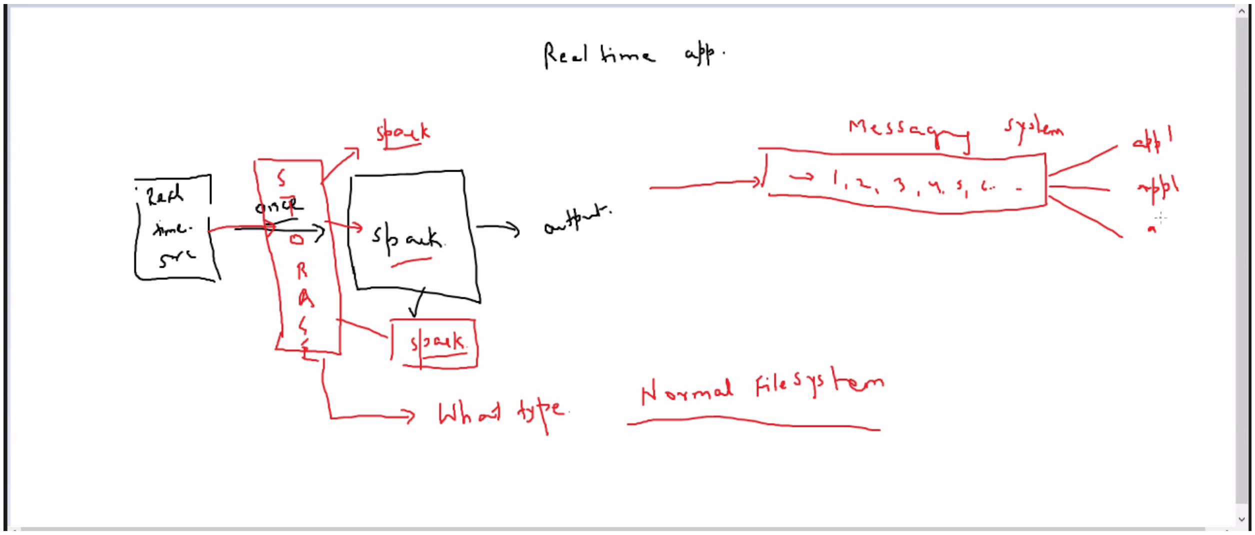
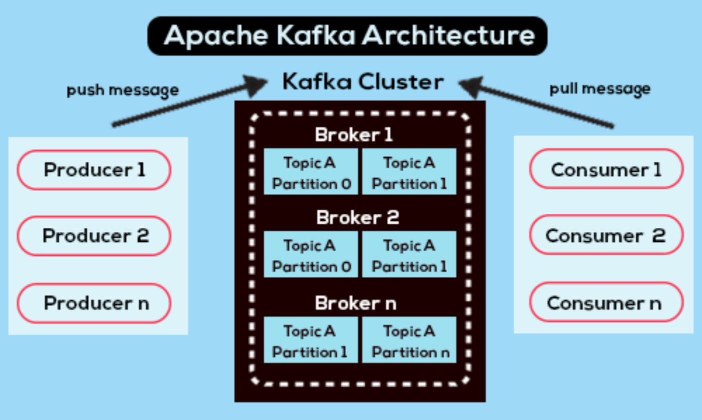
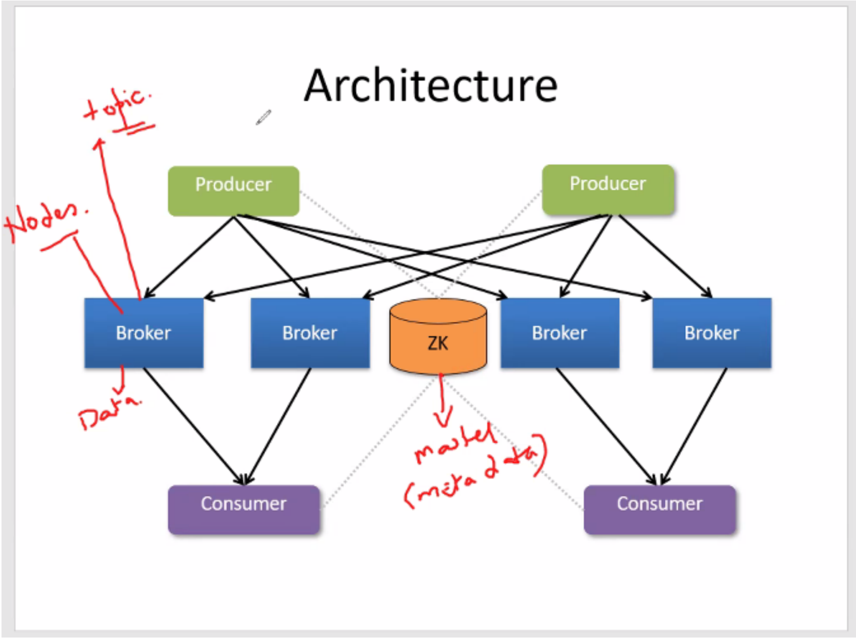
→ Kafka

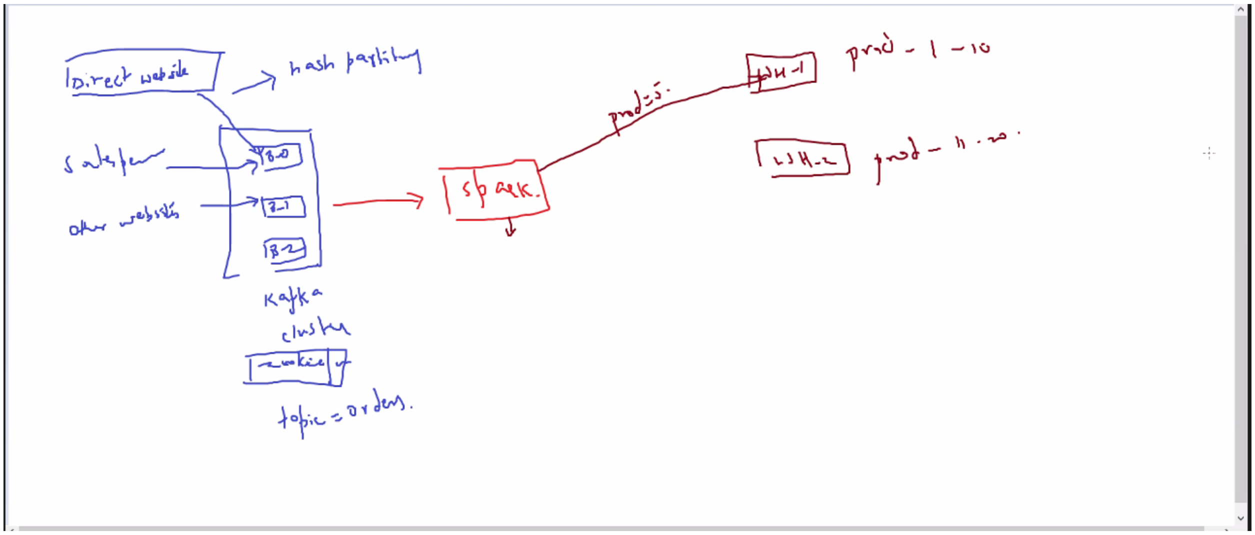


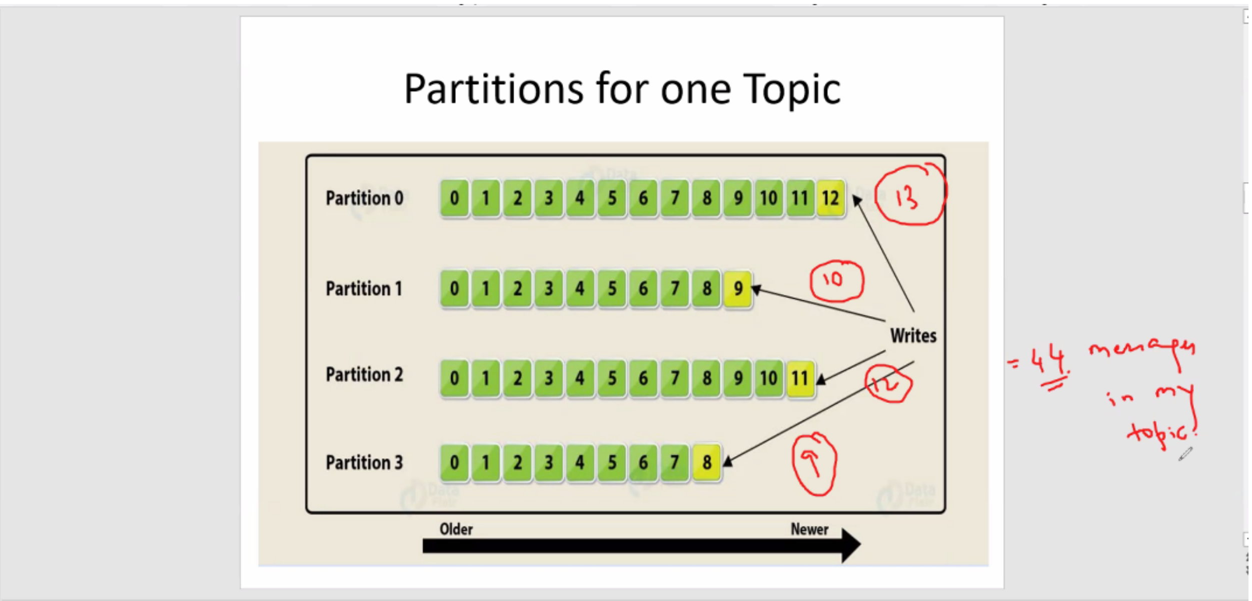
a. Kafka cluster, called cluster because it has multiple brokers
b. Using kafka, we collect real-time data from different publishers
c. distributed publish-subscribe messaging system
d. Publisher is sender and Consumer/Subscriber is receiver
e. Topic:
    i. an object where data is stored representing similar types of data
    ii. Can have many K, V paired messages
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. A kafka server is called a Broker, a bridge between producers & consumers
i. Kafka does not process data, it's just a storage layer
j. Zookeeper is used to store information about Kafka Cluster & details of consumer clients, manages brokers, mandatory to run zookeeper server
k. 

l.

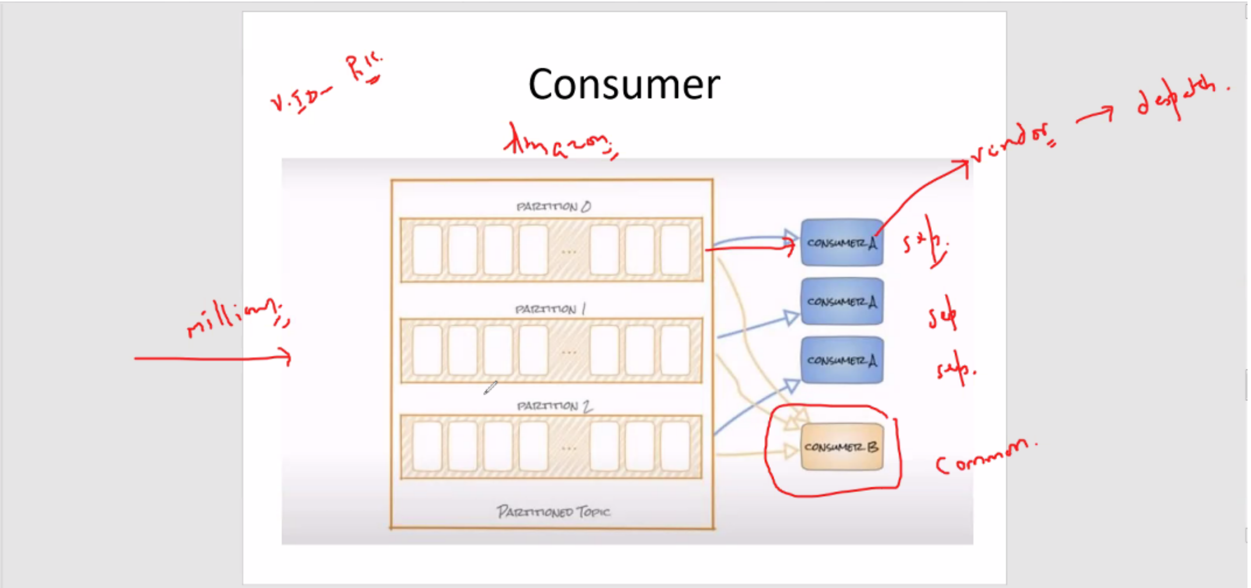


Architecture





Partitions for one Topic



Consumer

→ Oozie

a. Java Web App to schedule hadoop jobs
b. Combines multiple jobs sequentially into one logical unit of work
c. Three types of oozie jobs
    i. Oozie Workflow
        1. Directed Acyclic Graphs, specifying a sequence of actions to execute
    ii. Oozie Coordinator
        1. Jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability
    iii. Oozie Bundle

1. Provides a way to package multiple coordinator and workflow jobs and to manage the lifecycle of those jobs