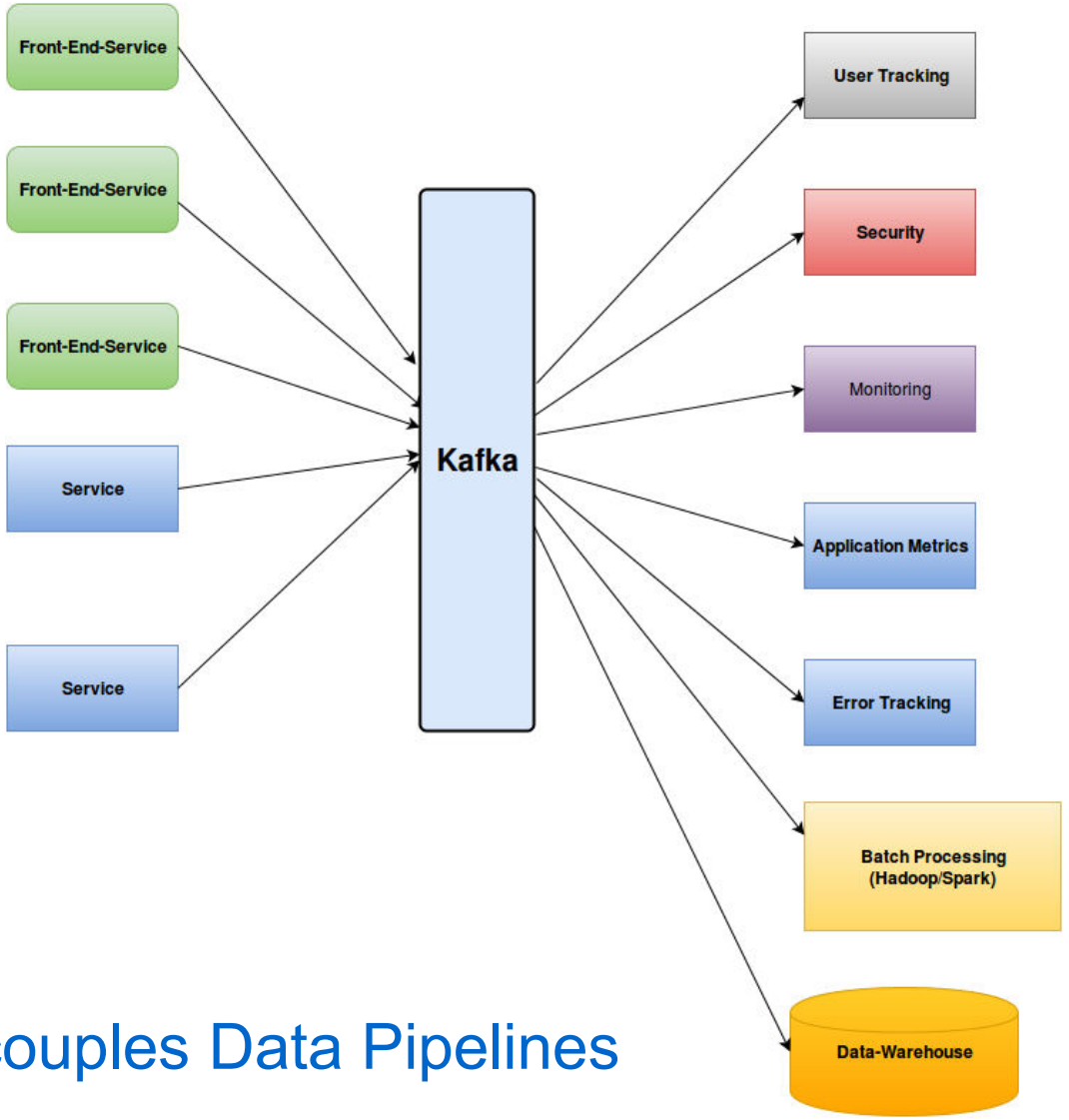
• Building Data Pipelines



• Building Data Pipelines

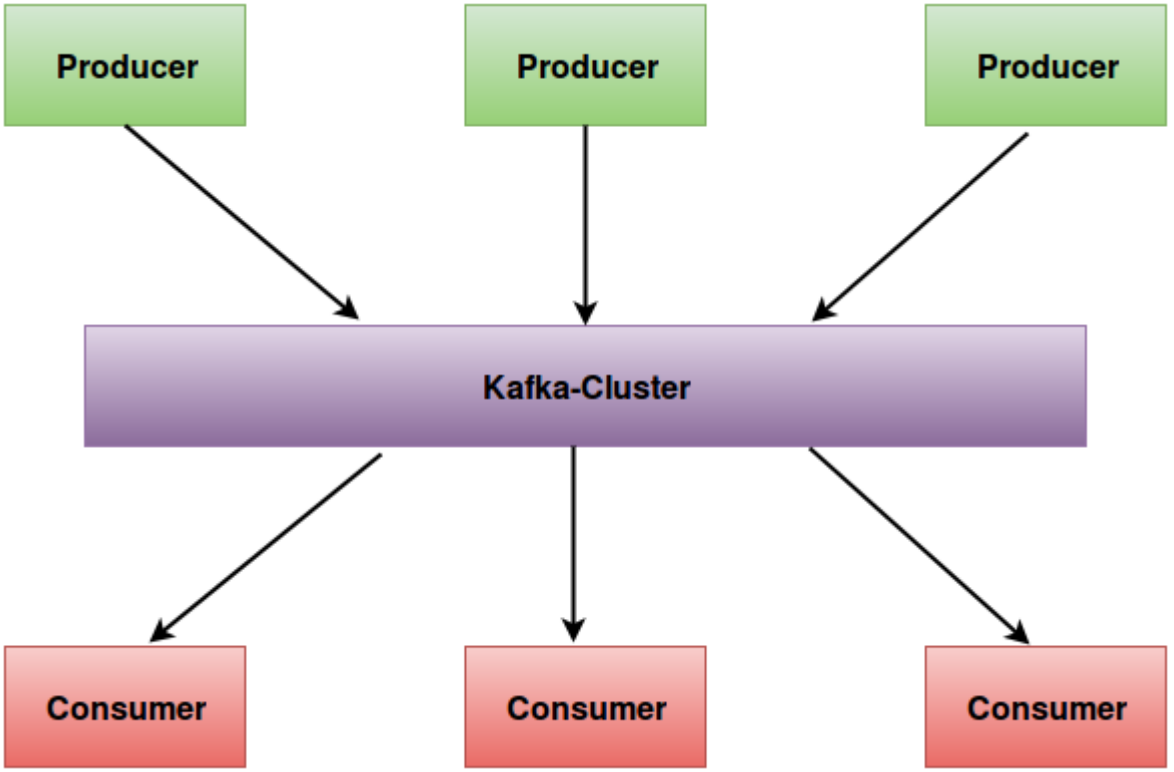


• Building Data Pipelines

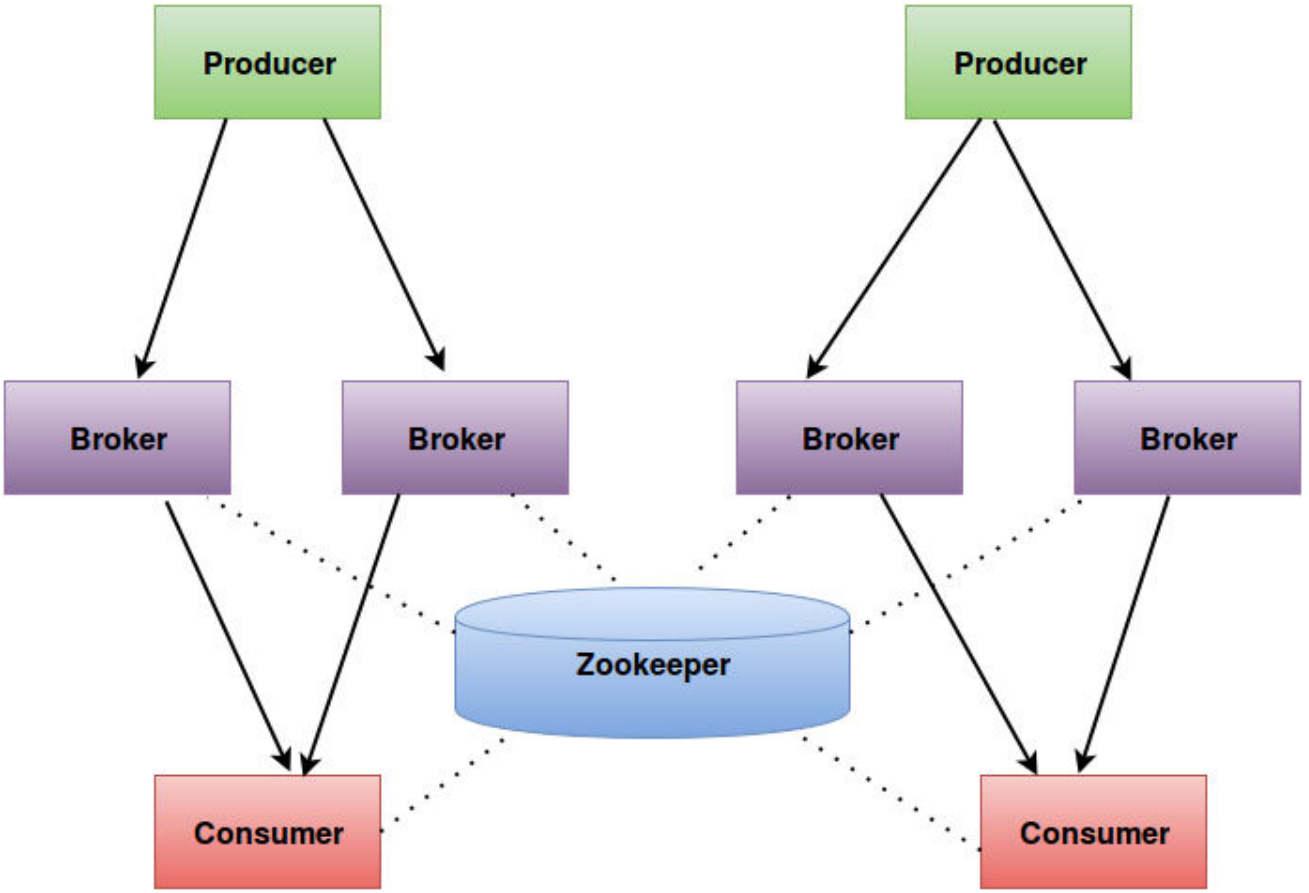


- Kafka decouples Data Pipelines

- High level overview



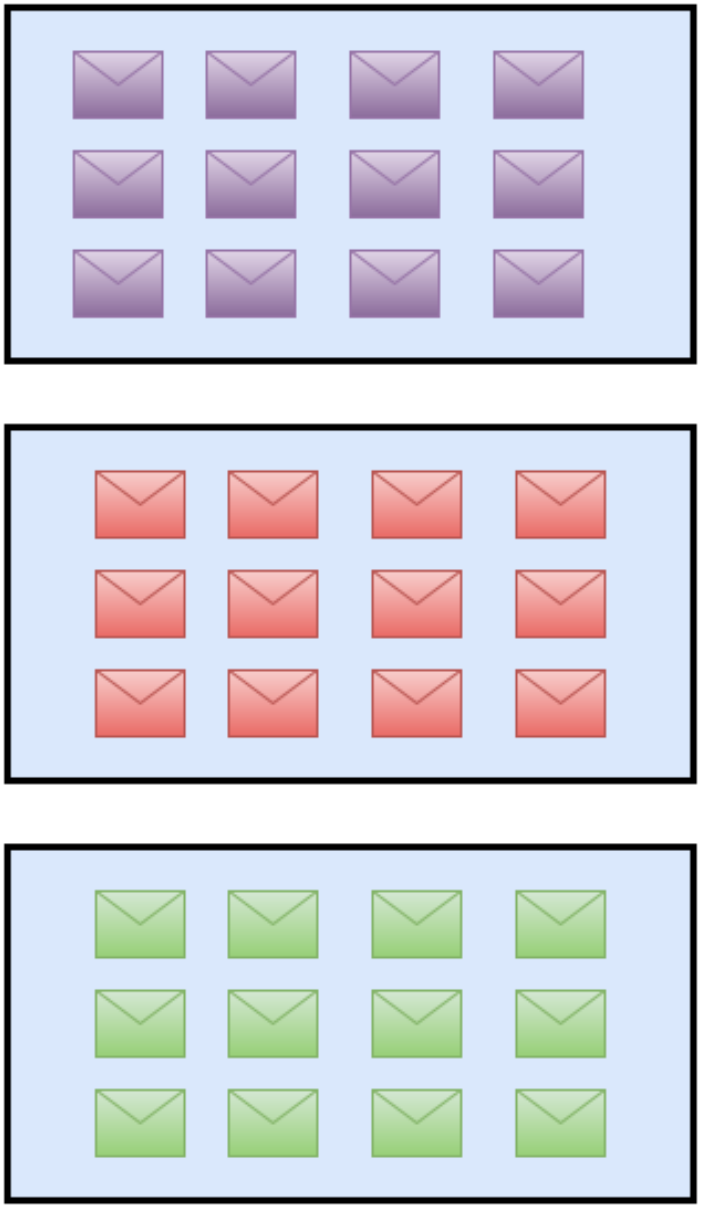
- High level overview

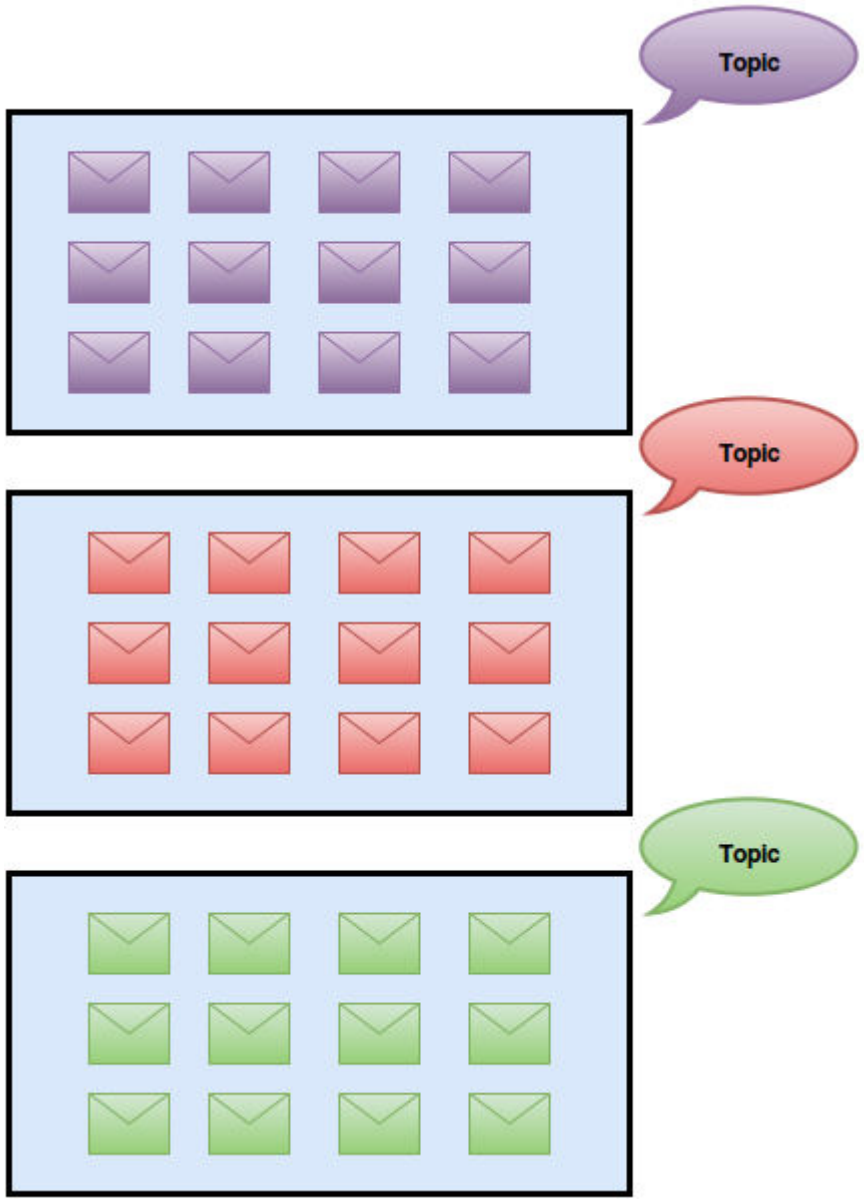


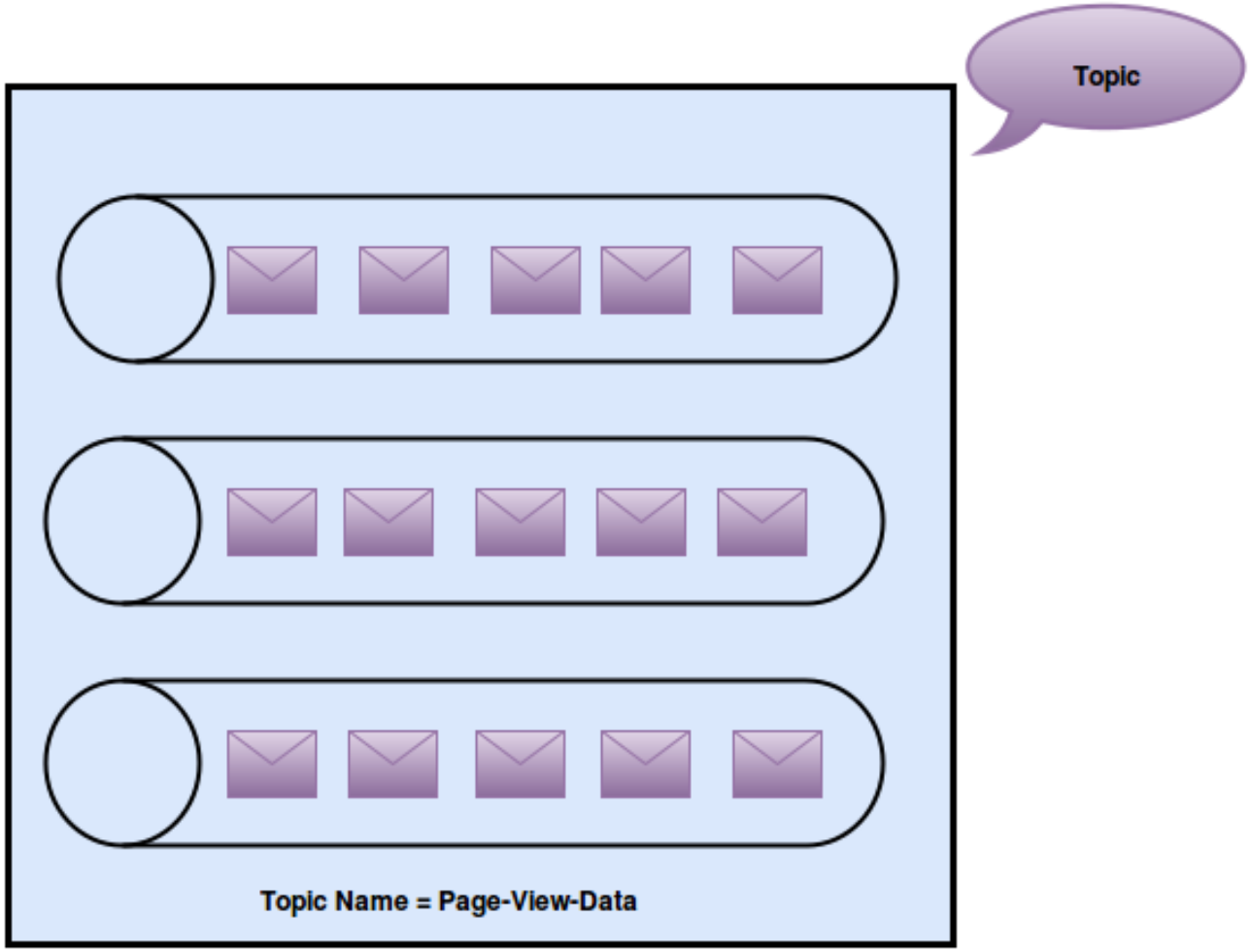
Use cases

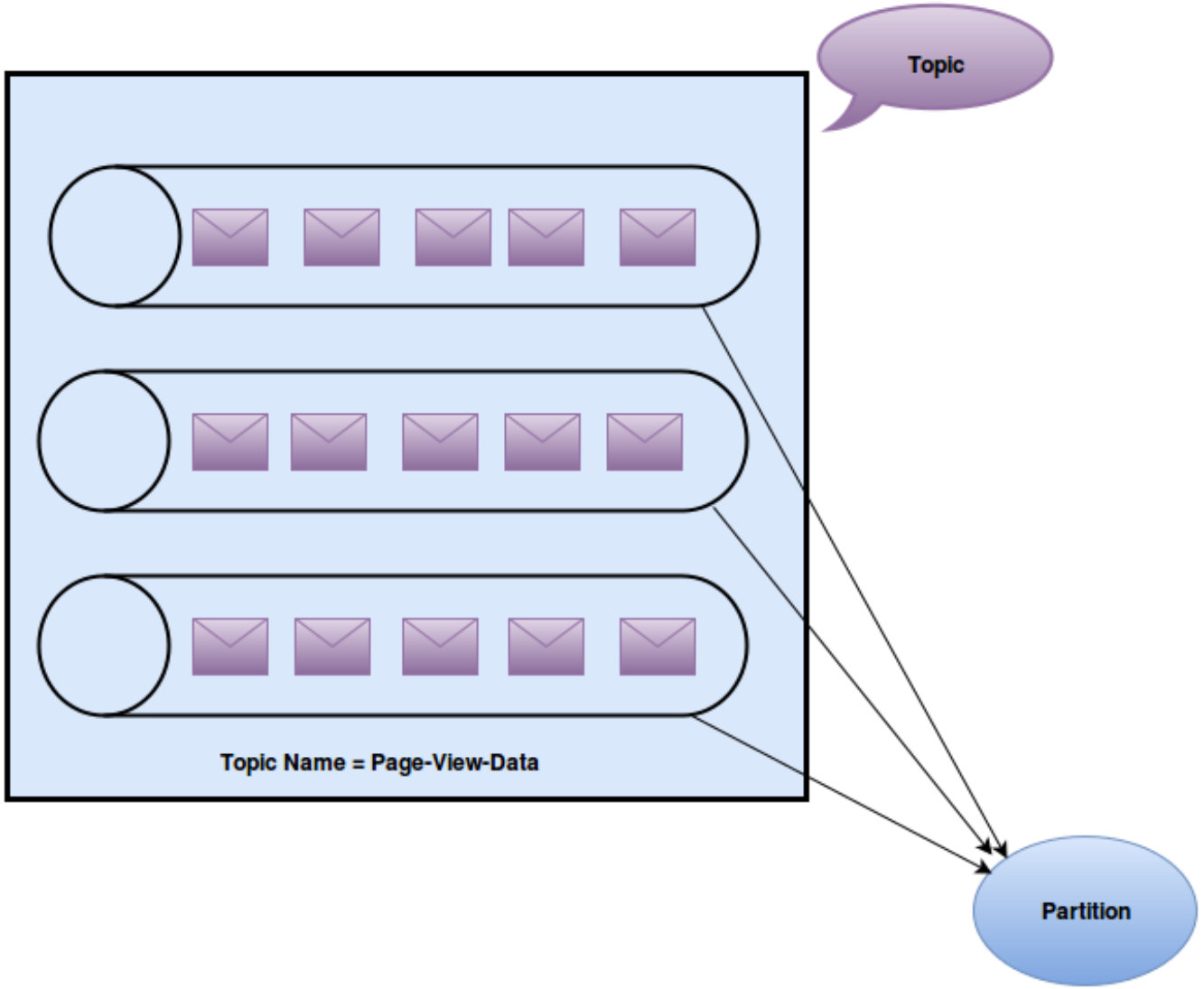
- Messaging
- Website Activity Tracking
- Metrics
- Log Aggregation
- **Real-Time Stream Processing**
- Event Sourcing
- Commit Log
- Internet Of Things (IoT)

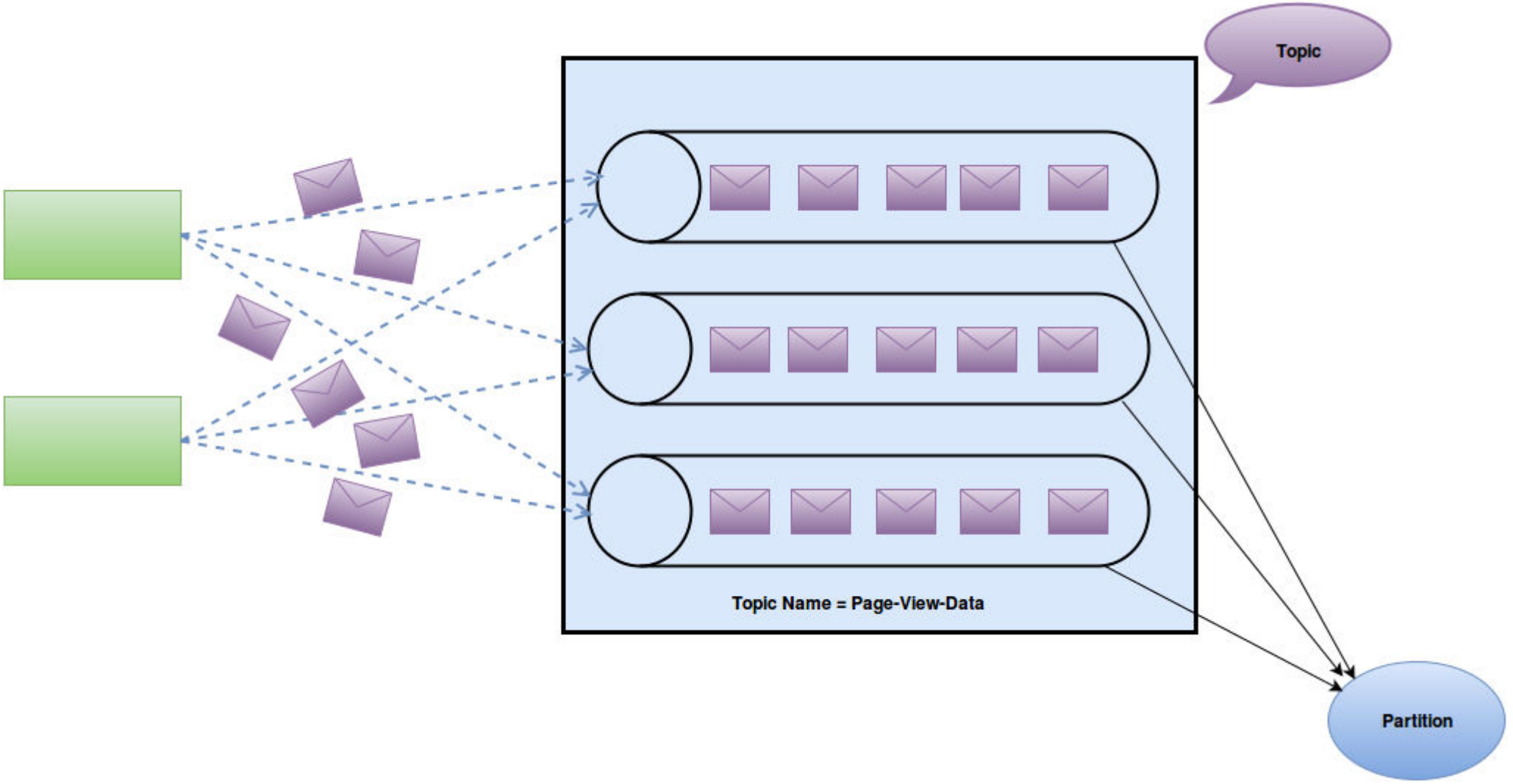
- Key Terminology

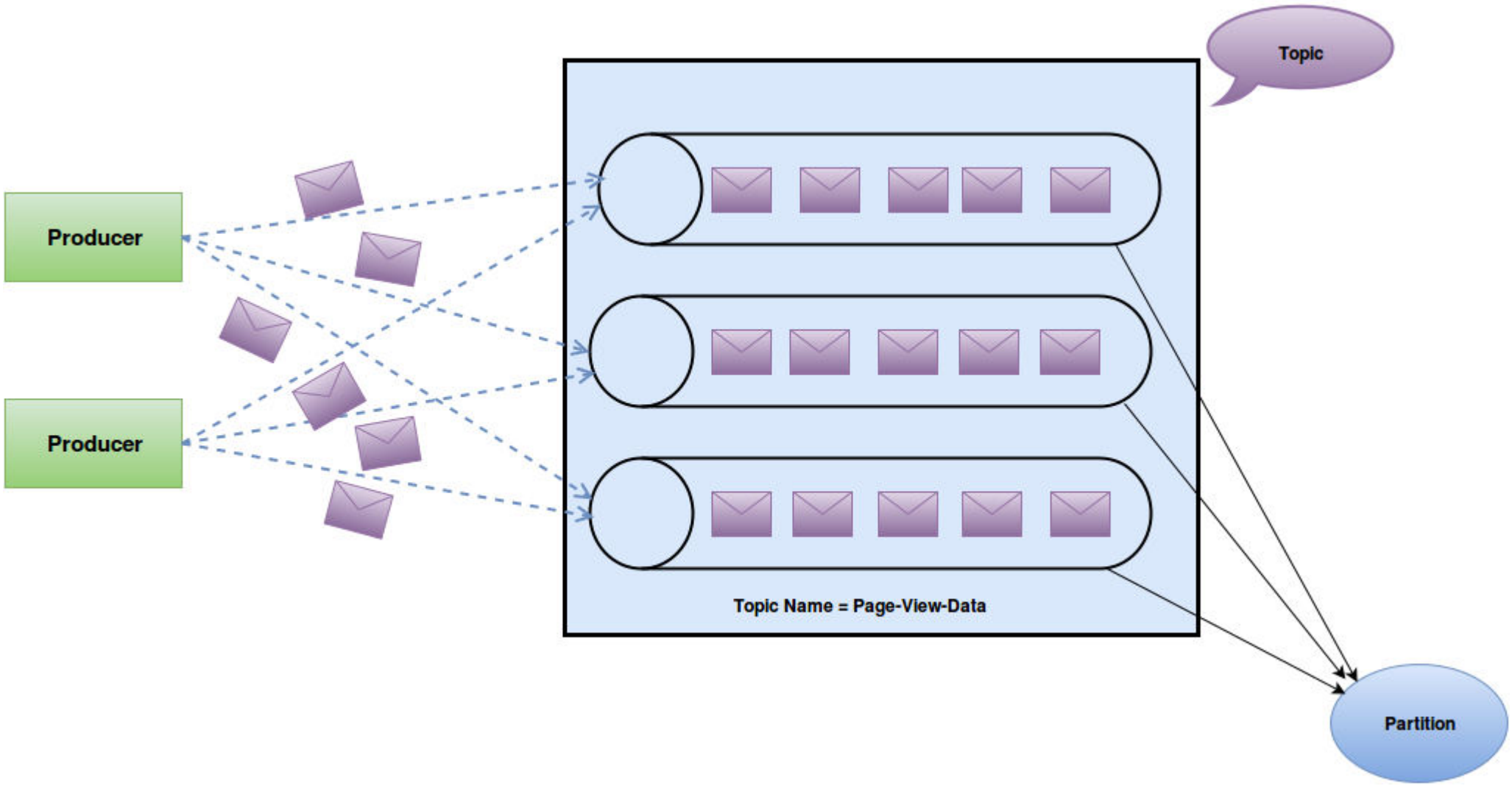


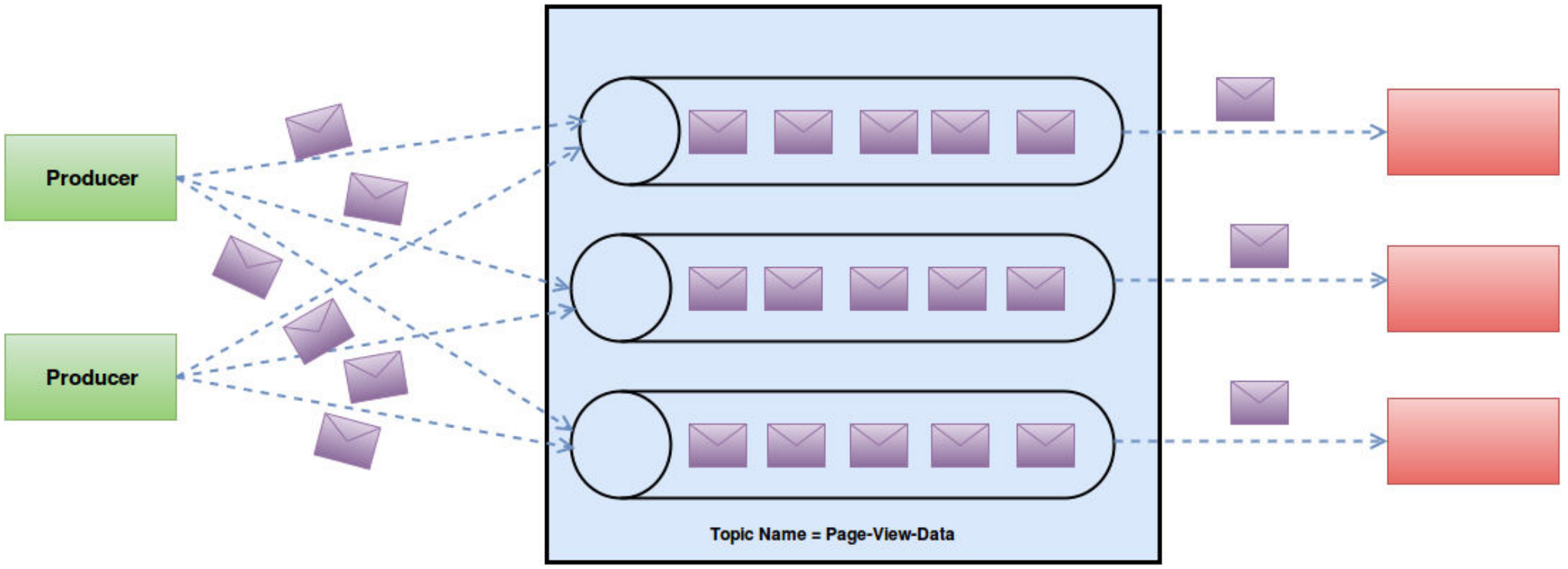

Topic
Topic
Topic

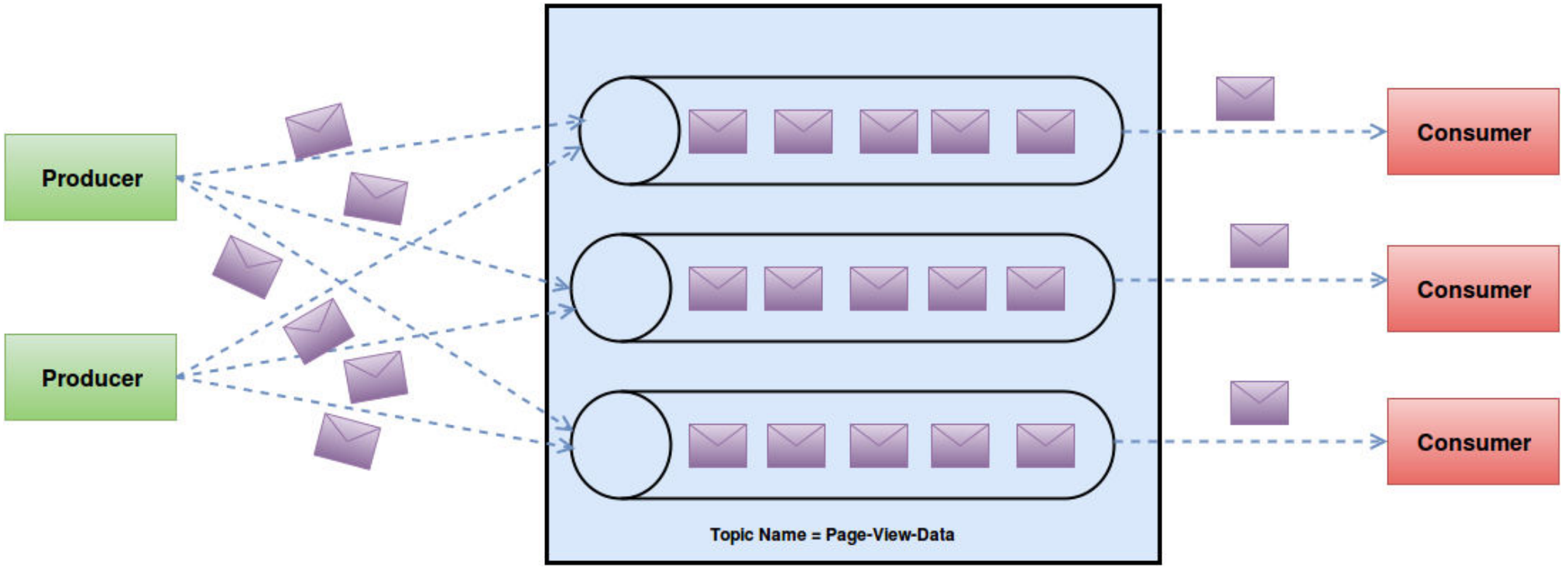

Topic
Topic Name = Page-View-Data

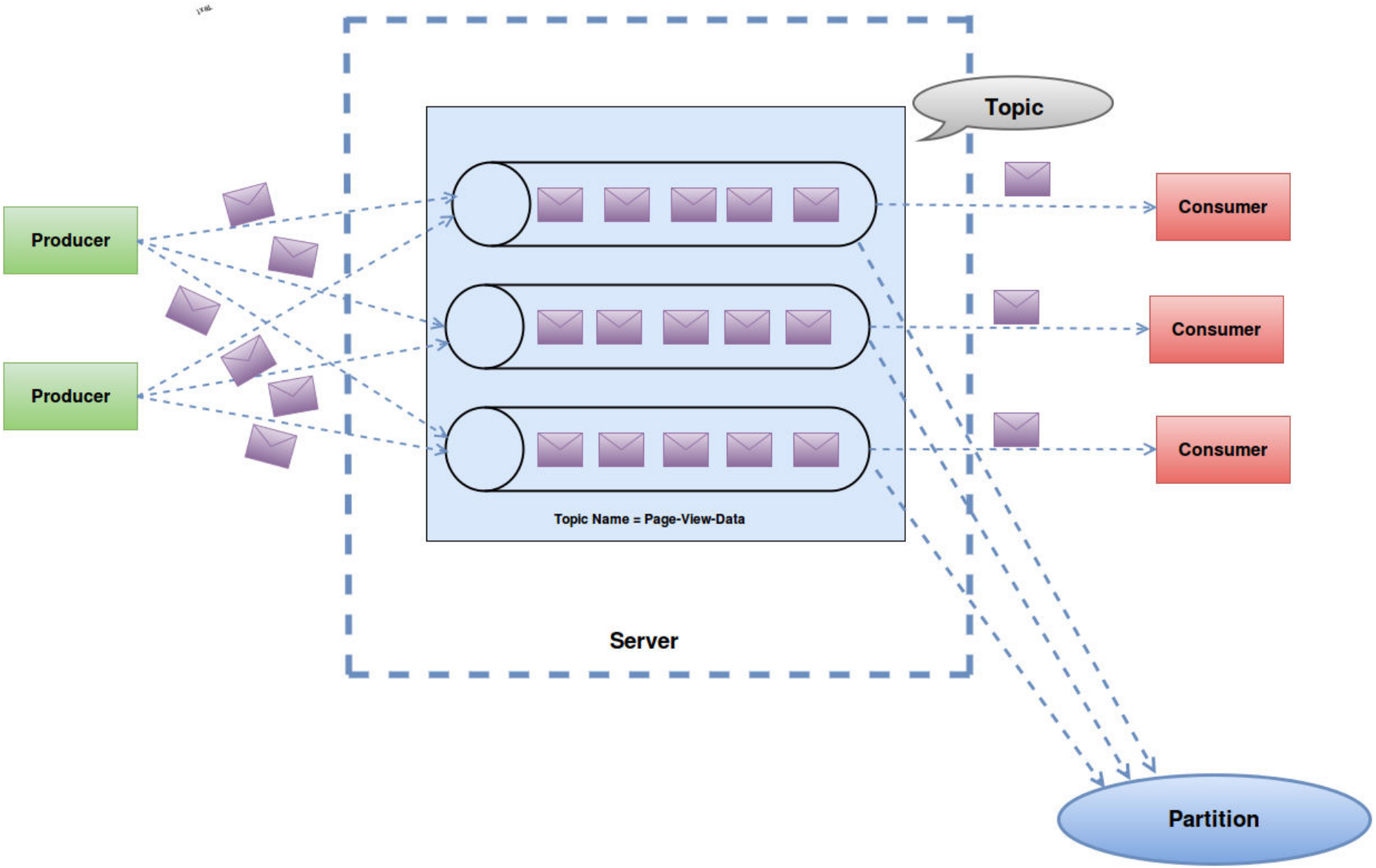

Topic
Topic Name = Page-View-Data
Partition

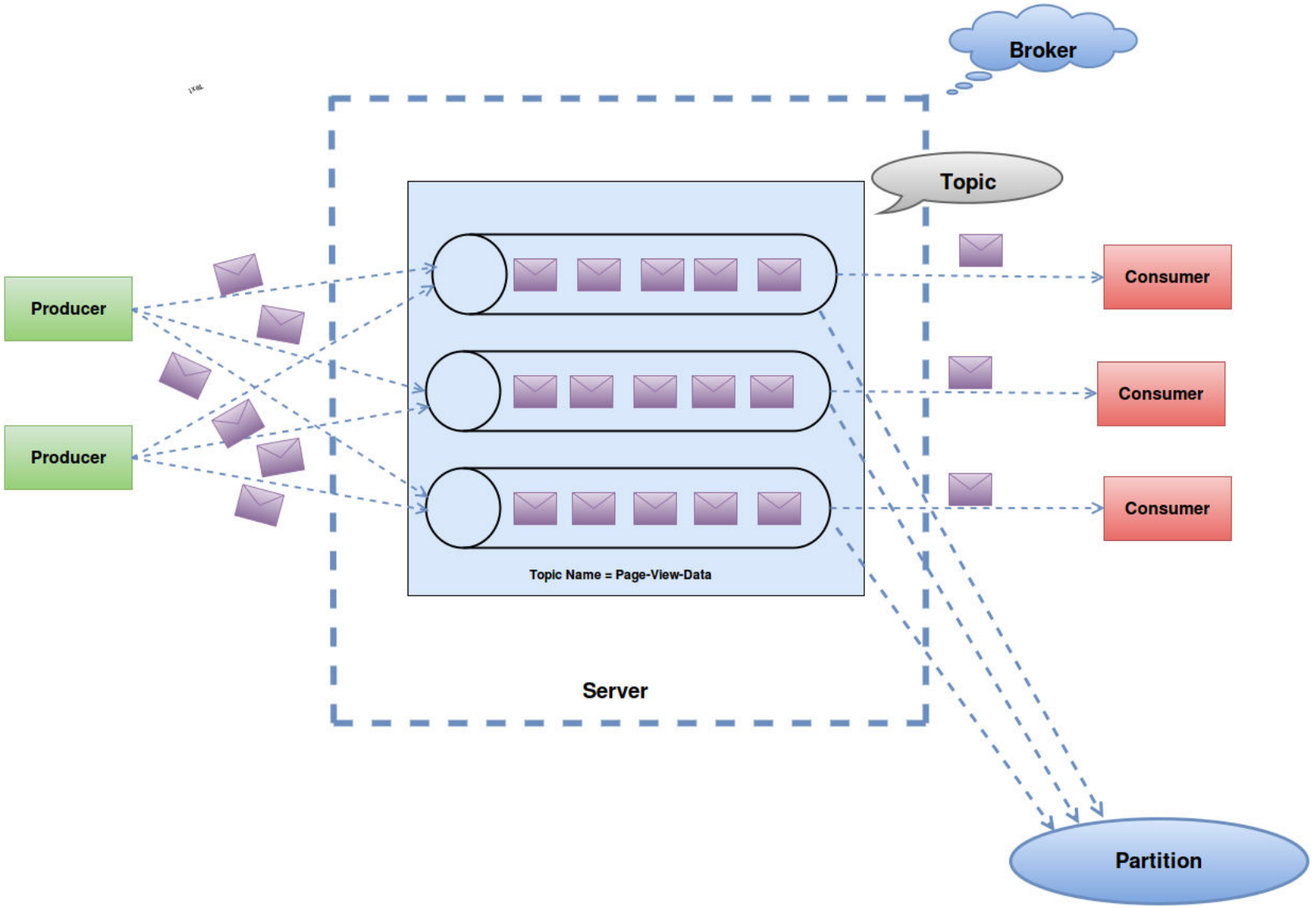

Topic
Topic Name = Page-View-Data
Partition


Topic
Producer
Producer
Topic Name = Page-View-Data
Partition


Producer
Producer
Topic Name = Page-View-Data

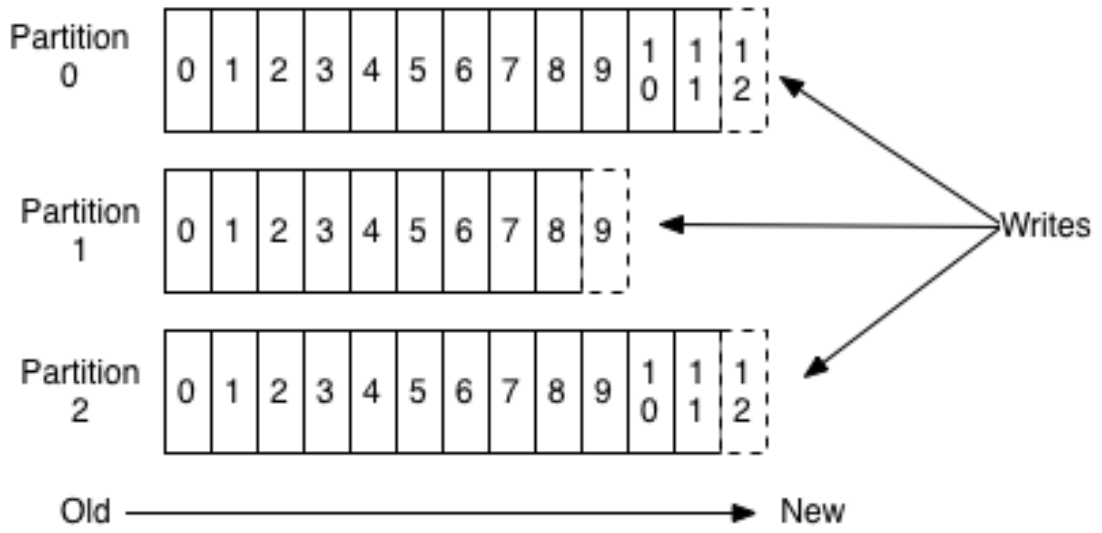

Producer
Producer
Topic Name = Page-View-Data
Consumer
Consumer
Consumer


Topic
Producer
Producer
Consumer
Consumer
Consumer
Topic Name = Page-View-Data
Server
Partition


Broker
Topic
Producer
Producer
Consumer
Consumer
Consumer
Topic Name = Page-View-Data
Server
Partition

Anatomy of a Topic

For each topic, the Kafka cluster maintains a partitioned log that looks like this:

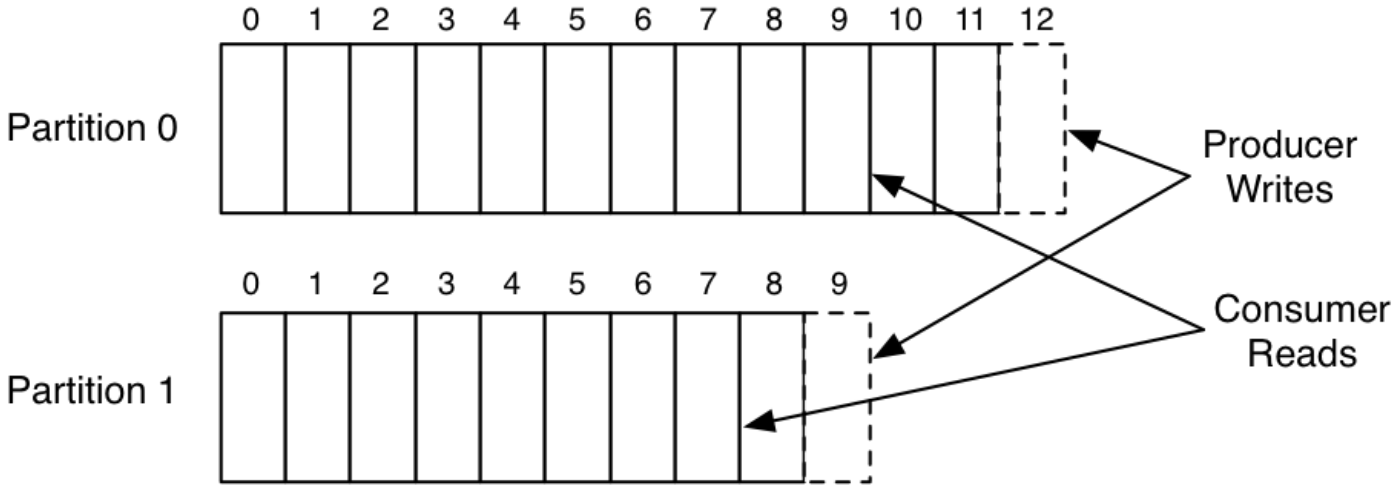


- Number of partition for a Topic is configurable. In this example number of partition are three.

http://kafka.apache.org/images/log_anatomy.png

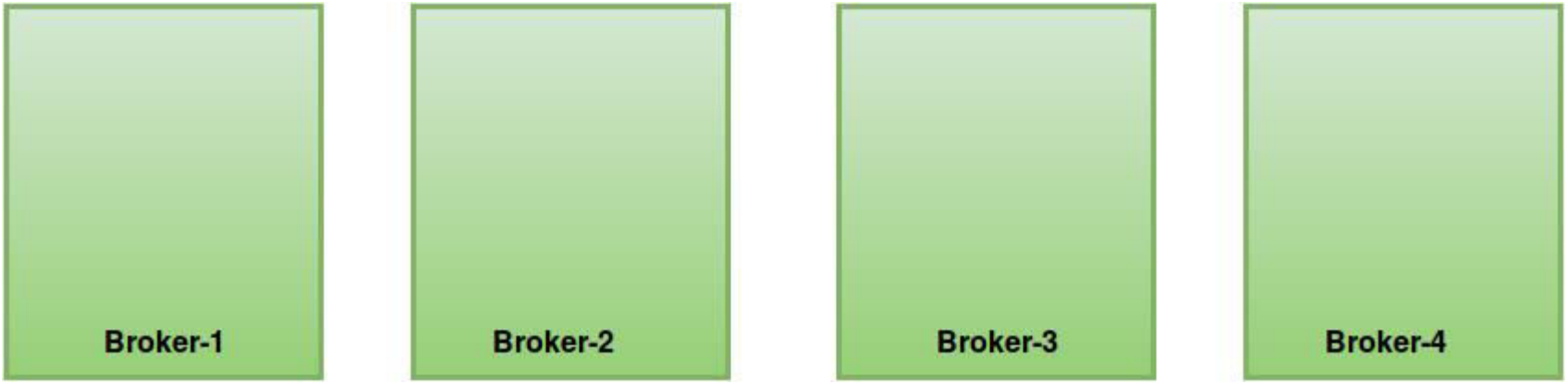
Reading & Writing From Topic

Topic with two partition:

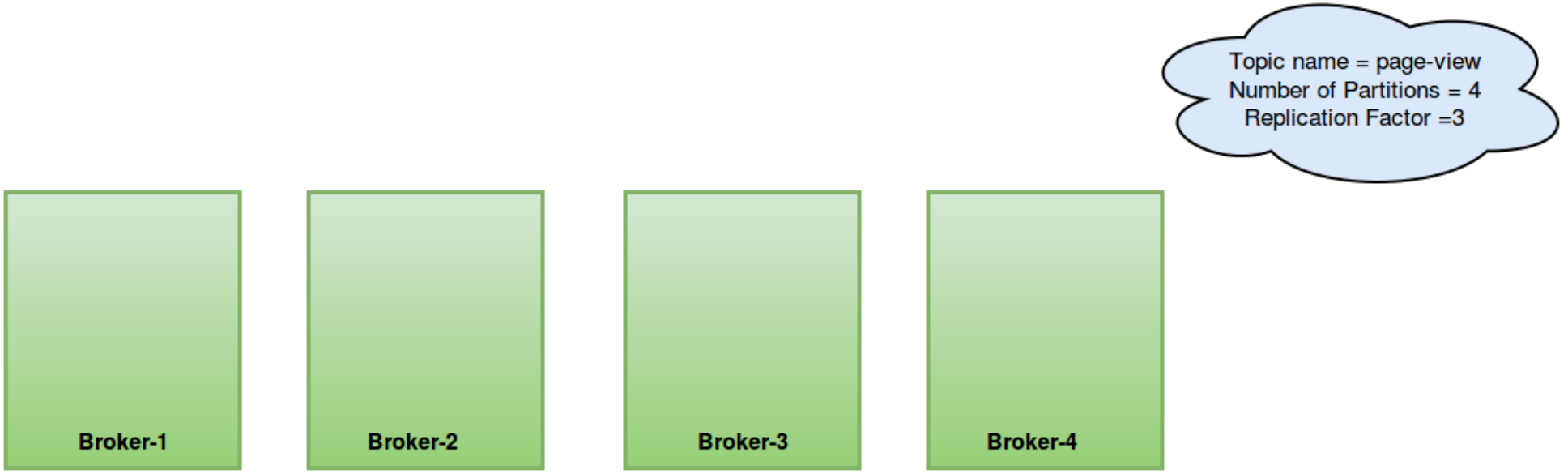


https://content.linkedin.com/content/dam/engineering/en-us/blog/migrated/partitioned_log_0.png

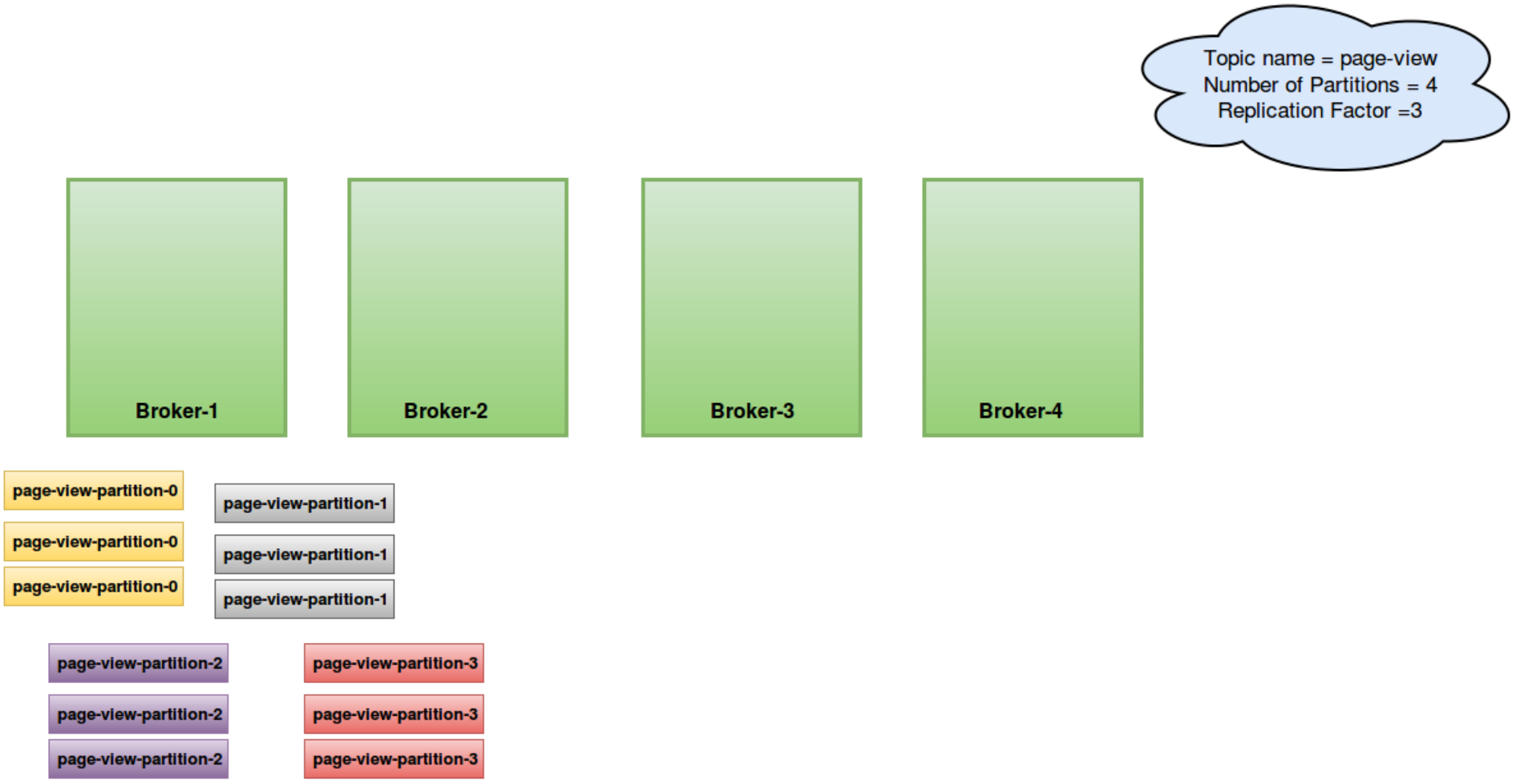
Partitions distribution



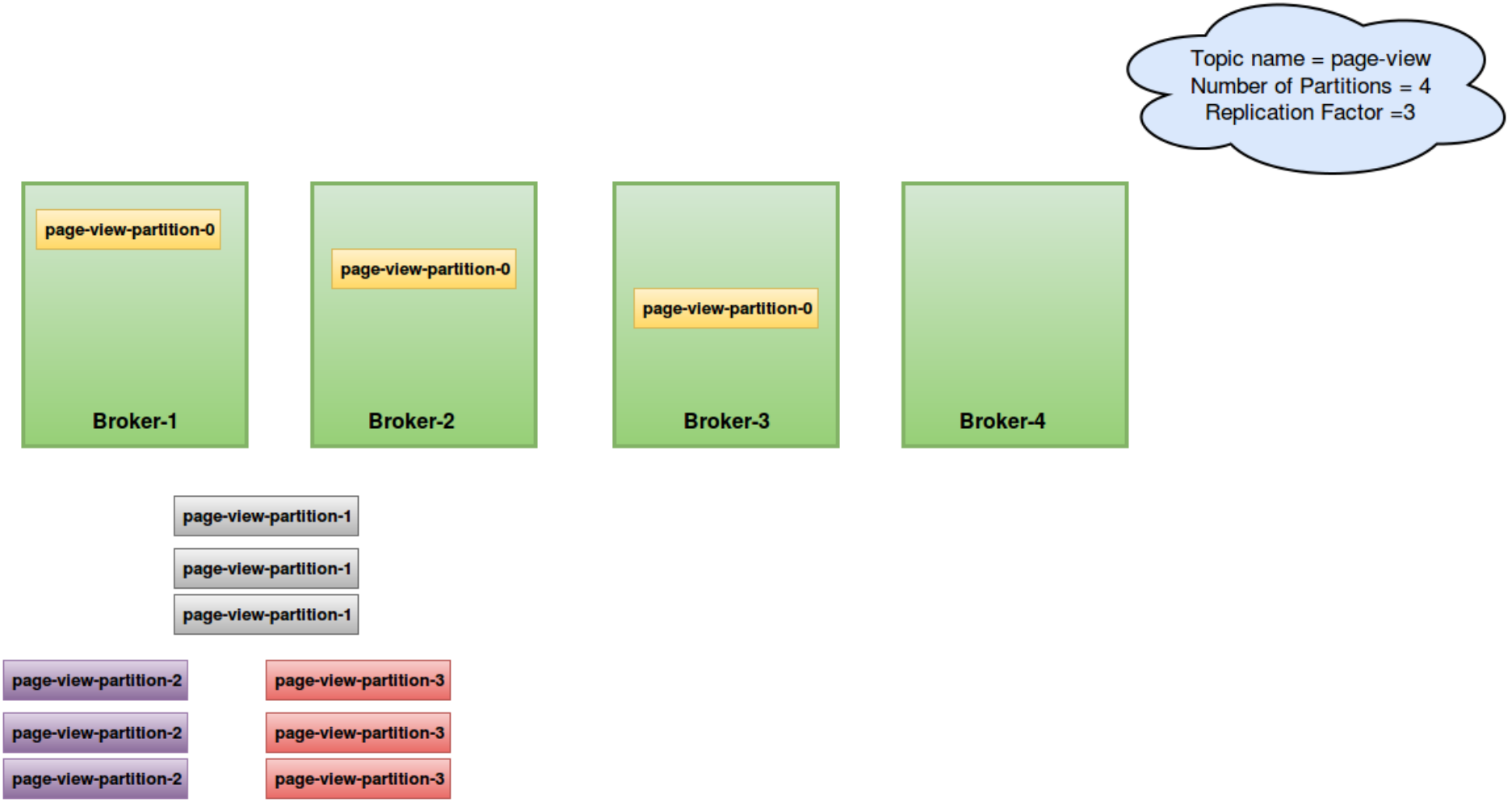
Partitions distribution



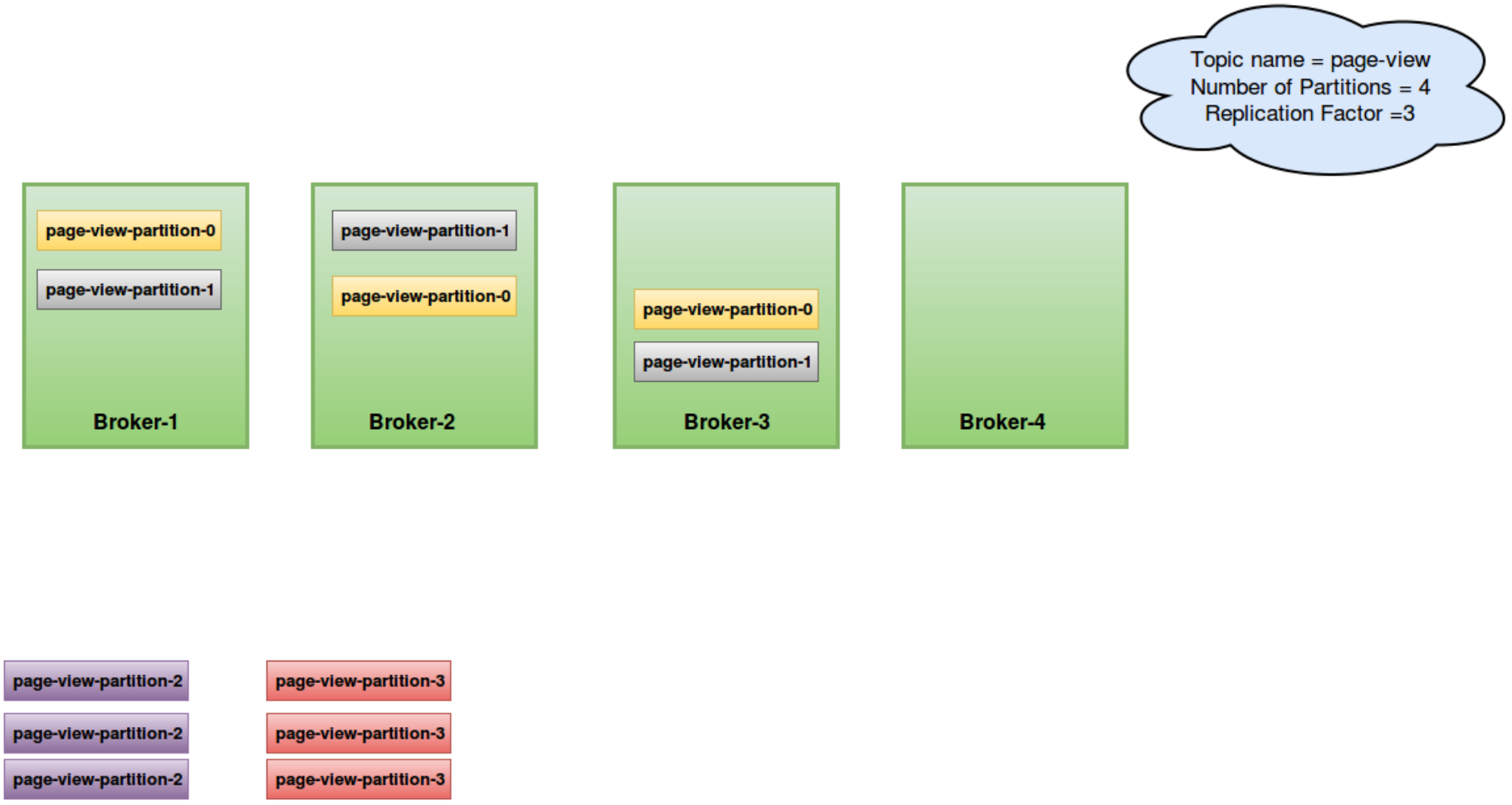
Partitions distribution



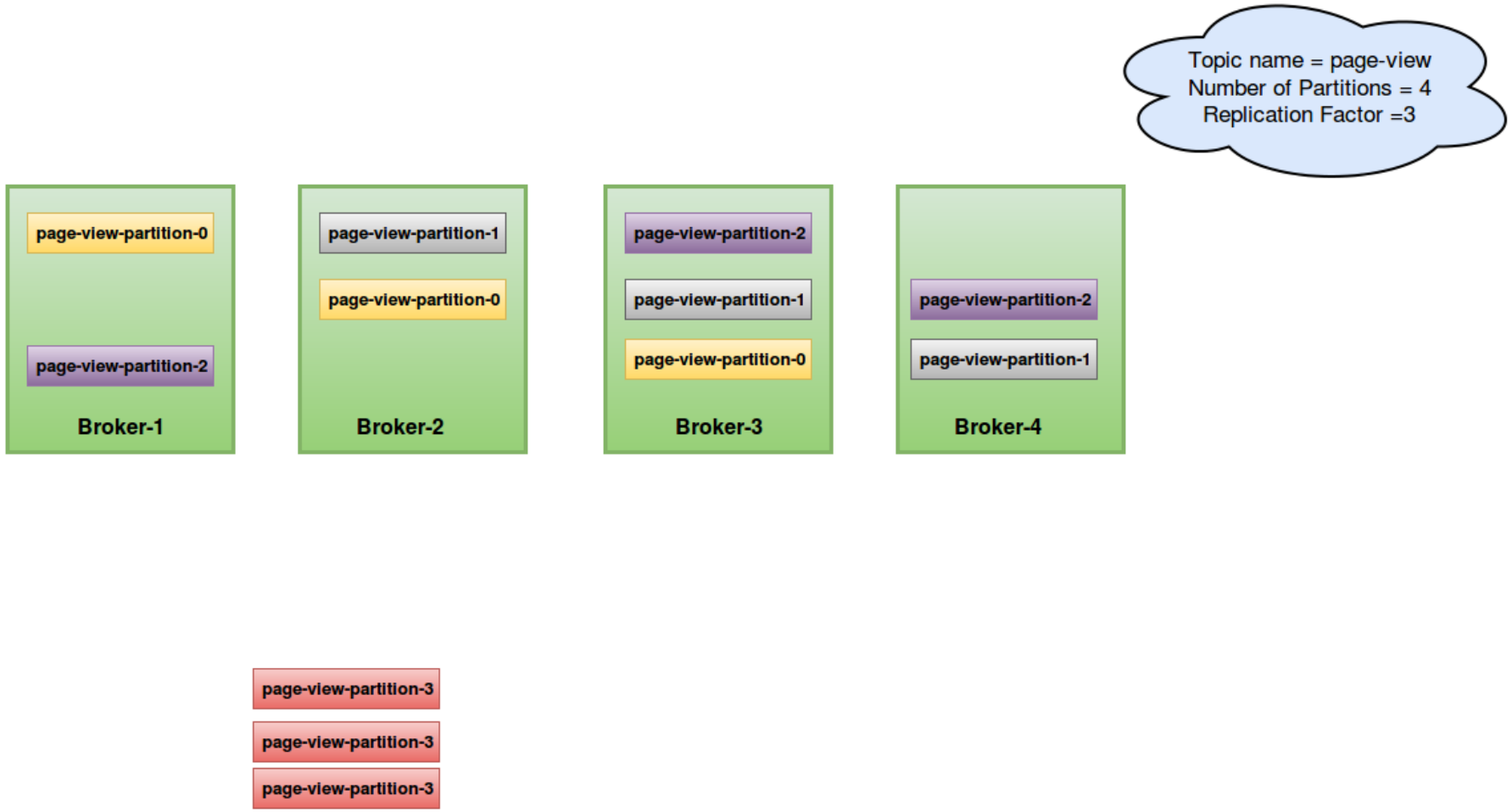
Partitions distribution



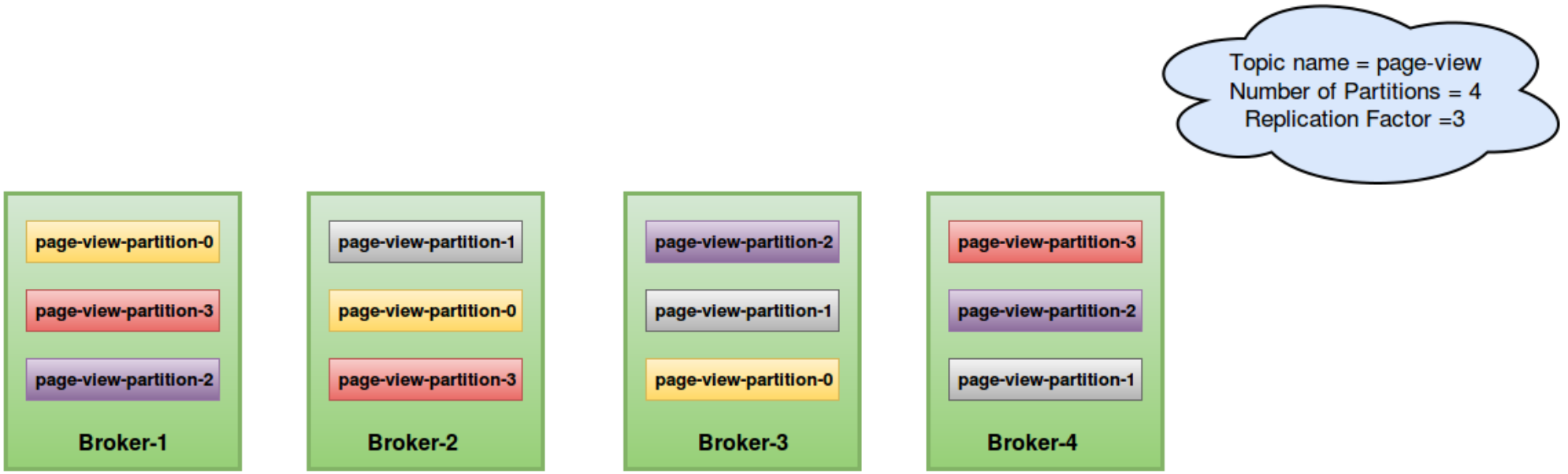
Partitions distribution



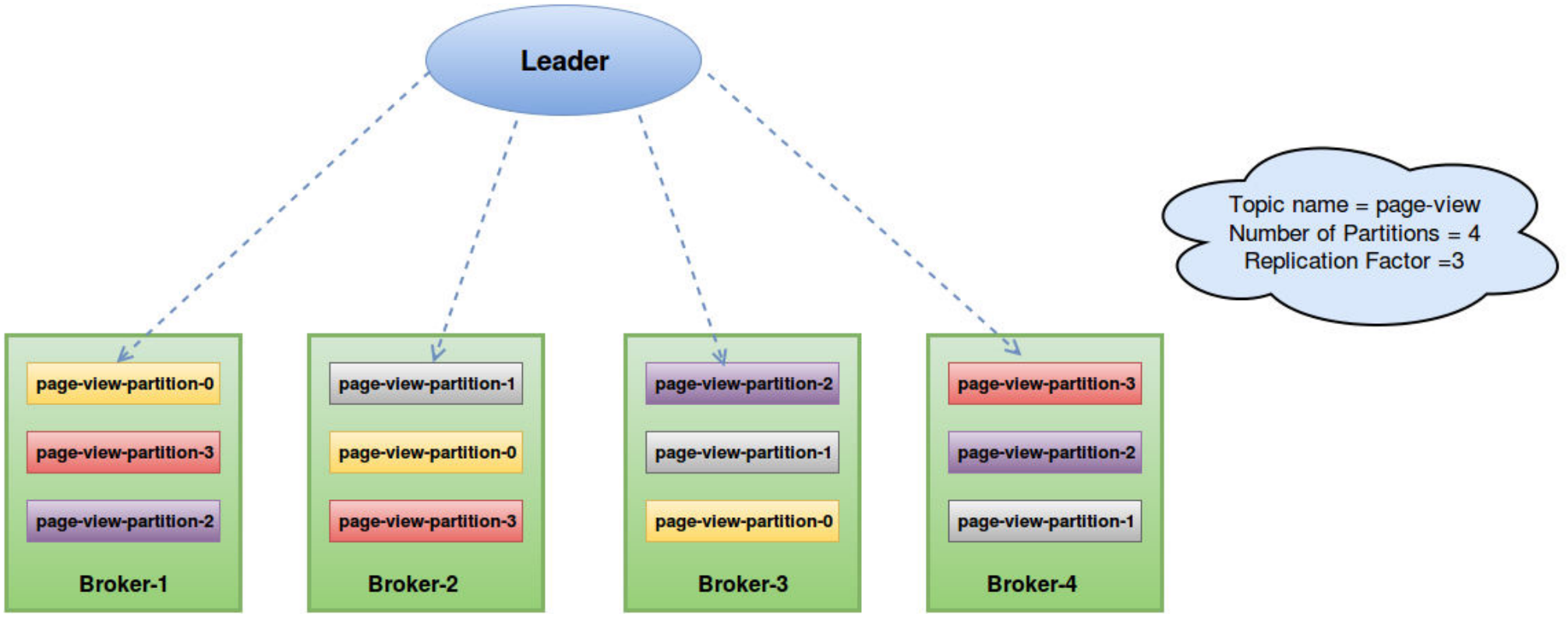
Partitions distribution



Partitions distribution

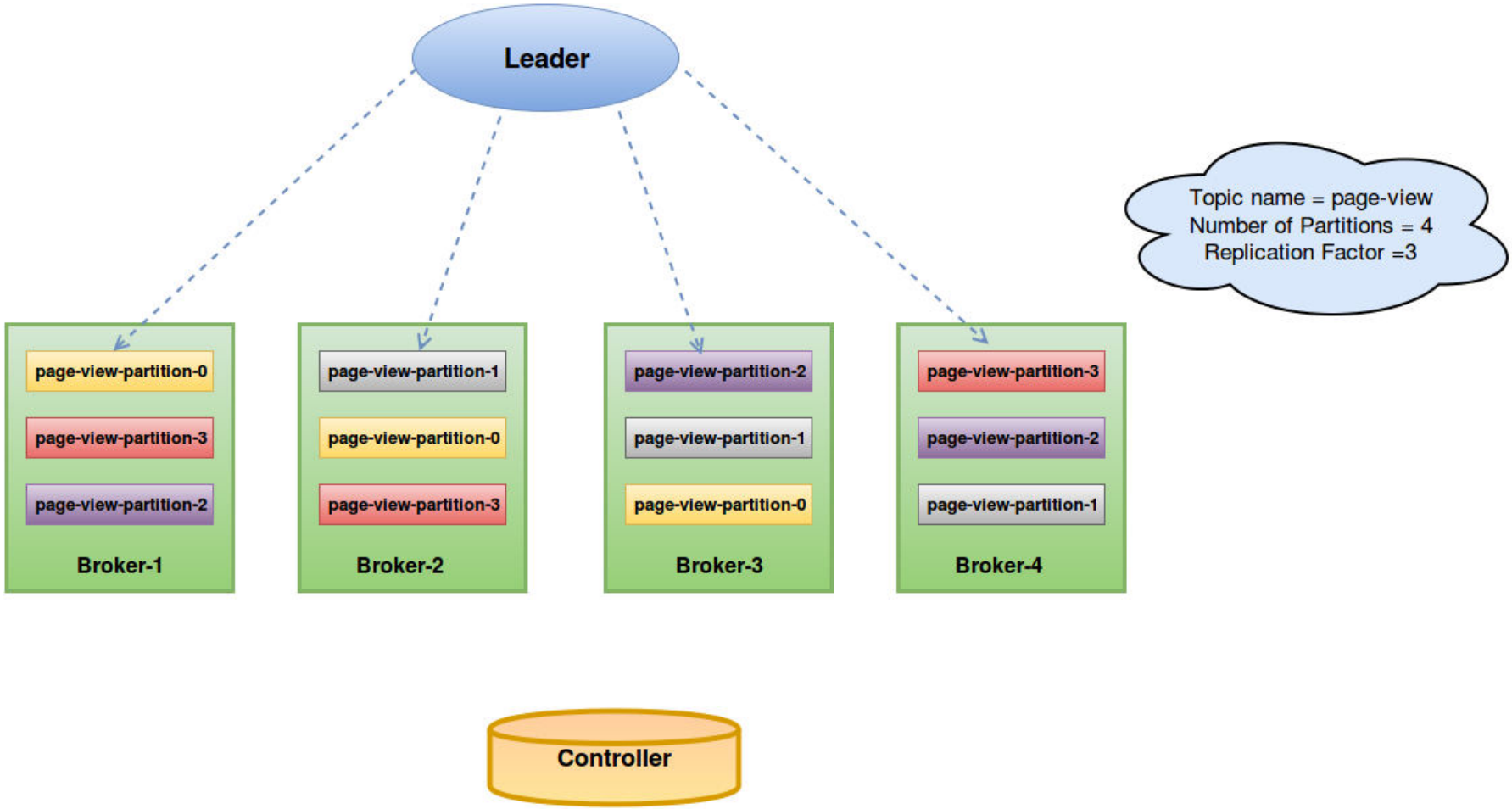


Partitions Distribution

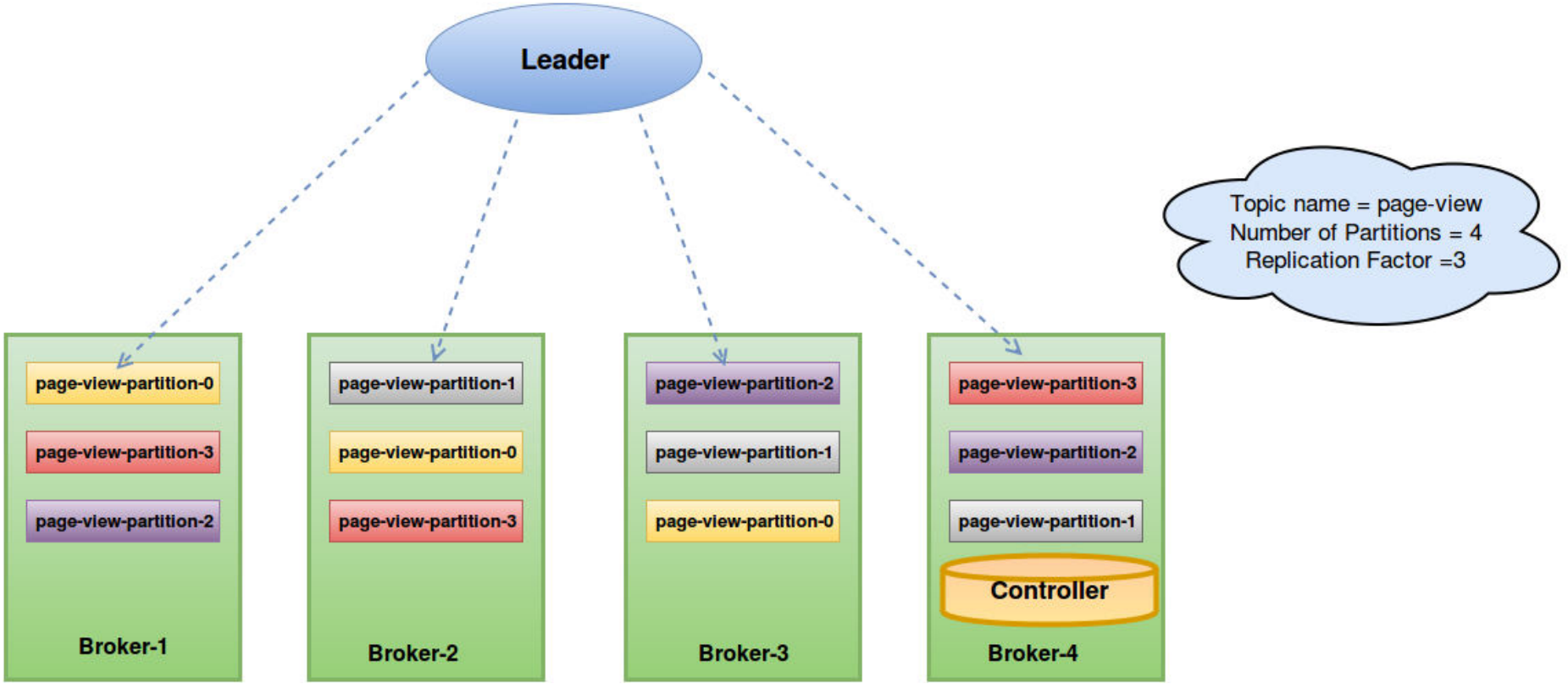


Who is responsible for these tasks ?

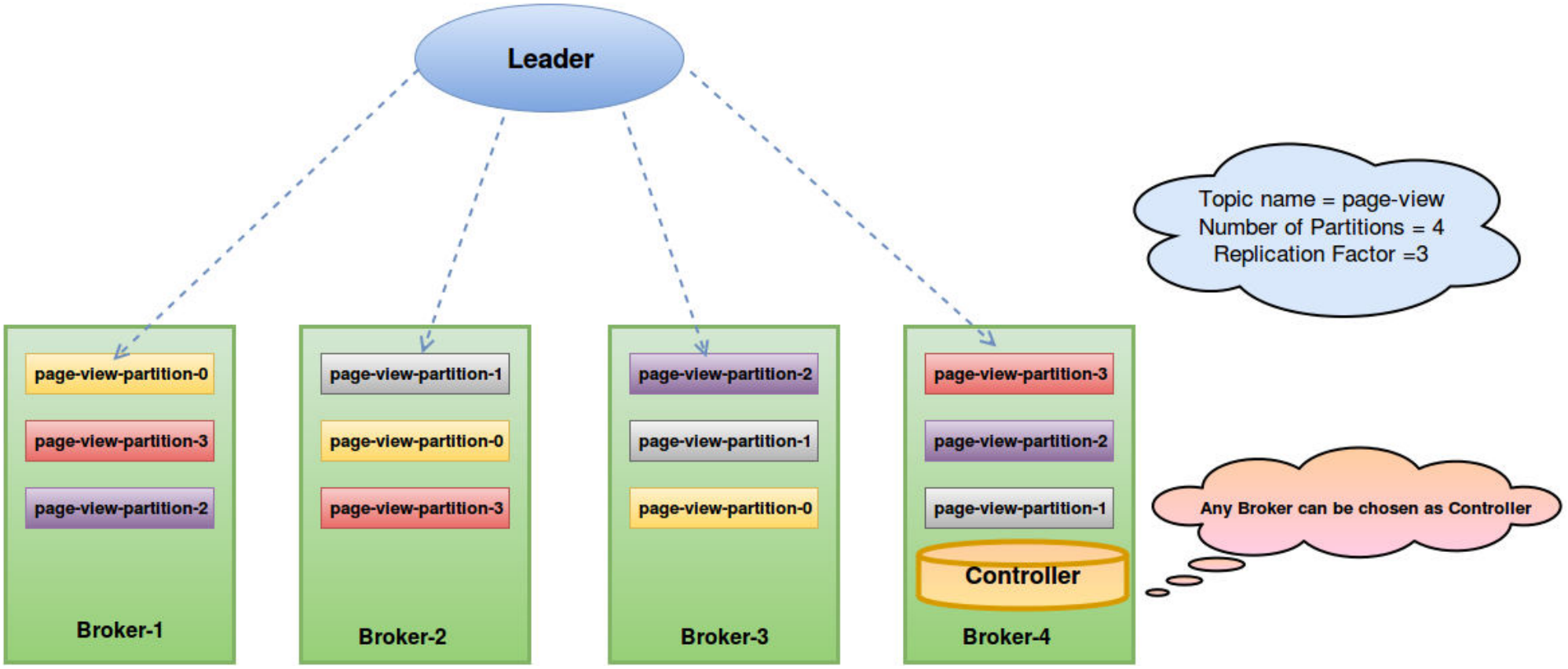
Partitions Distribution



Partitions Distribution



Partitions Distribution



Responsibility Of Controller

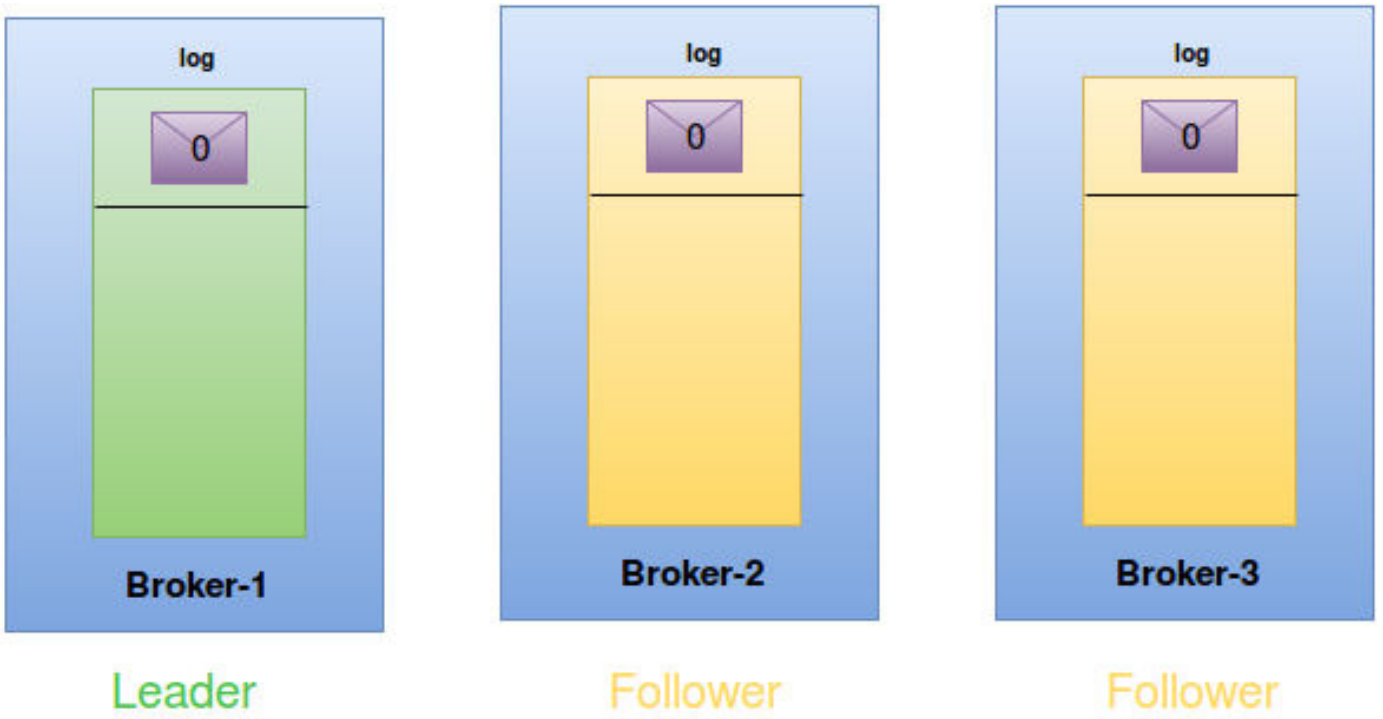
- managing the states of partitions and replicas
- performing administrative tasks like reassigning partitions

Roles For Partition

- Each partition has one server which acts as the "leader" and zero or more servers which act as "followers".
- The leader handles all read and write requests for the partition while the followers passively replicate the leader.
- If the leader fails, one of the followers will automatically become the new leader.
- Each server acts as a leader for some of its partitions and a follower for others so load is well balanced within the cluster.

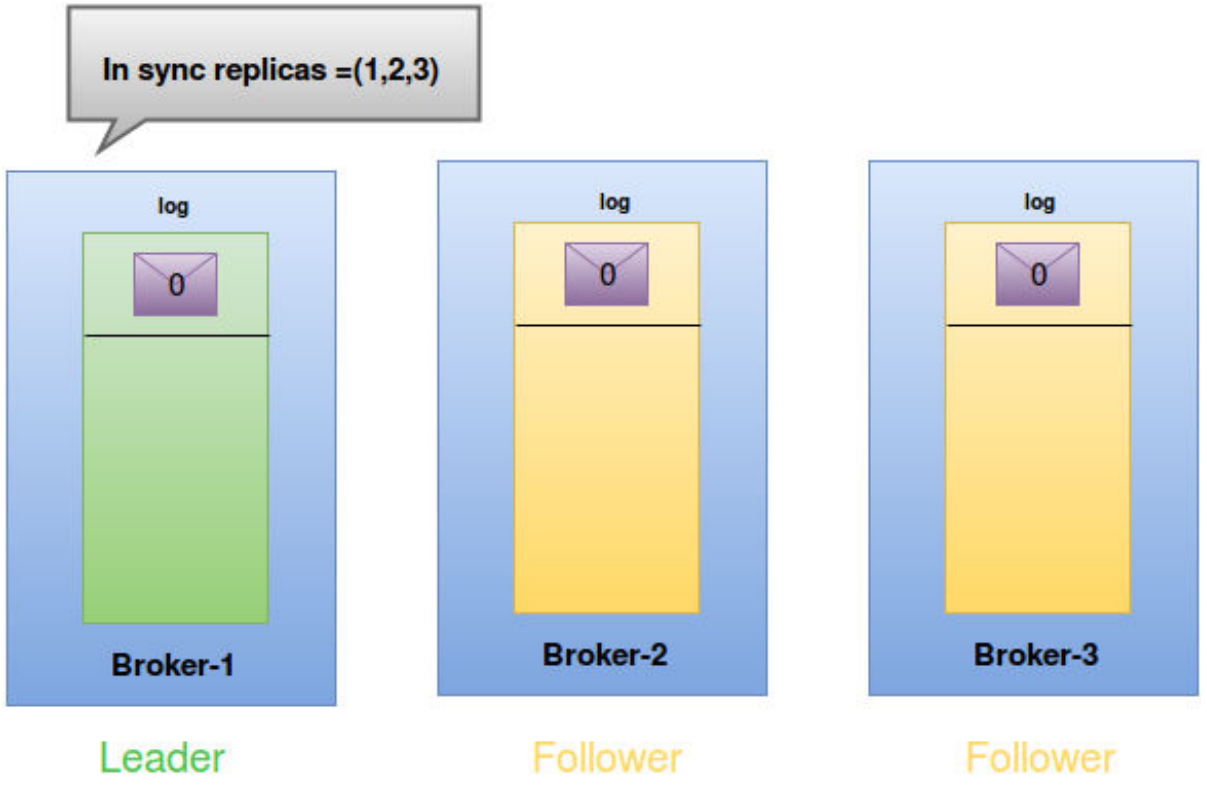
Replication Protocol

Replication Protocol



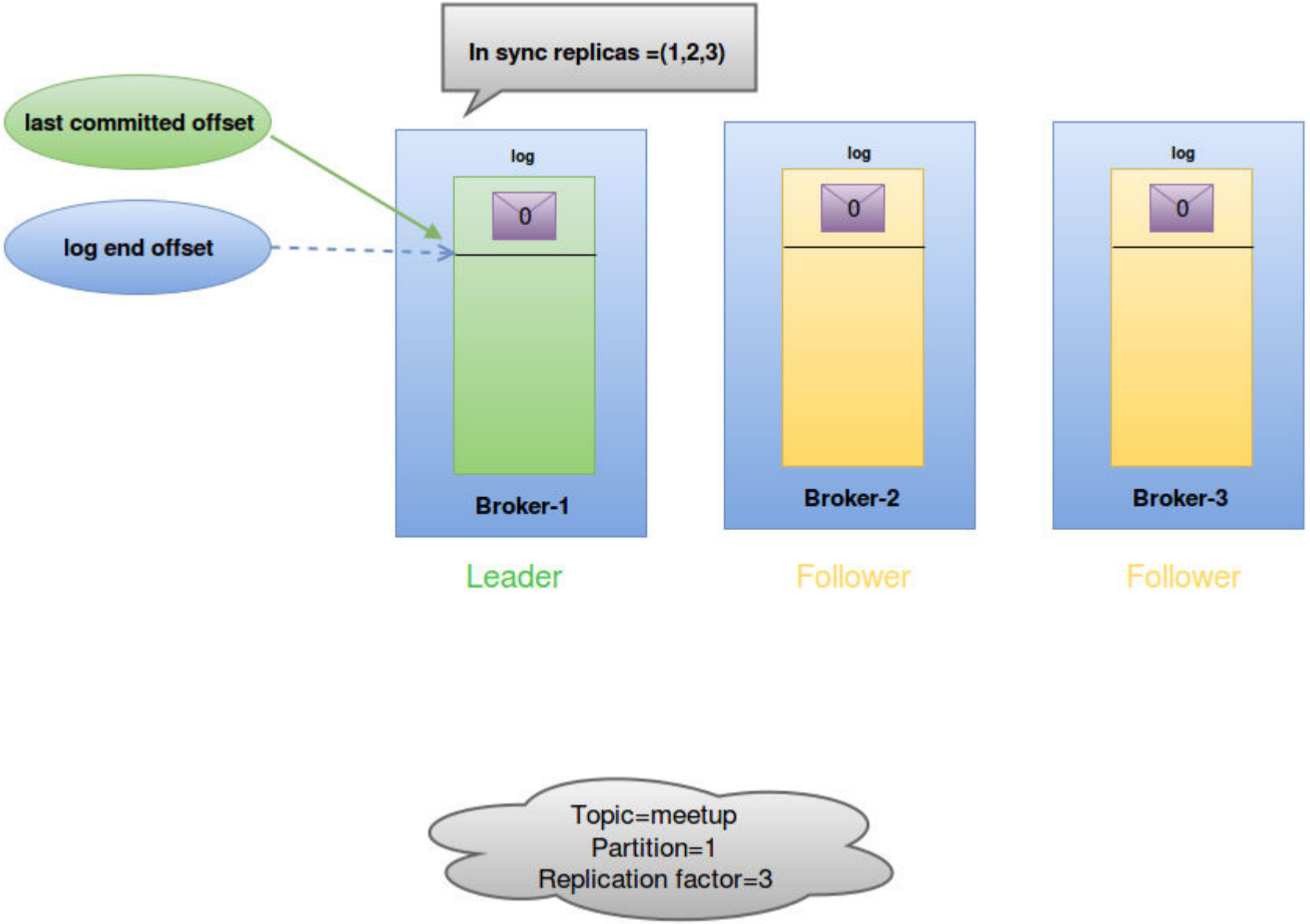
Topic=meetup
Partition=1
Replication factor=3

Replication Protocol



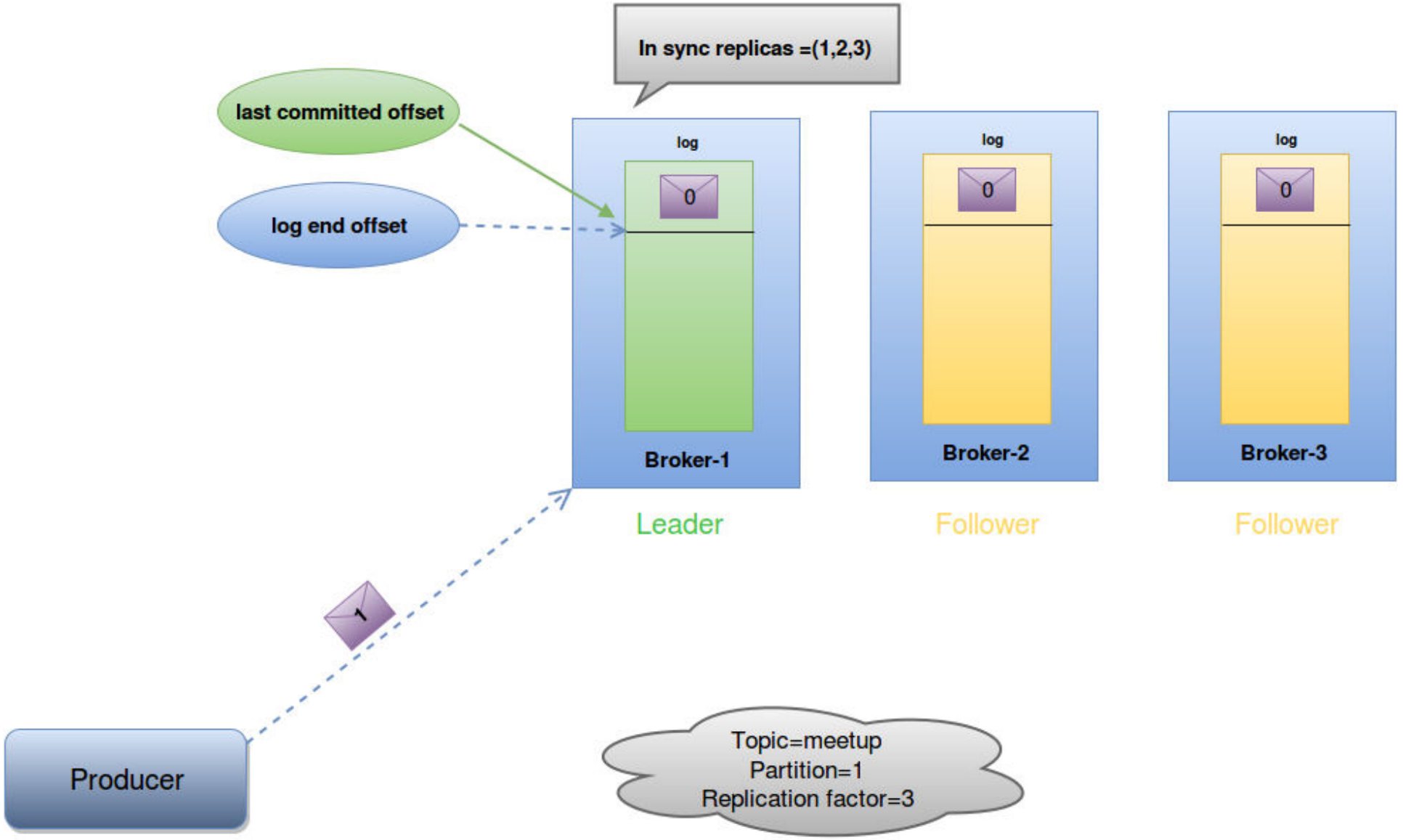
Topic=meetup
Partition=1
Replication factor=3

Replication Protocol

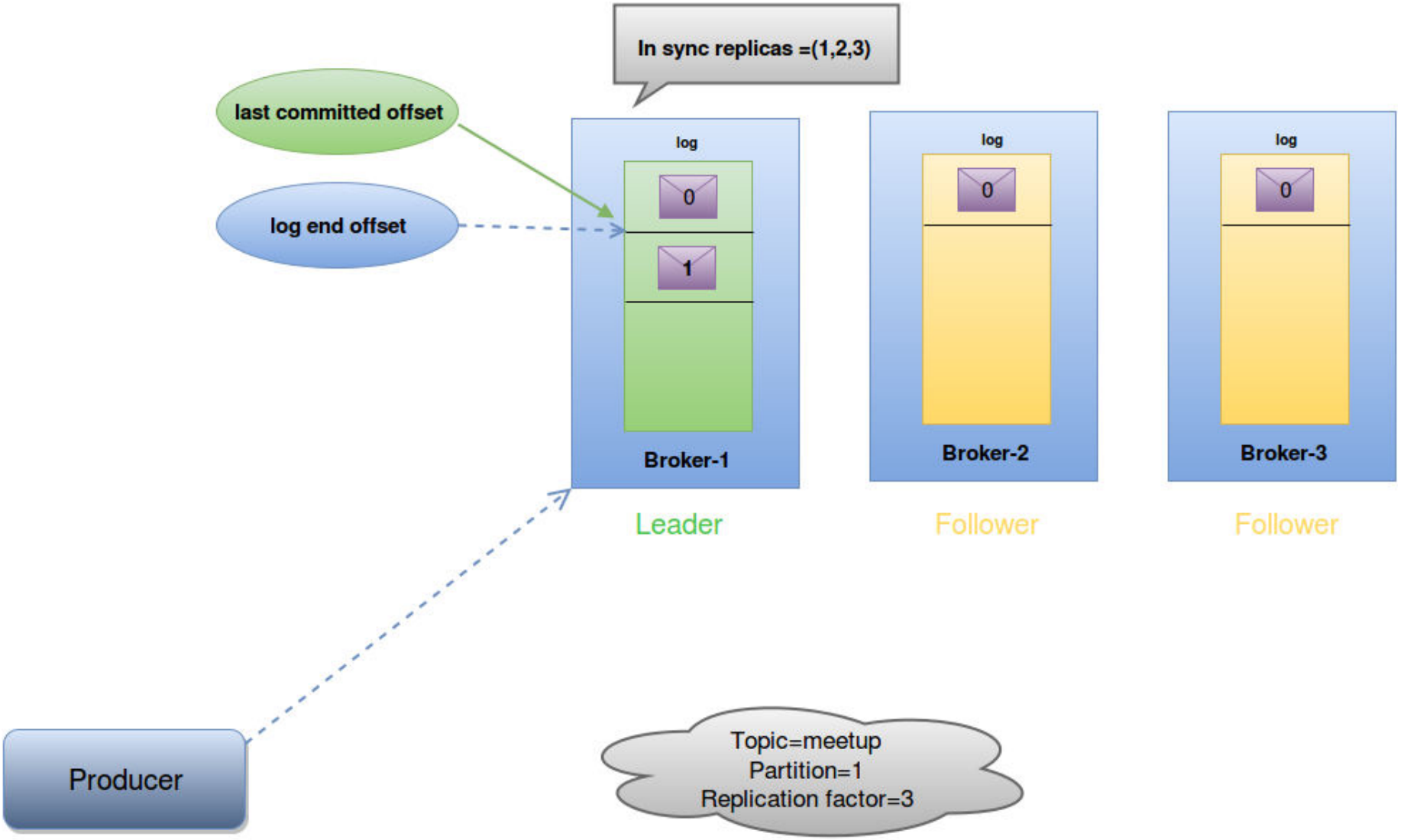


Topic=meetup
Partition=1
Replication factor=3

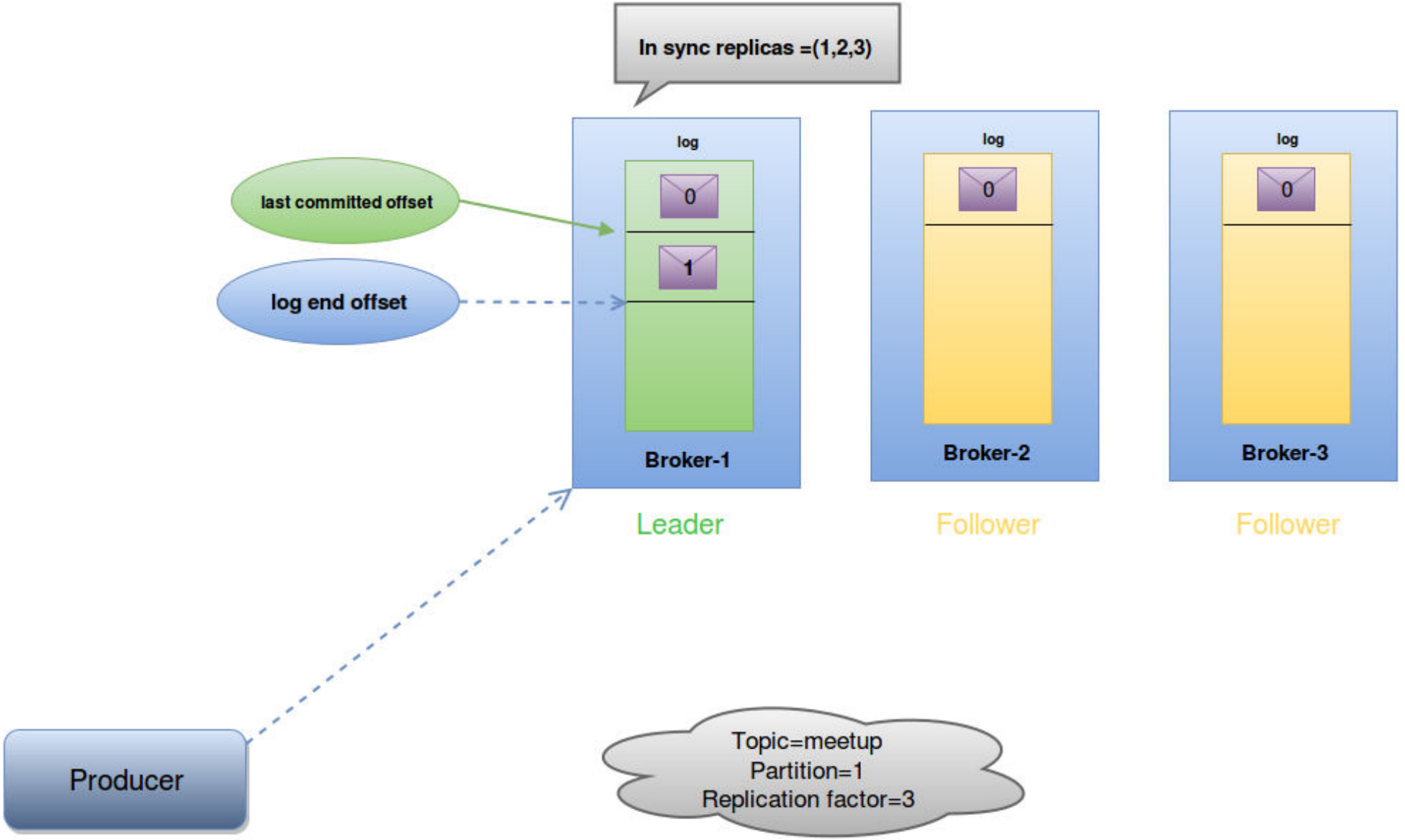
Replication Protocol



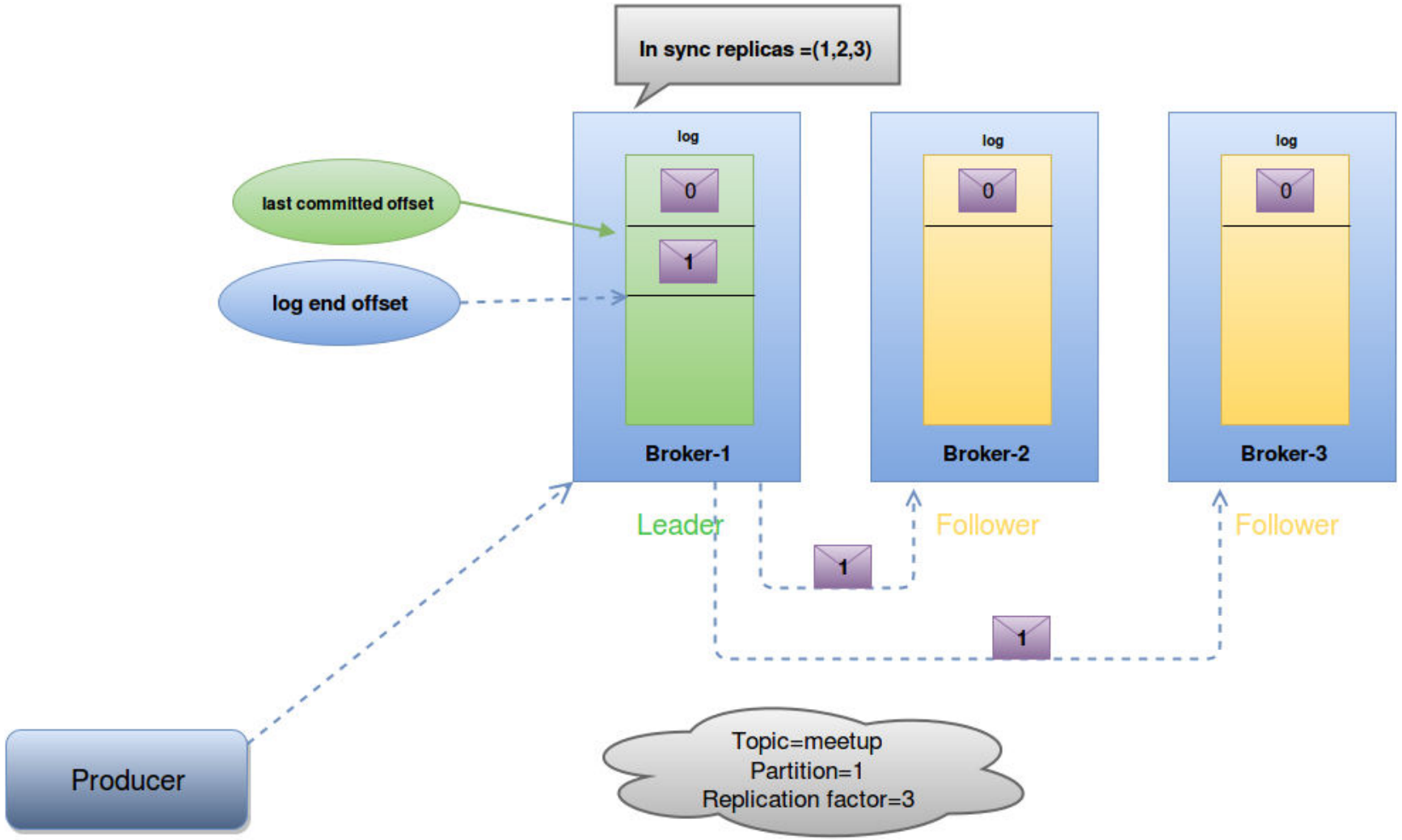
Replication Protocol



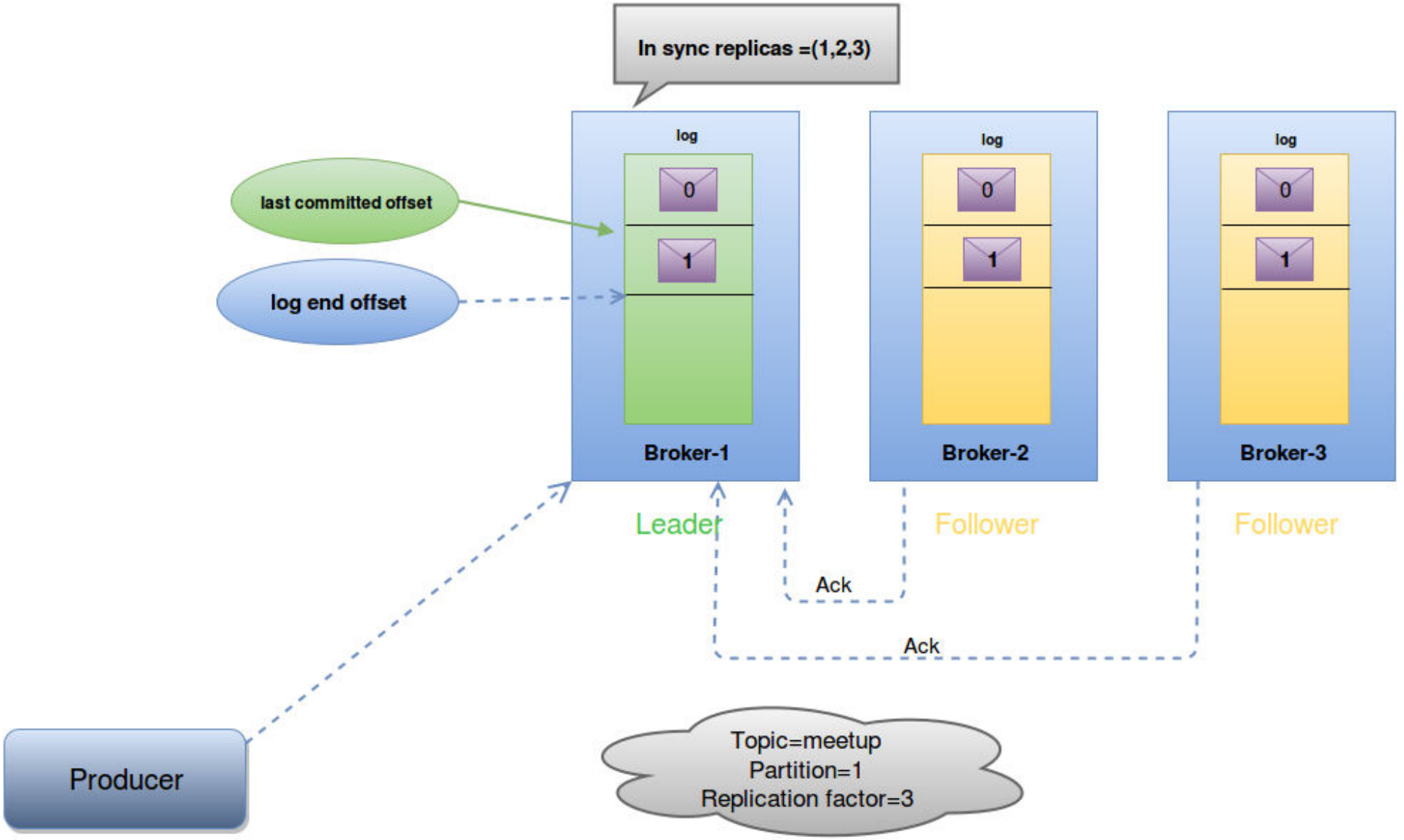
Replication Protocol



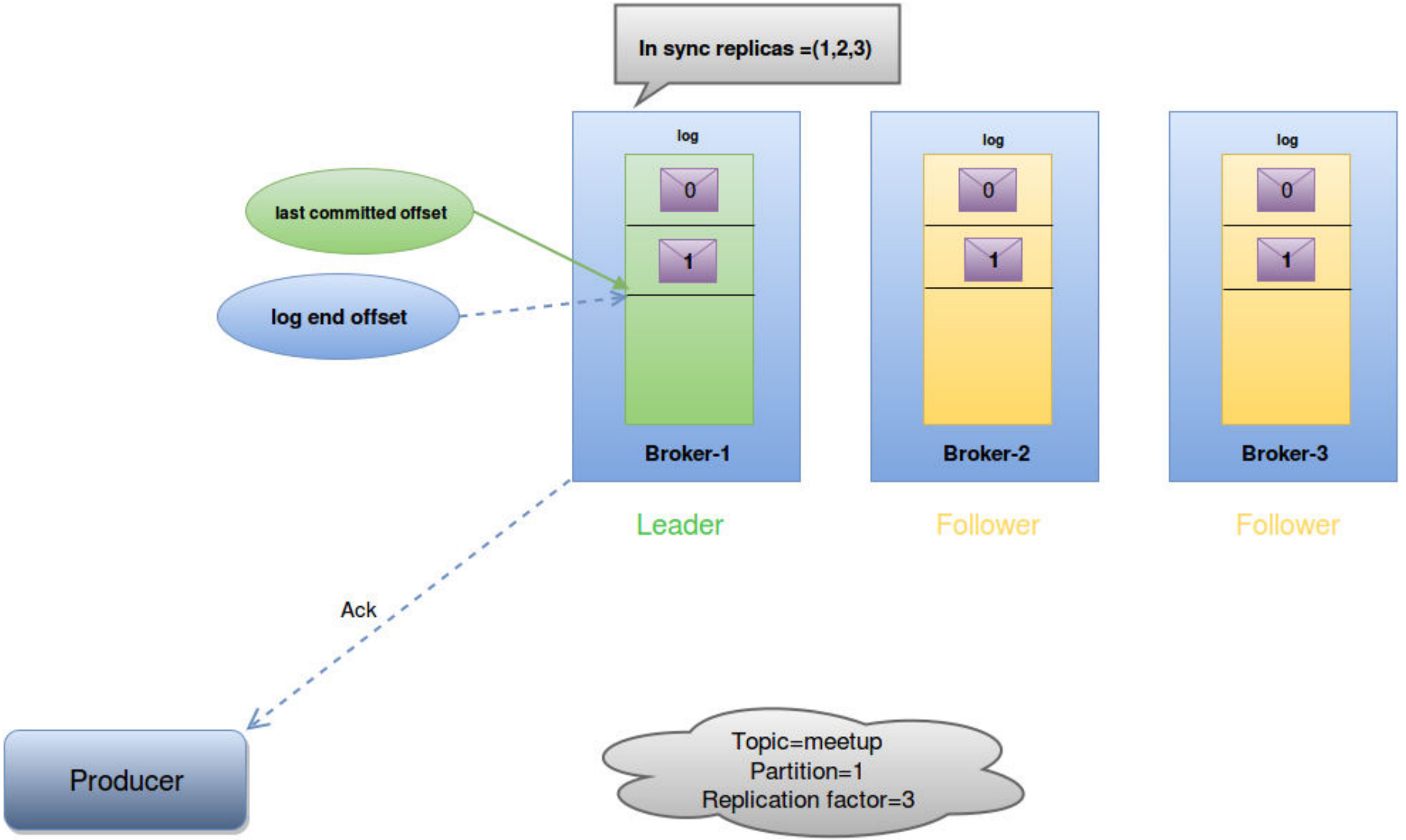
Replication Protocol



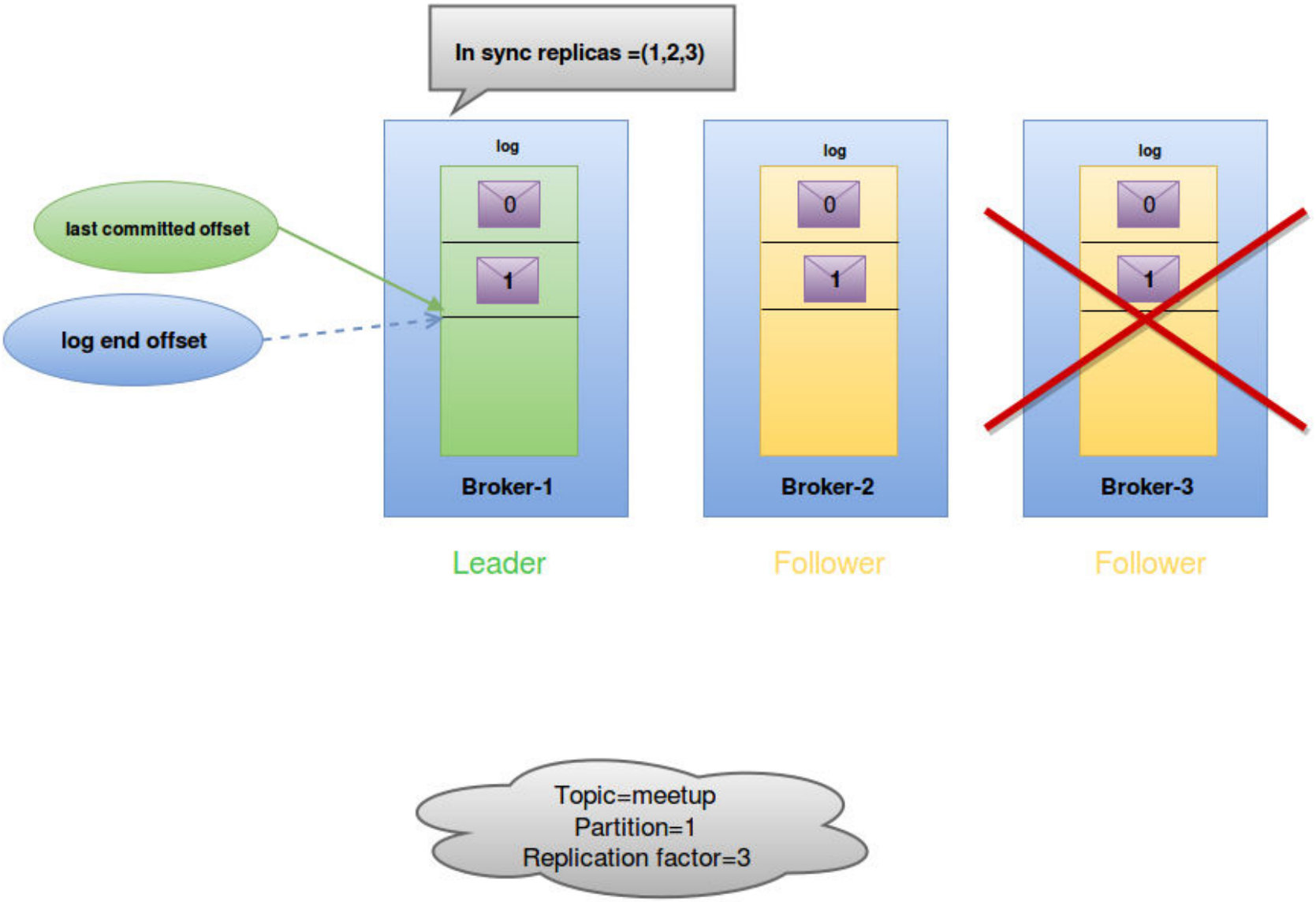
Replication Protocol



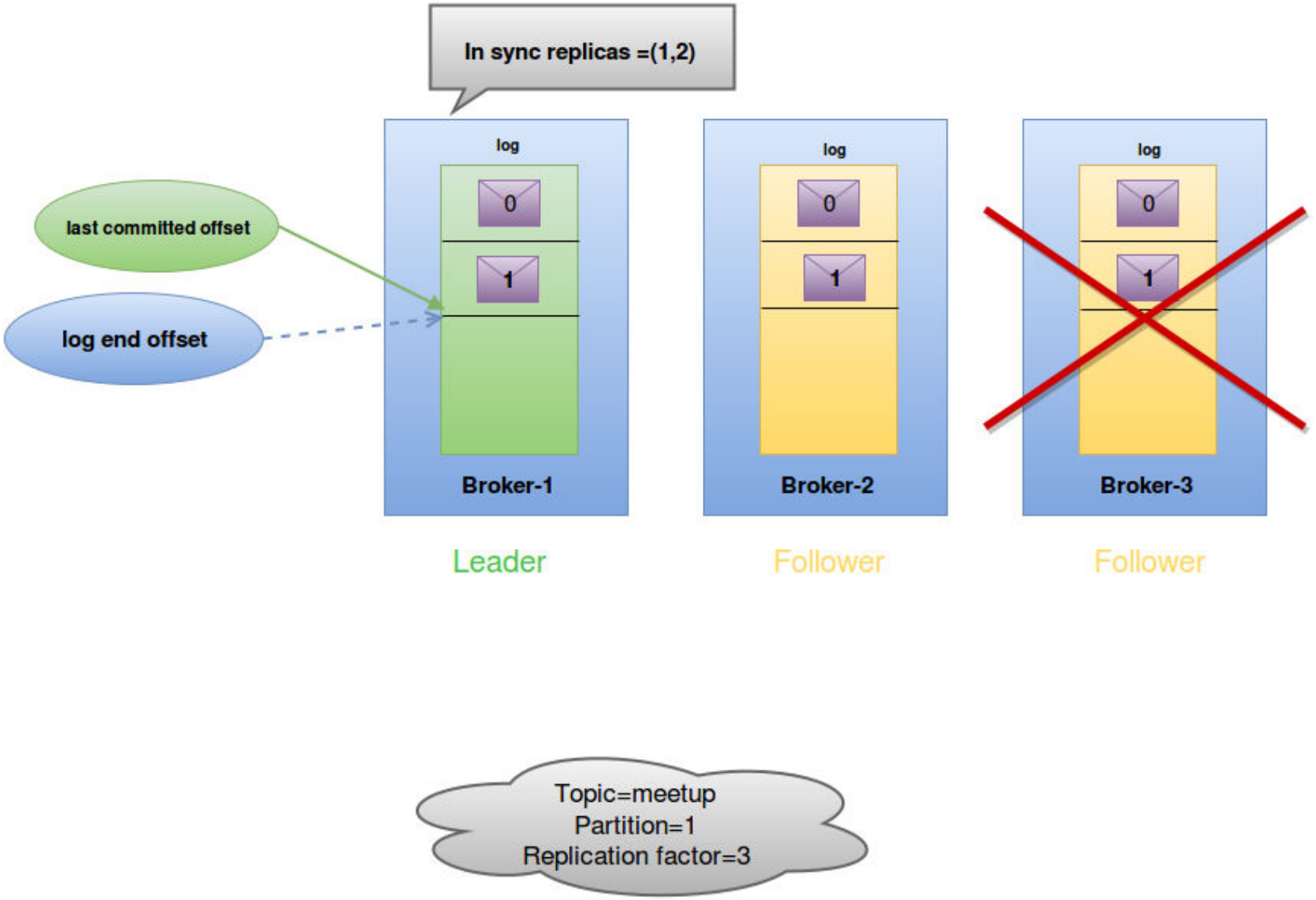
Replication Protocol



Replication Protocol

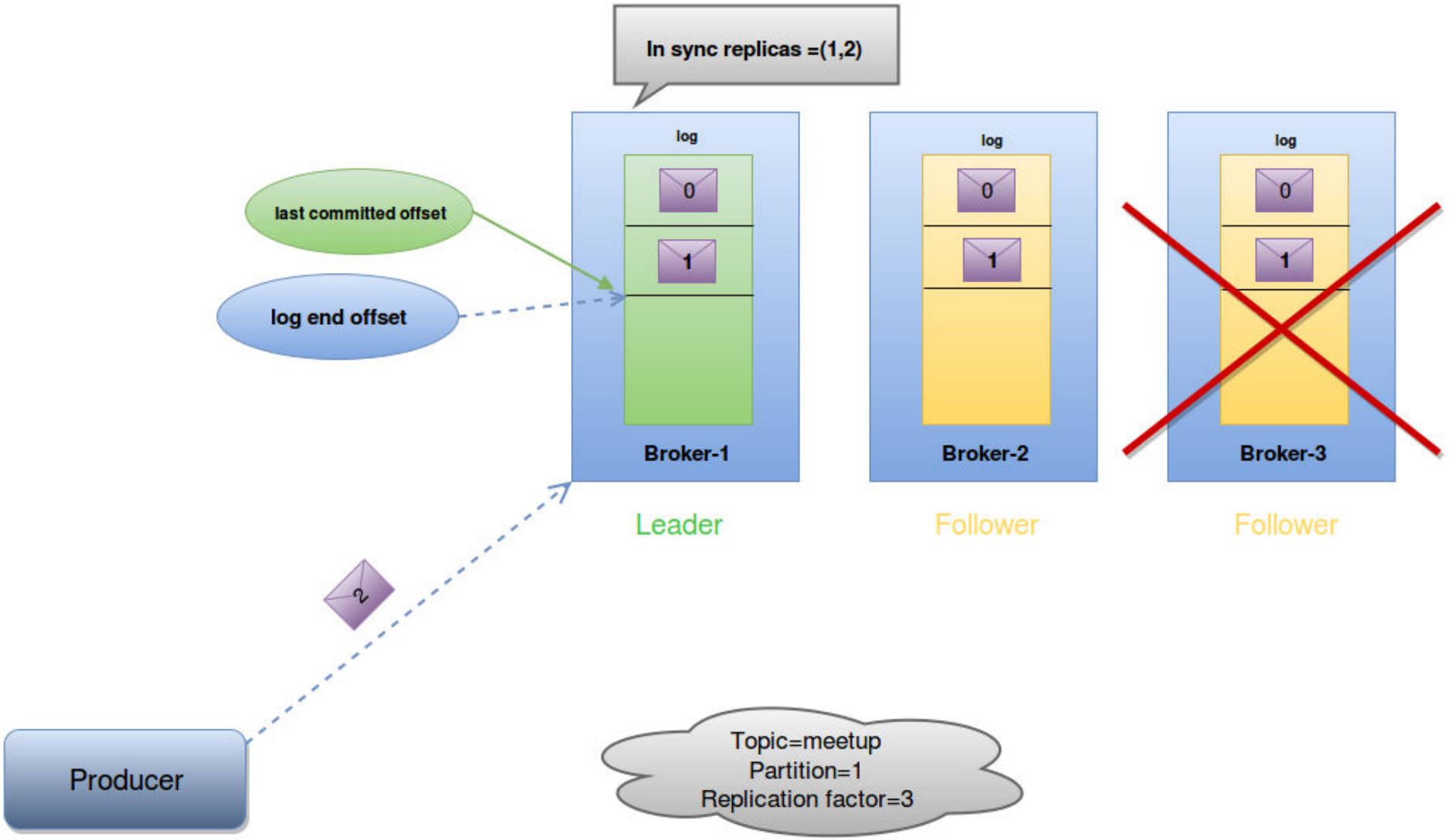


Replication Protocol

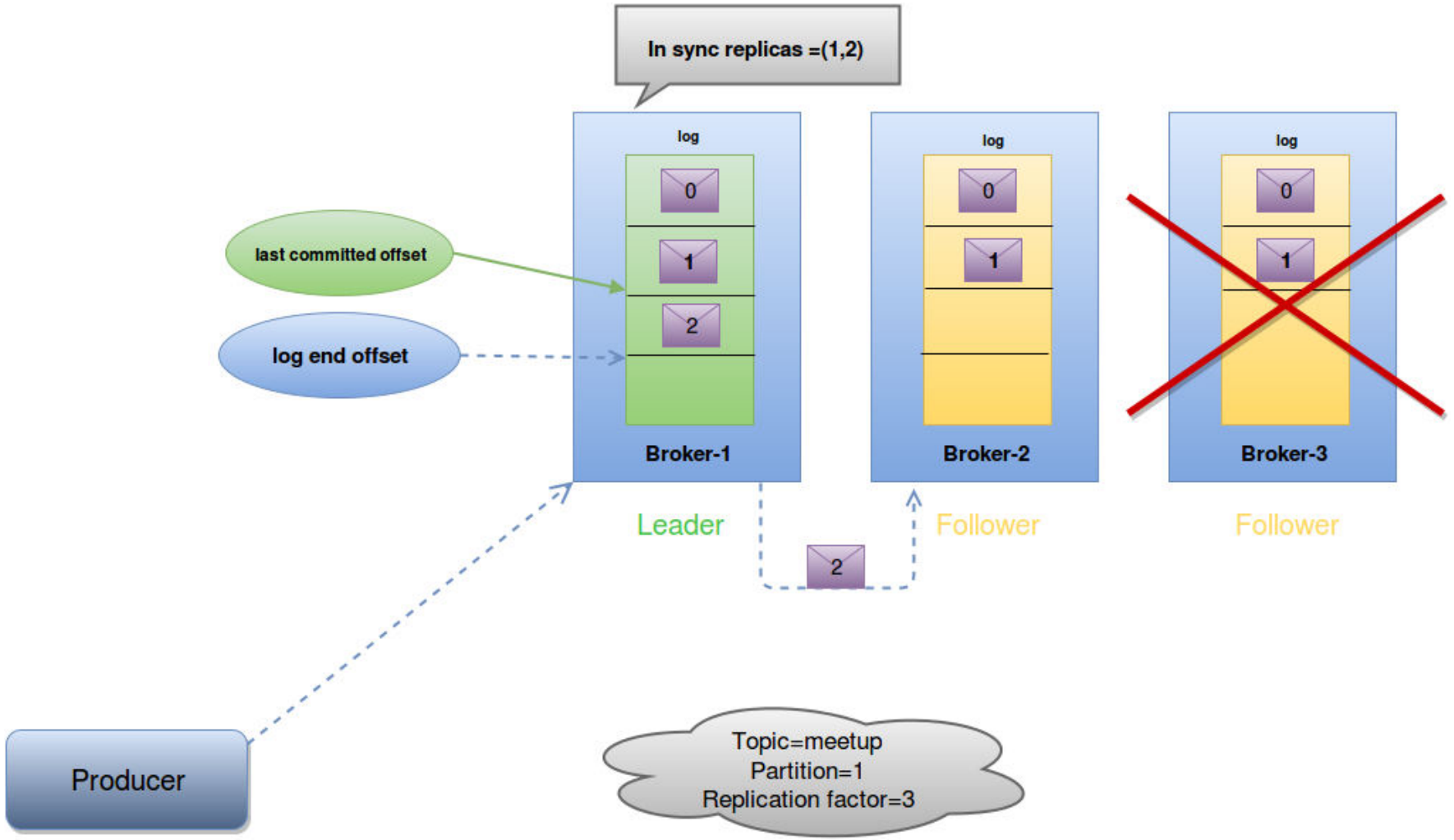


Topic=meetup
Partition=1
Replication factor=3

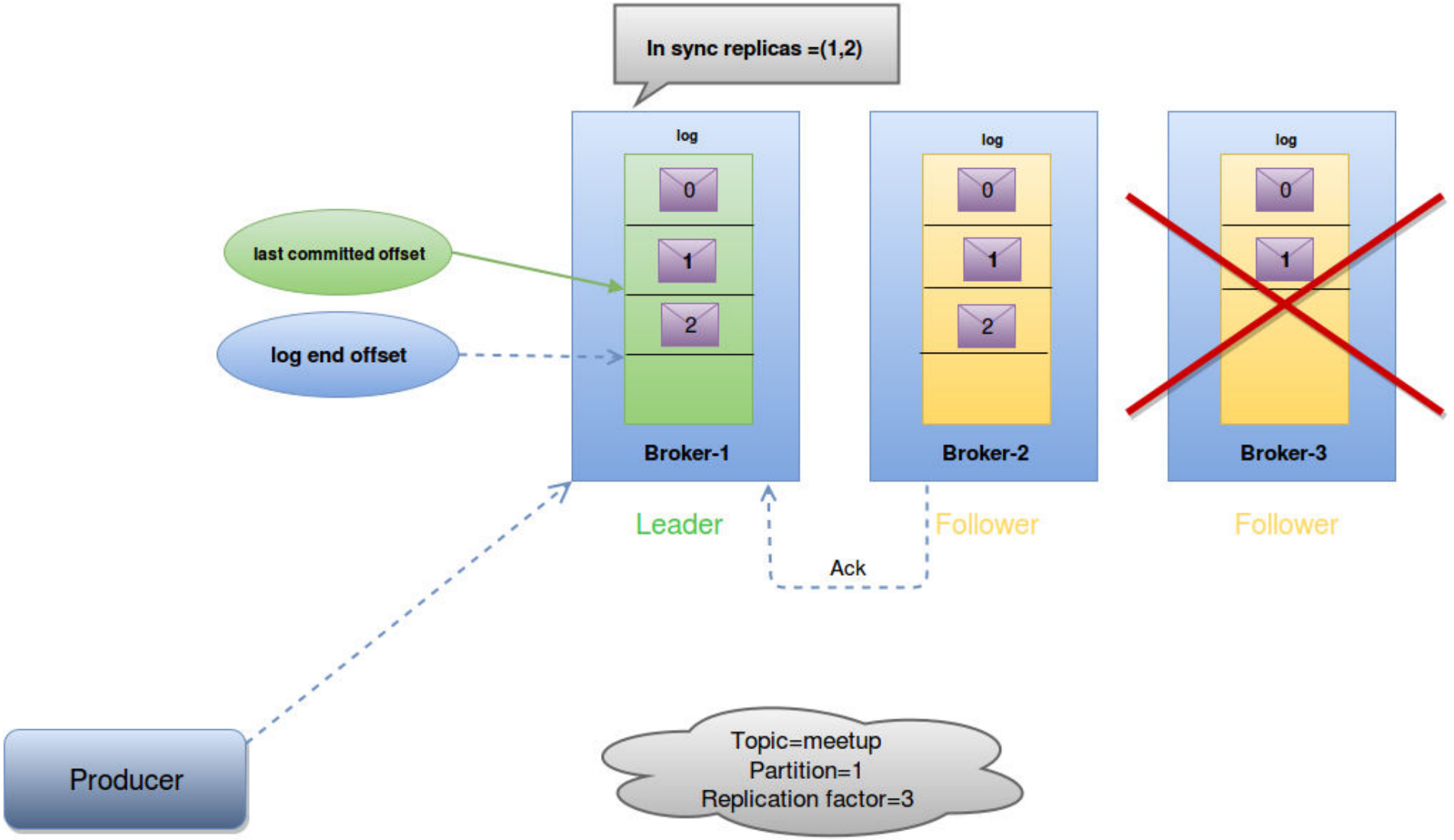
Replication Protocol



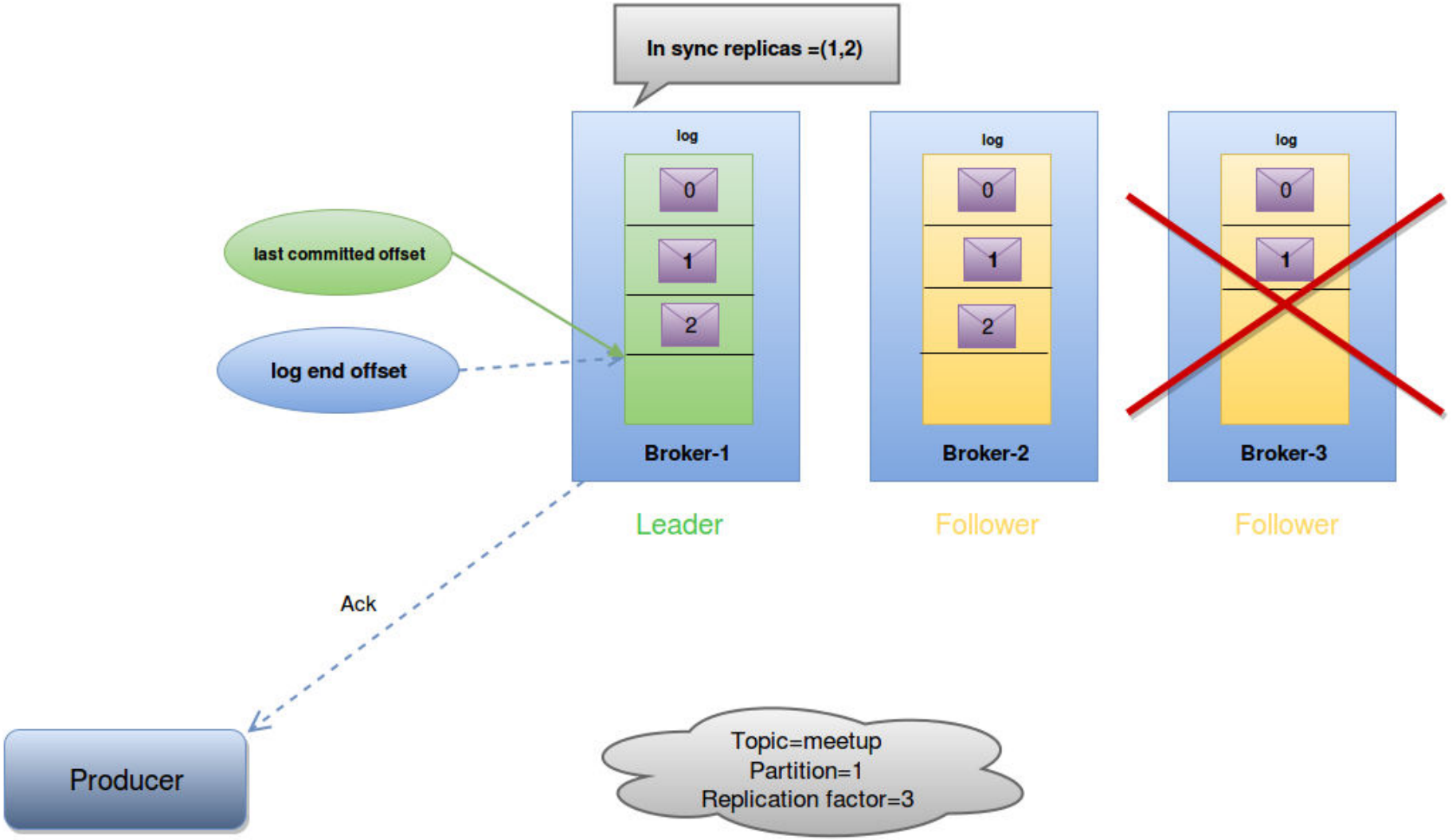
Replication Protocol



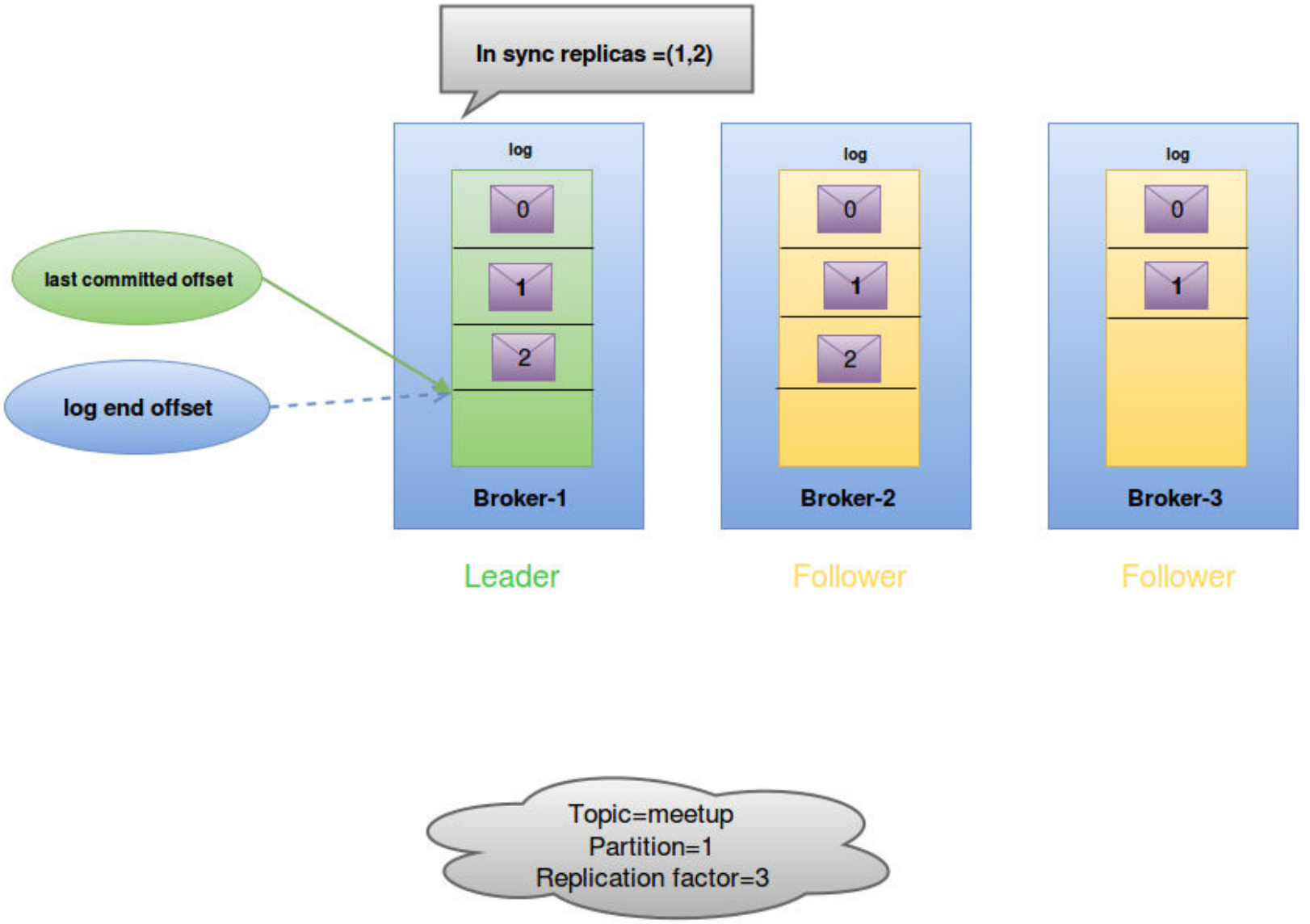
Replication Protocol



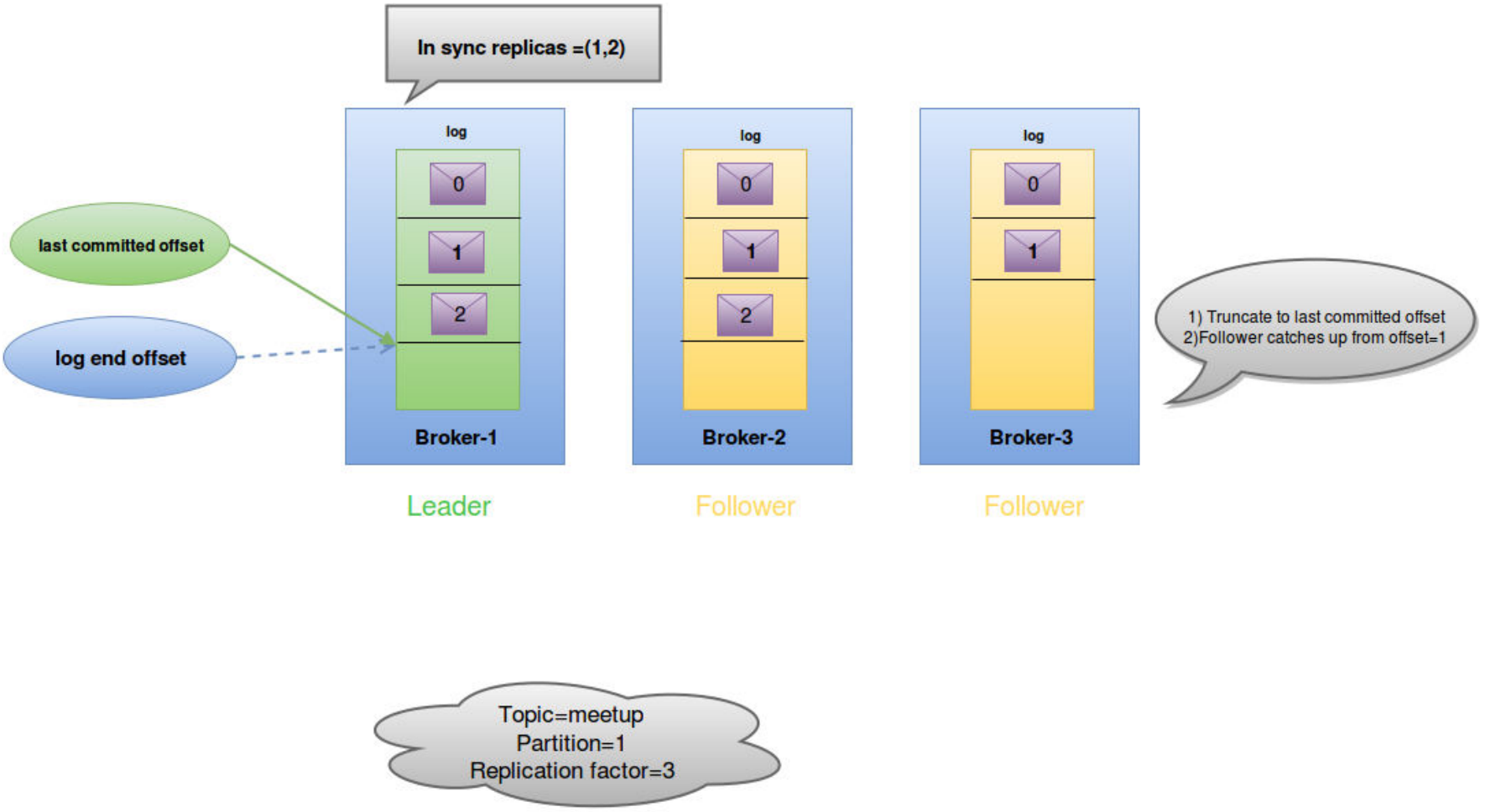
Replication Protocol



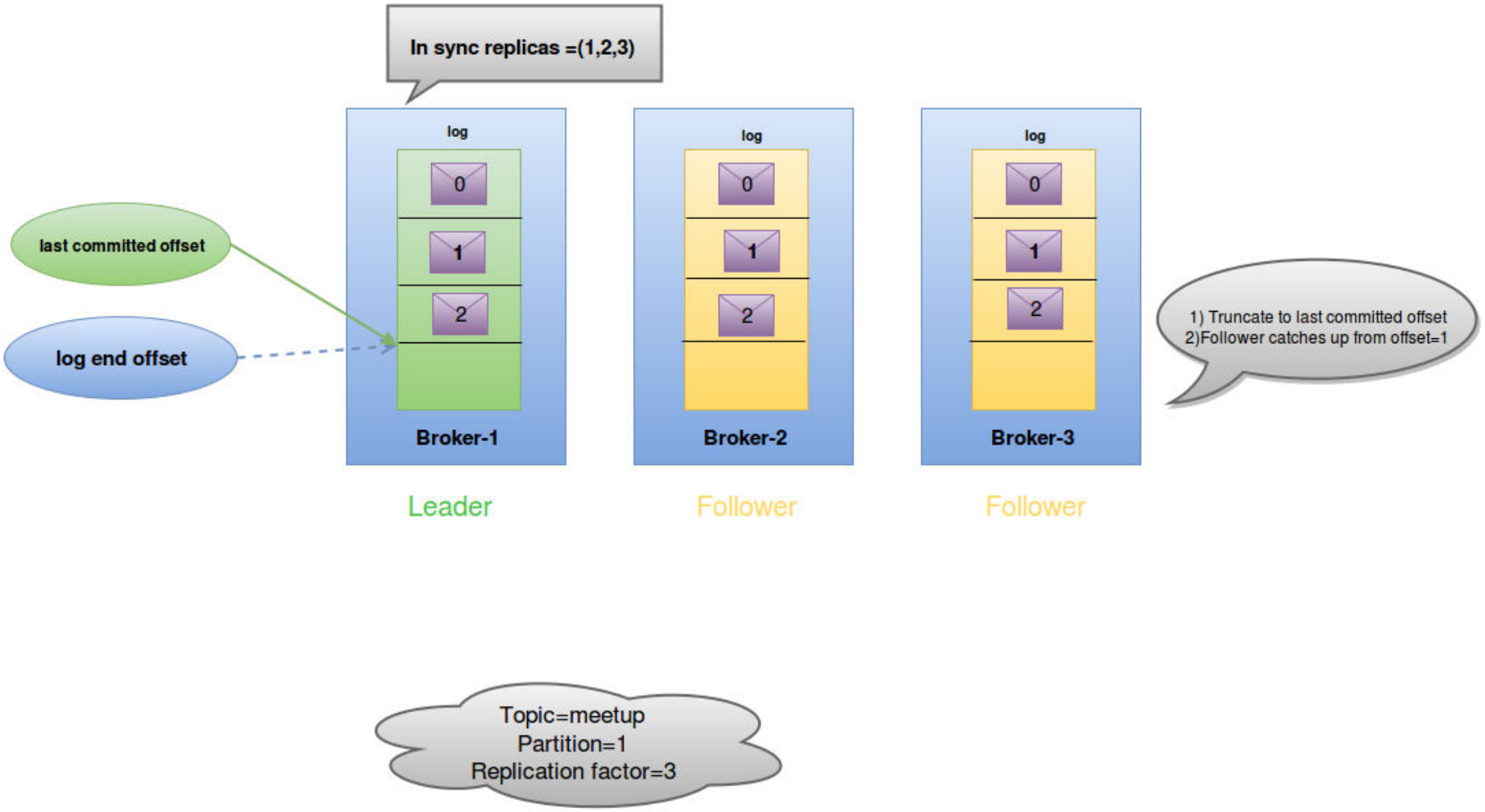
Replication Protocol



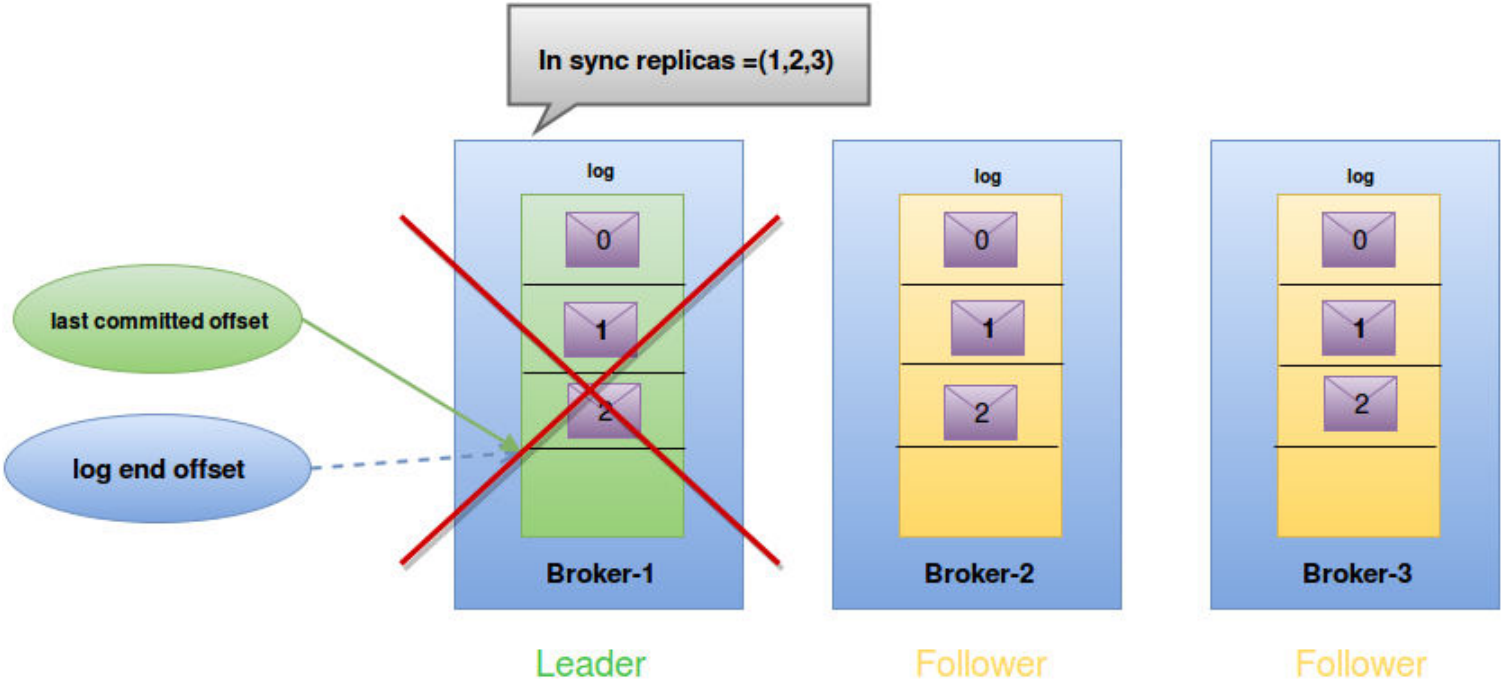
Replication Protocol



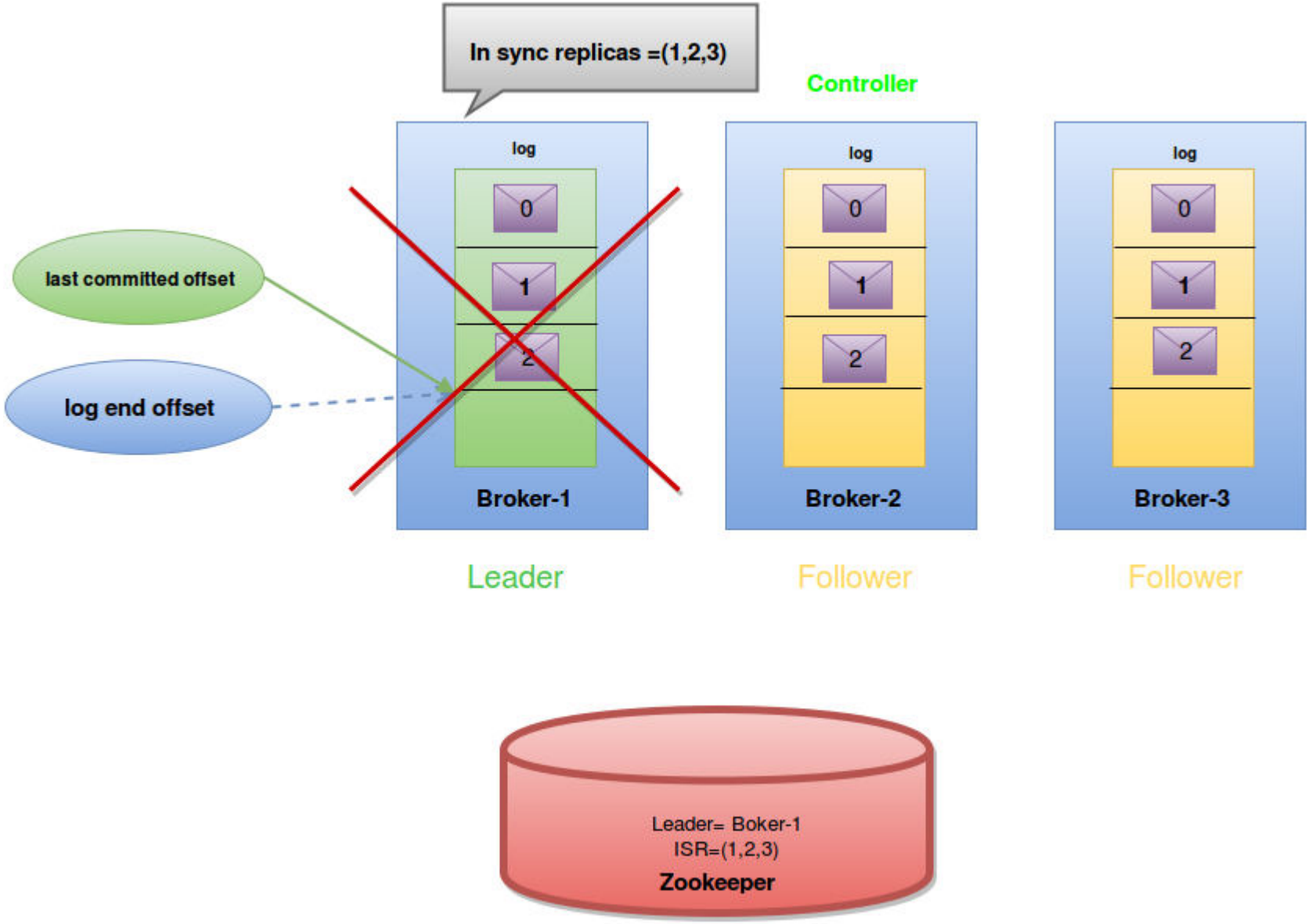
Replication Protocol



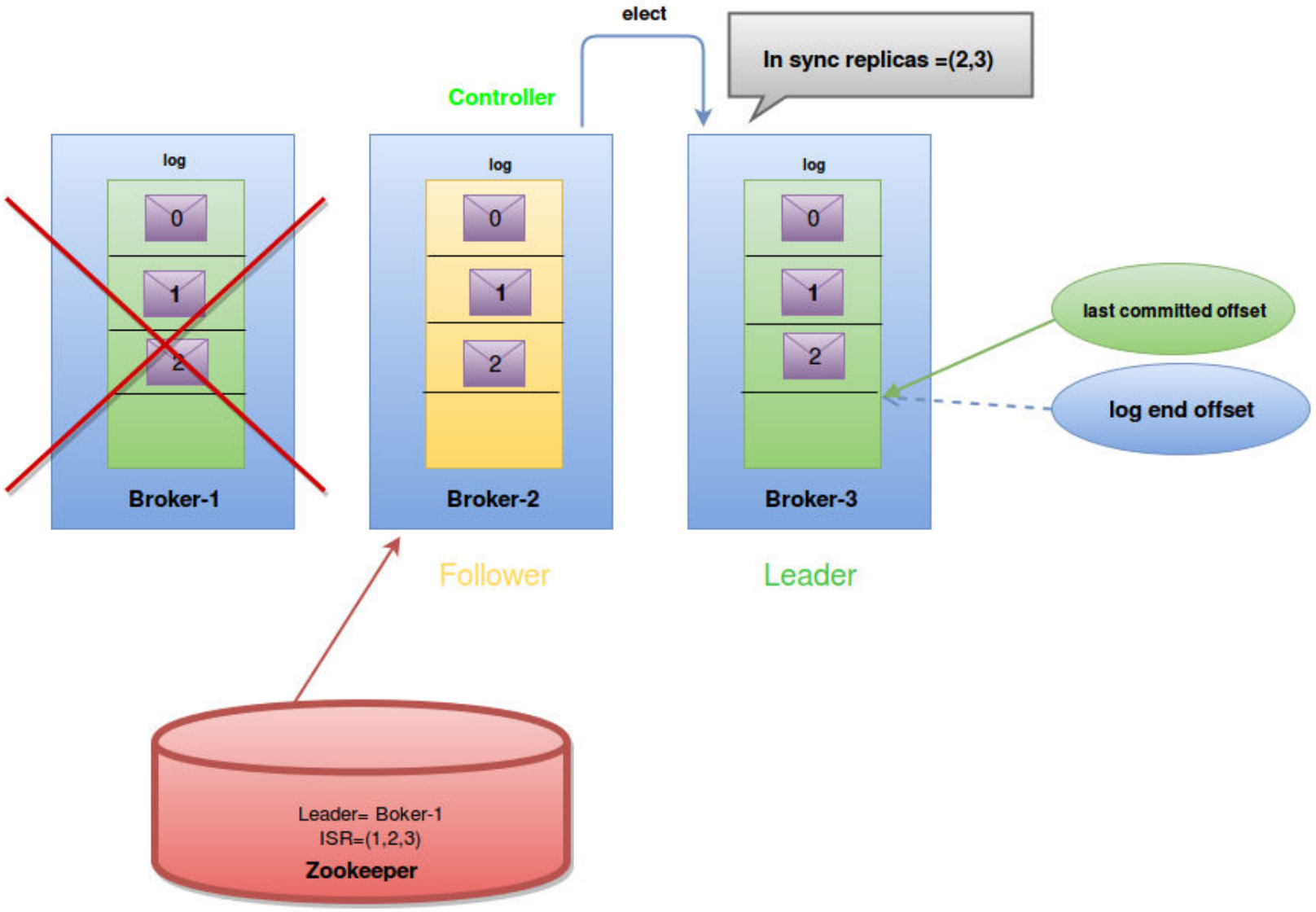
Replication Protocol



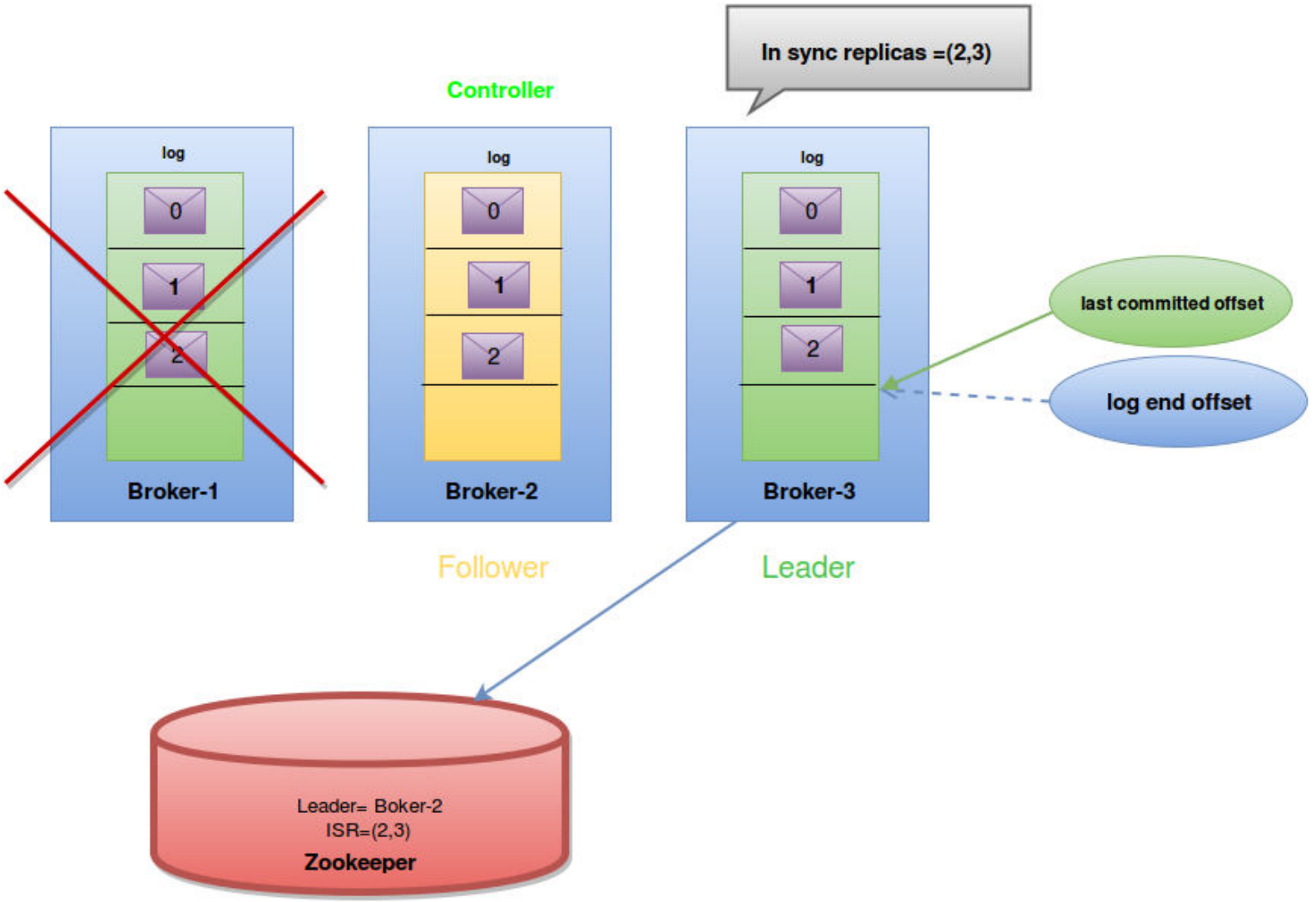
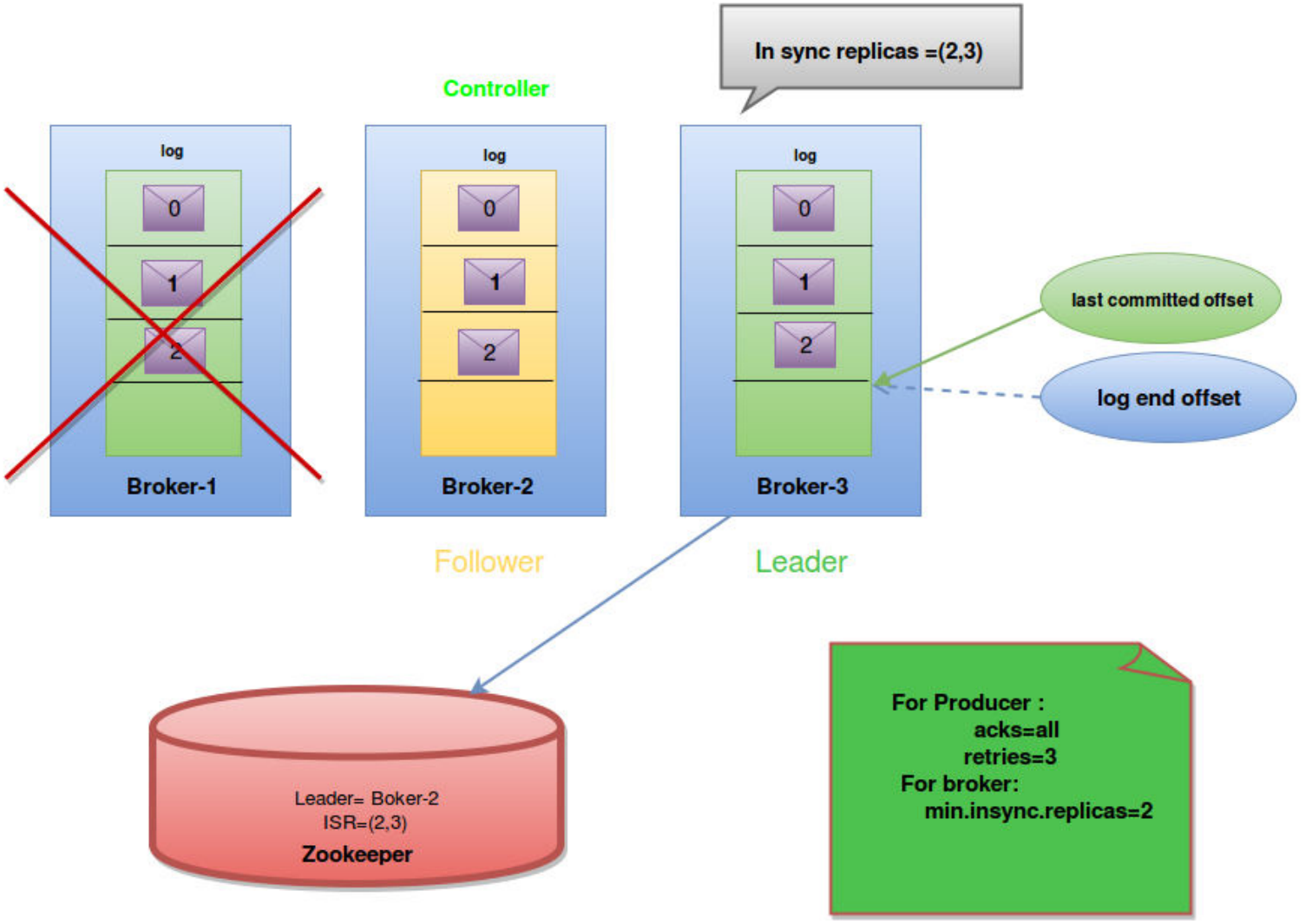
Replication Protocol



Replication Protocol



Replication Protocol



Replication Protocol