About Kafka

About Kafka

Kafka is a distributed publish-subscribe messaging system.

It is designed to support the following:

- Persistent messaging with O(1) disk structures that provide constant time performance even with many TB of stored messages.
- High-throughput: even with very modest hardware Kafka can support hundreds of thousands of messages per second.
- Explicit support for partitioning messages over Kafka servers and distributing consumption over a cluster of consumer machines while maintaining per-partition ordering semantics.
- Support for parallel data load into Hadoop.

Kafka provides a publish-subscribe solution that can handle all activity stream data and processing on a consumer-scale web site. This kind of activity (page views, searches, and other user actions) are a key ingredient in many of the social feature on the modern web. This data is typically handled by "logging" and ad hoc log aggregation solutions due to the throughput requirements. This kind of ad hoc solution is a viable solution to providing logging data to an offline analysis system like Hadoop, but is very limiting for building real-time processing. Kafka aims to unify offline and online processing by providing a mechanism for parallel load into Hadoop as well as the ability to partition real-time consumption over a cluster of machines.

The use for activity stream processing makes Kafka comparable to Facebook's Scribe or Cloudera's Flume, though the architecture and primitives are very different for these systems and make Kafka more comparable to a traditional messaging system.