Module 9

Introduction to Kafka

Thanachart Numnonda, Executive Director, IMC Institute

Thanisa Numnonda, Faculty of Information Technology,

King Mongkut’s Institute of Technology Ladkrabang

Introduction

Open-source message broker project

An open-source message broker project developed by the Apache Software Foundation written in Scala. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. It is, in its essence, a "massively scalable pub/sub message queue architected as a distributed transaction log", making it highly valuable for enterprise infrastructures.

What is Kafka?

- An apache project initially developed at LinkedIn
- Distributed publish-subscribe messaging system
- Designed for processing of real time activity stream data
  e.g. logs, metrics collections
- Written in Scala
- Does not follow JMS Standards, neither uses JMS APIs

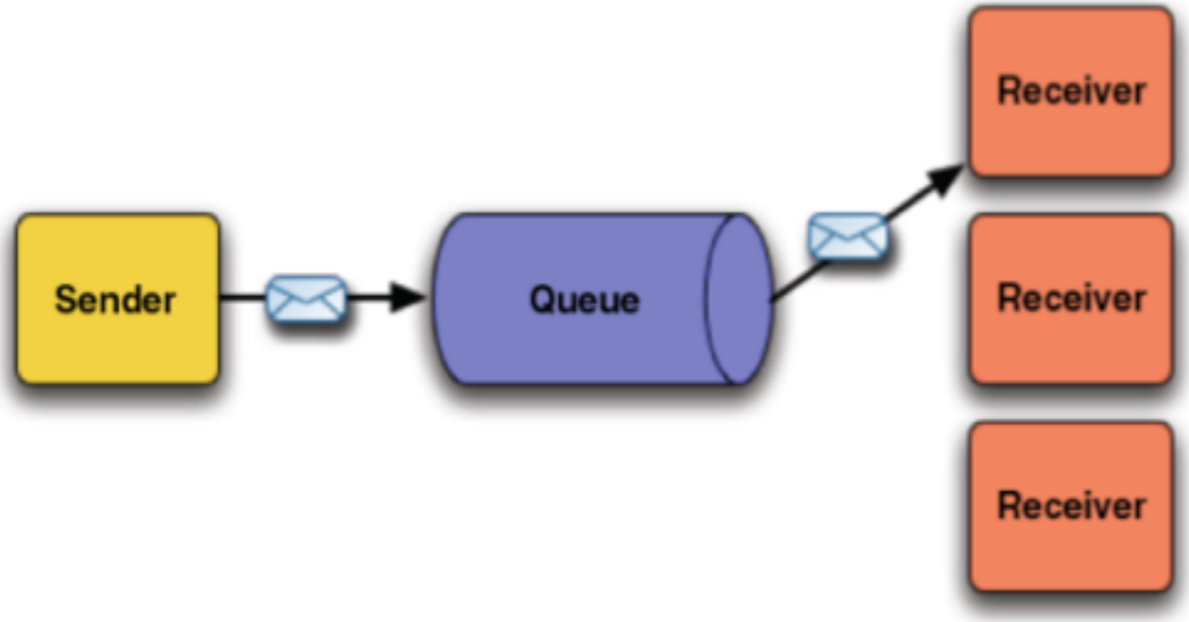
Kafka: Features

- Persistent messaging
- High-throughput
- Supports both queue and topic semantics
- Uses Zookeeper for forming a cluster of nodes (producer/consumer/broker) and many more…

Why Kafka?

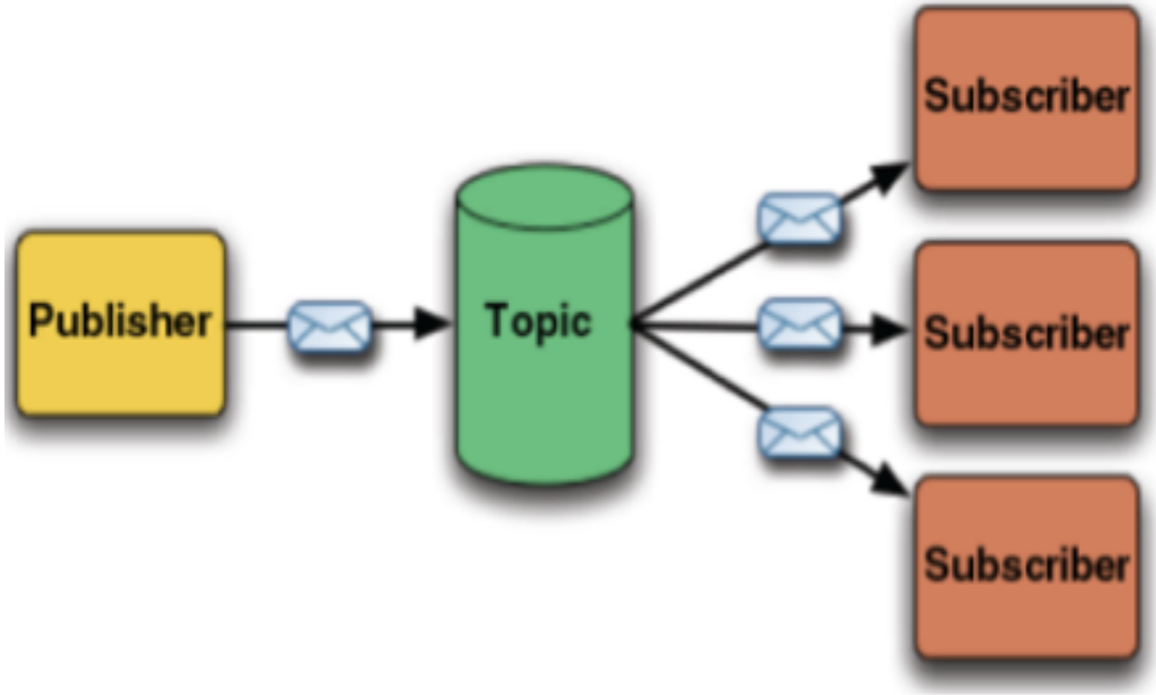
- Built with speed and scalability in mind.
- Enabled near real-time access to any data source
- Empowered hadoop jobs
- Allowed us to build real-time analytics
- Vastly improved our site monitoring and alerting capability
- Enabled us to visualize and track our call graphs.

Messaging System Concept: Queue



Source: Real time Analytics with Apache Kafka and Spark, Rahul Jain

Messaging System Concept: Topic

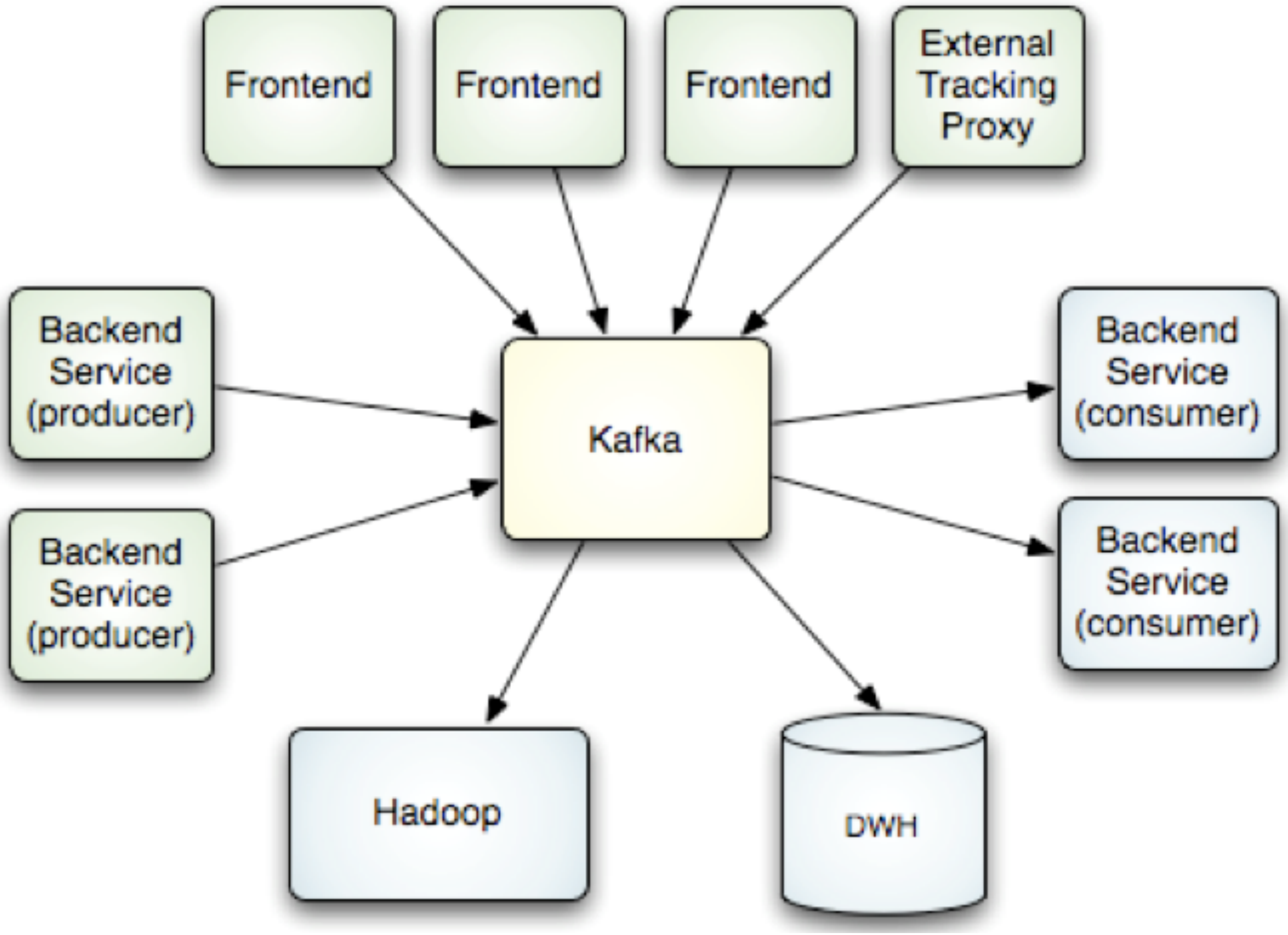


Source: Real time Analytics with Apache Kafka and Spark, Rahul Jain

Terminology

- Kafka maintains feeds of messages in categories called topics.
- Processes that publish messages to a Kafka topic are called producers.
- Processes that subscribe to topics and process the feed of published messages are called consumers.
- Kafka is run as a cluster comprised of one or more servers each of which is called a broker.

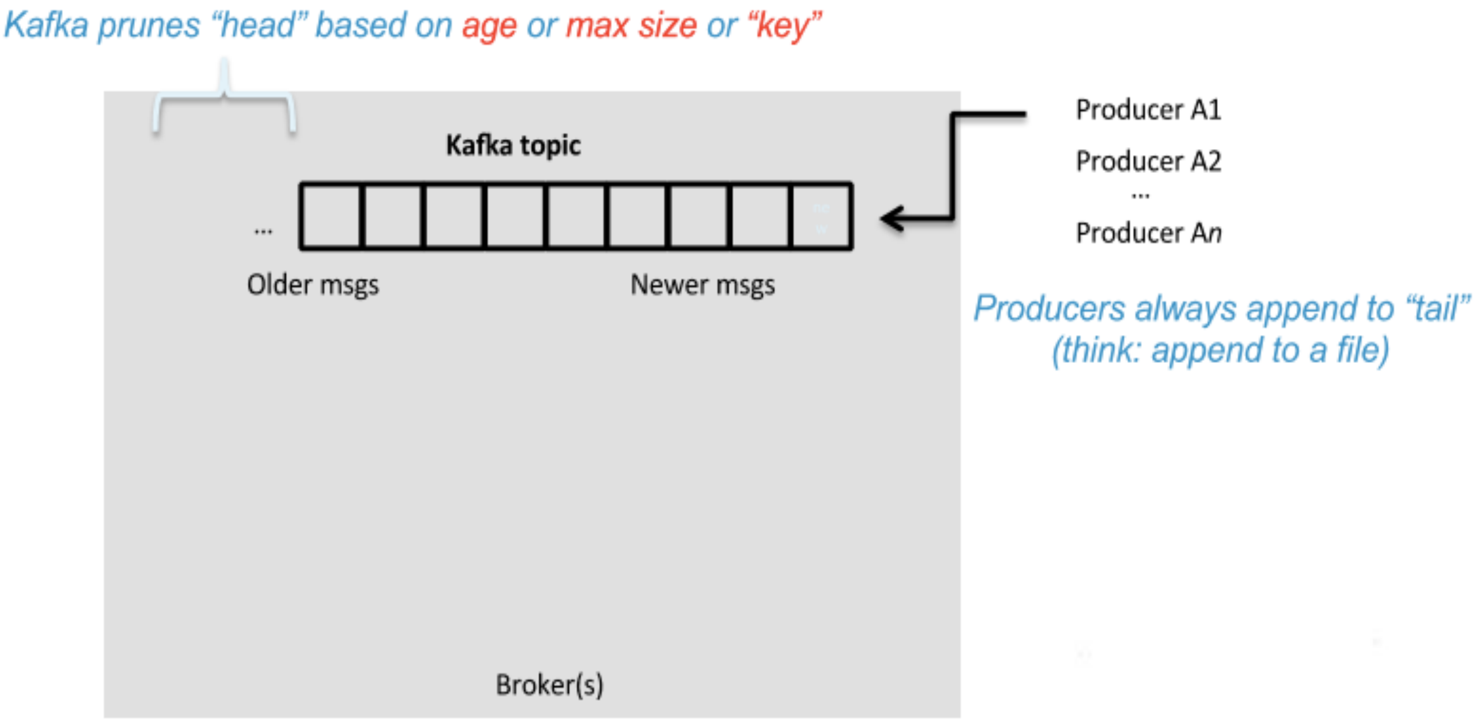
Kafka


Frontend
Frontend
Frontend
External Tracking Proxy
Backend Service (producer)
Backend Service (producer)
Kafka
Backend Service (consumer)
Backend Service (consumer)
Hadoop
DWH

Source: Real time Analytics with Apache Kafka and Spark, Rahul Jain

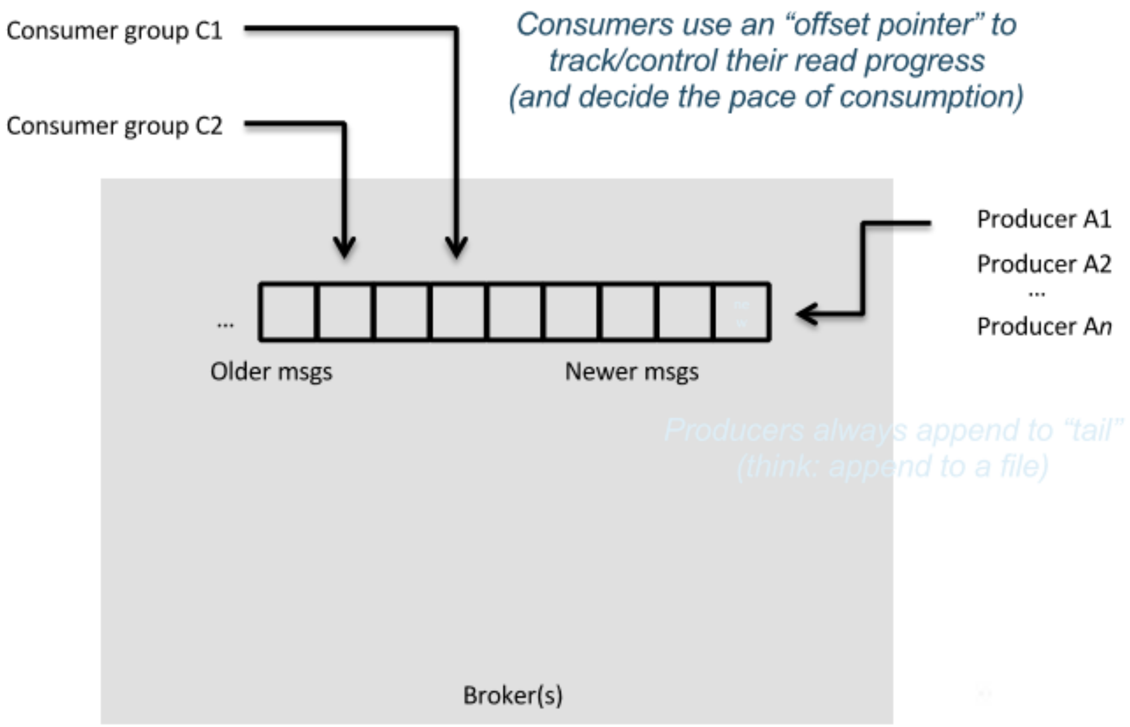
Topics

- Topic: feed name to which messages are published


Kafka prunes "head" based on age or max size or "key"
Kafka topic
...
Older msgs
Newer msgs
Broker(s)
Producer A1
Producer A2
...
Producer An
Producers always append to "tail"
(think: append to a file)

Source: Apache Kafka with Spark Streaming - Real Time Analytics Redefined

Topics



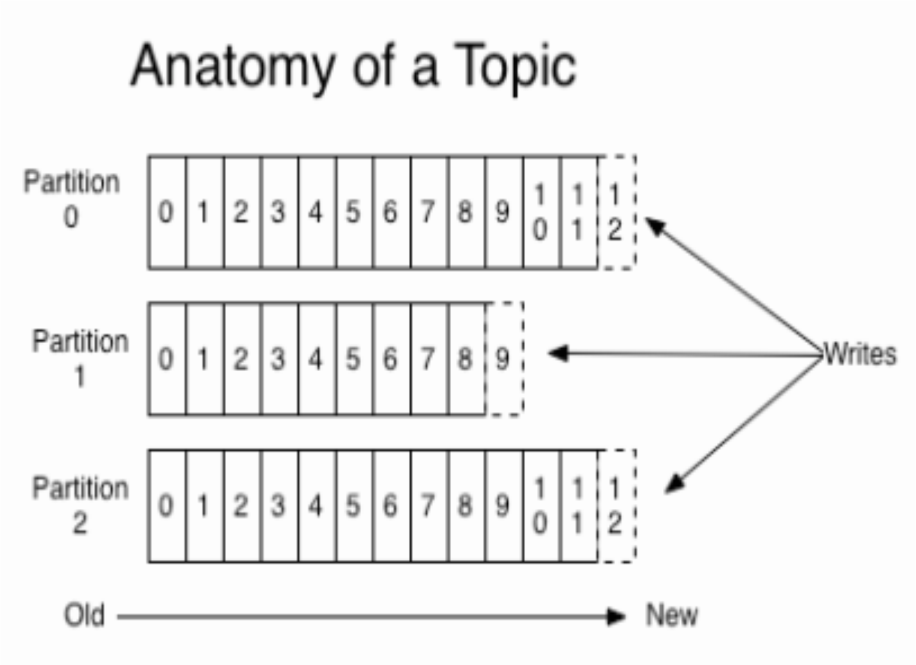
Consumers use an "offset pointer" to track/control their read progress (and decide the pace of consumption)

Producers always append to "tail" (think: append to a file)

Source: Apache Kafka with Spark Streaming - Real Time Analytics Redefined

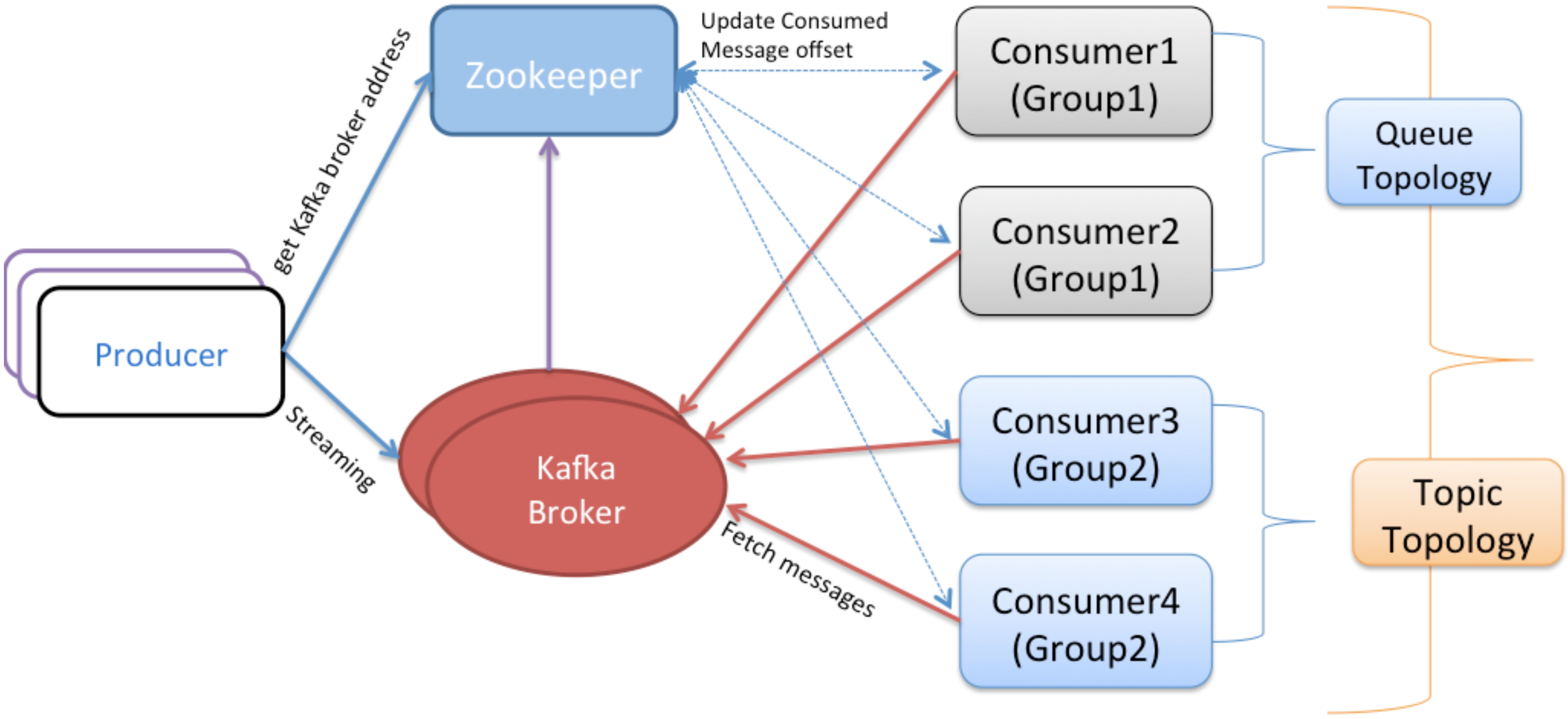
Topics

- A topic consists of partitions.
- Partition: ordered + immutable sequence of messages that is continually appended



Source: Apache Kafka with Spark Streaming - Real Time Analytics Redefined

Kafka Architecture



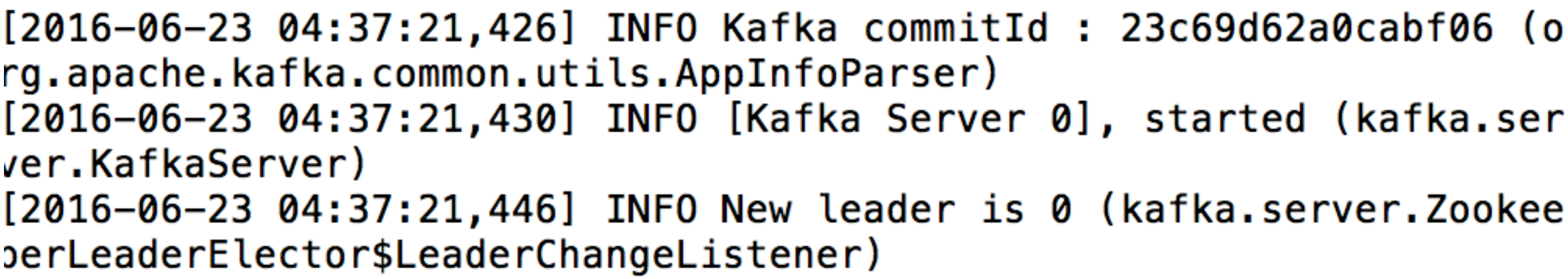
Source: Real time Analytics with Apache Kafka and Spark, Rahul Jain

Hands-on

SparkStreaming with Kafka

Install & Start Kafka Server

```
# wget http://www-us.apache.org/dist/kafka/0.9.0.1/kafka_2.10-
0.9.0.1.tgz
# tar xzf kafka_2.10-0.9.0.1.tgz
# cd kafka_2.10-0.9.0.1
# bin/kafka-server-start.sh config/server.properties&
```



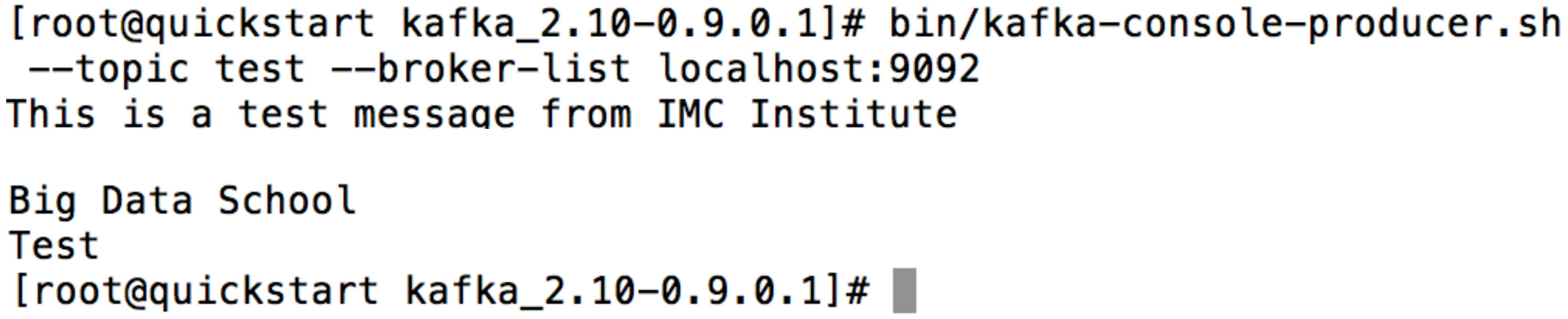
```
[2016-06-23 04:37:21,426] INFO Kafka commitId : 23c69d62a0cabf06 (o
rg.apache.kafka.common.utils.AppInfoParser)
[2016-06-23 04:37:21,430] INFO [Kafka Server 0], started (kafka.ser
ver.KafkaServer)
[2016-06-23 04:37:21,446] INFO New leader is 0 (kafka.server.Zookee
perLeaderElector$LeaderChangeListener)
```

Running Kafka Producer

Suggestion: Please open new terminal before run any following commands

```
# bin/kafka-console-producer.sh --topic test --broker-list
localhost:9092
```

type some random messages followed by Ctrl-D to finish

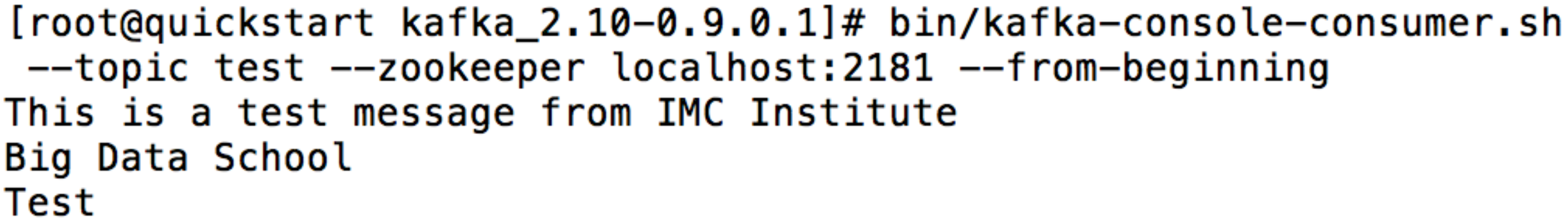


```
[root@quickstart kafka_2.10-0.9.0.1]# bin/kafka-console-producer.sh
 --topic test --broker-list localhost:9092
This is a test message from IMC Institute

Big Data School
Test
[root@quickstart kafka_2.10-0.9.0.1]# 
```

Running Kafka Consumer

```
# bin/kafka-console-consumer.sh --topic test --zookeeper
localhost:2181 --from-beginning
```



```
[root@quickstart kafka_2.10-0.9.0.1]# bin/kafka-console-consumer.sh
 --topic test --zookeeper localhost:2181 --from-beginning
This is a test message from IMC Institute
Big Data School
Test
```

Suggestion: Press Ctrl+c (ONLY 1 TIMES) to exit.