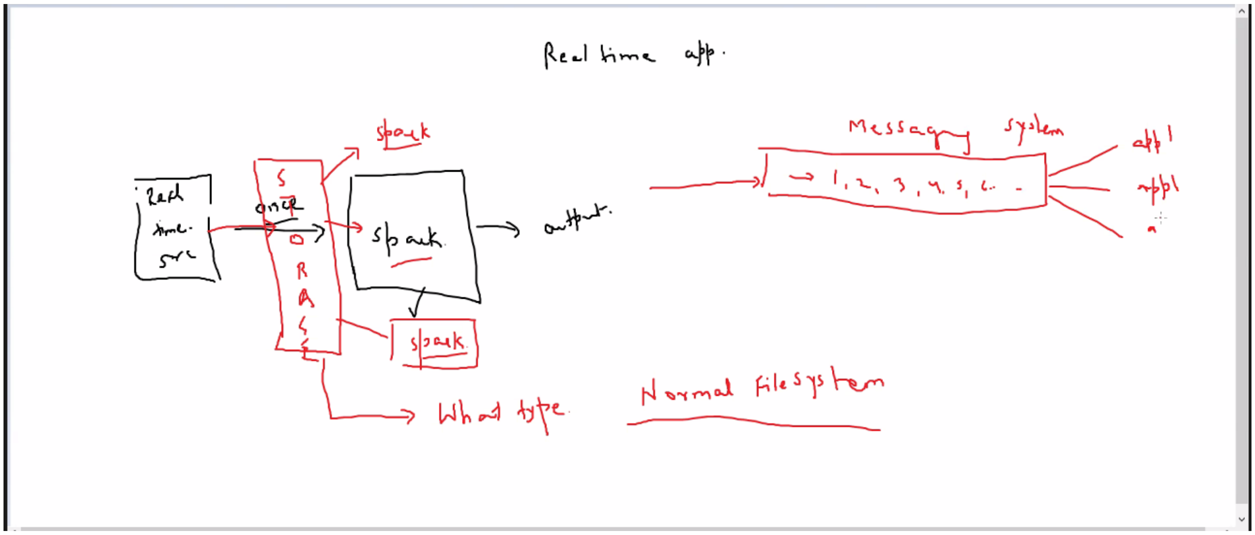
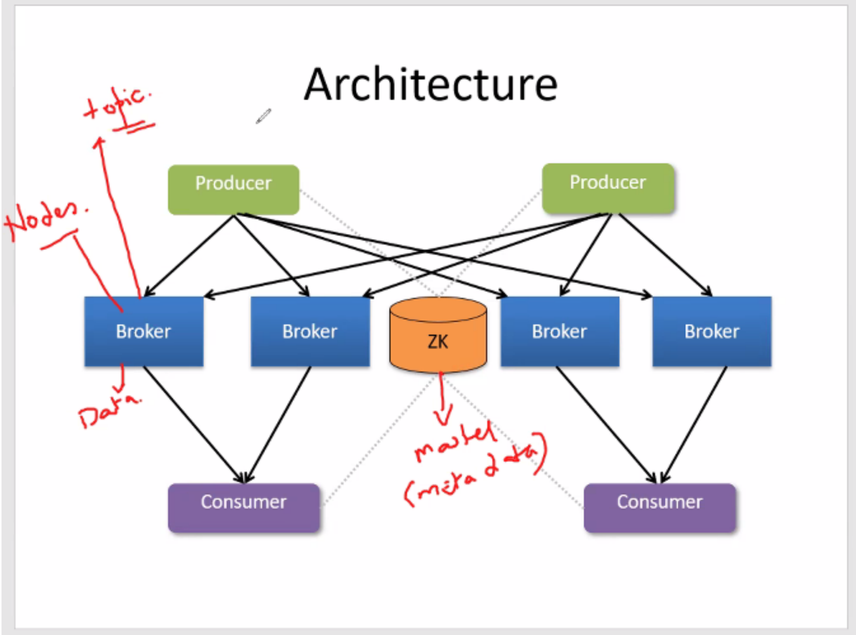
→ Kafka



a. Kafka cluster, called cluster because it has multiple brokers
b. Using kafka, we collect real-time data from different publishers
c. distributed publish-subscribe messaging system
d. Publisher is sender and consumer/subscriber is receiver
e. Topic:
    i. an object where data is stored
    ii. Can have many K, V paired messages
    iii.
f. Once order is collected from publishers, then it is sent to spark app which is a subscriber, and then spark app sends to different data warehouses
g. One partition for each broker, but each partition can have different number of messages
h. Kafka does not process data, it’s just a storage layer
i.



Architecture

Producer

Producer

Broker

Broker

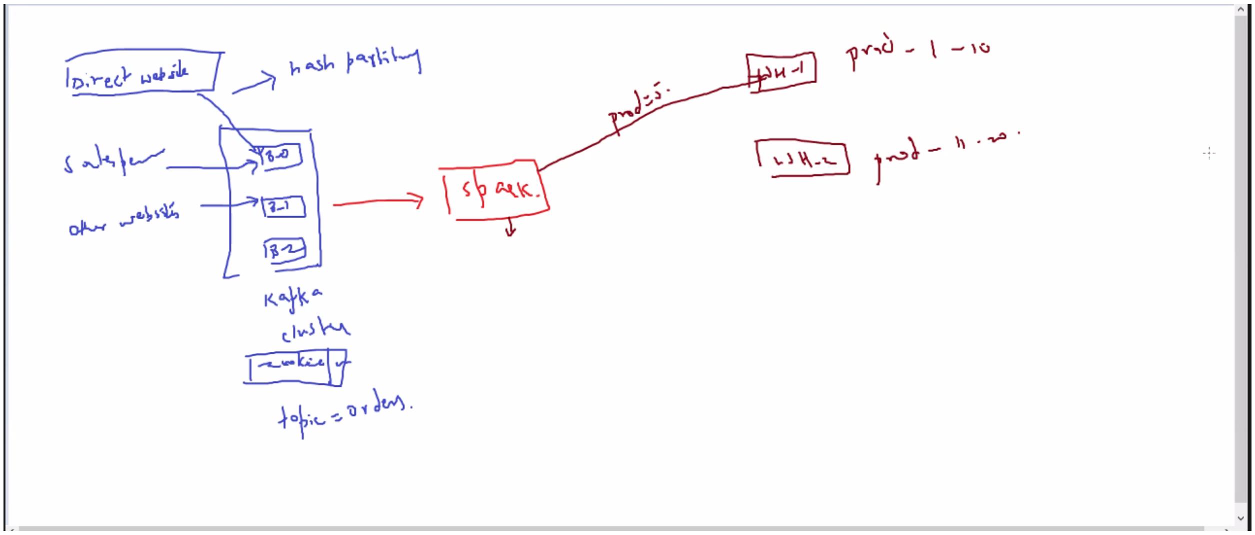
ZK

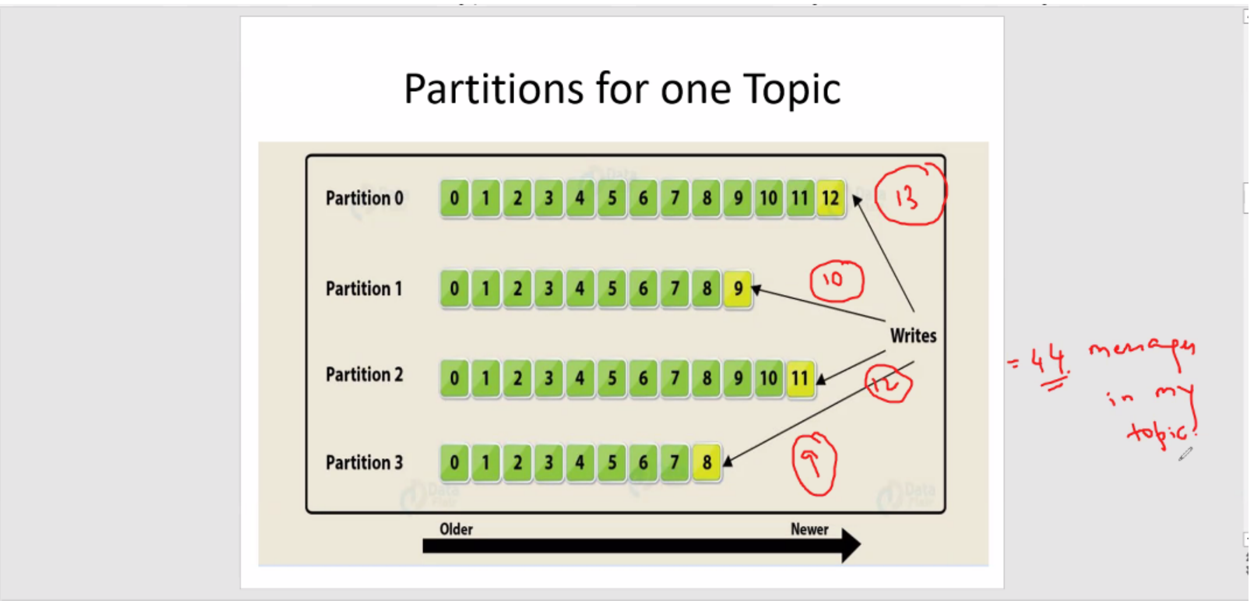
Broker

Broker

Consumer

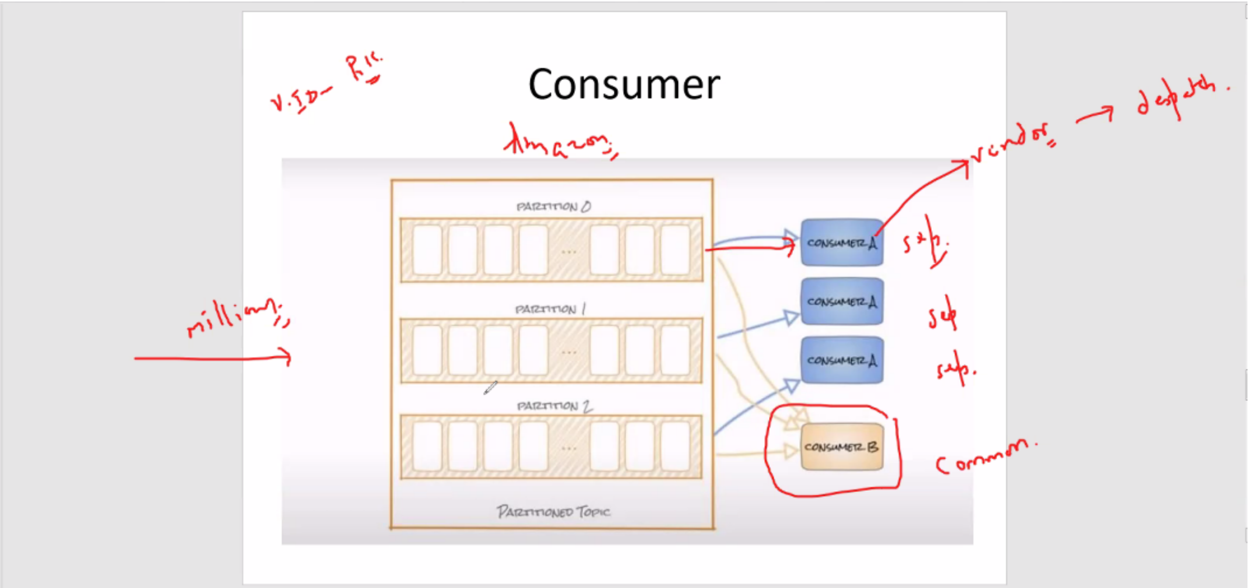
Consumer





Partitions for one Topic

= 44 messages in my topic?



Consumer

→ Oozie

  a.

→

→

→