Using Apache Kafka for Integration and Data Processing Pipelines with Spring

JOSH LONG (/TEAM/JLONG) APRIL 15, 2015

30 COMMENTS (/BLOG/2015/04/15/USING-APACHE-KAFKA-FOR-INTEGRATION-AND-DATA-PROCESSING-PIPELINES-WITH-SPRING#DISQUS_THREAD)

Applications generated more and more data than ever before and a huge part of the challenge - before it can even be analyzed - is accommodating the load in the first place. Apache's Kafka (http://kafka.apache.org) meets this challenge. It was originally designed by LinkedIn and subsequently open-sourced in 2011. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds. The design is heavily influenced by transaction logs. It is a messaging system, similar to traditional messaging systems like RabbitMQ, ActiveMQ, MQSeries, but it's ideal for log aggregation, persistent messaging, fast (_hundreds_ of megabytes per second!) reads and writes, and can accommodate numerous clients. Naturally, this makes it perfect for cloud-scale architectures!

Kafka powers many large production systems (https://cwiki.apache.org/confluence/display/KAFKA/Powered+By). LinkedIn uses it for activity data and operational metrics to power the LinkedIn news feed, and LinkedIn Today, as well as offline analytics going into Hadoop. Twitter uses it as part of their stream-processing infrastructure. Kafka powers online-to-online and online-to-offline messaging at Foursquare. It is used to integrate Foursquare monitoring and production systems with Hadoop-based offline infrastructures. Square uses Kafka as a bus to move all system events through Square's various data centers. This includes metrics, logs, custom events, and so on. On the consumer side, it outputs into Splunk, Graphite, or Esper-like real-time alerting. Netflix uses it for 300-600BN messages per day. It's also used by Airbnb, Mozilla, Goldman Sachs, Tumblr, Yahoo, PayPal, Coursera, Urban Airship, Hotels.com, and a seemingly endless list of other big-web stars. Clearly, it's earning its keep in some powerful systems!

Installing Apache Kafka

There are many different ways to get Apache Kafka installed. If you're on OSX, and you're using Homebrew, it can be as simple as `brew install kafka`. You can also download the latest distribution from Apache (http://kafka.apache.org/downloads.html). I downloaded `kafka_2.10-0.8.2.1.tgz`, unzipped it, and then within you'll find there's a distribution of