Kafka: A Distributed Messaging System for Log Processing

Kartikeya Upasani (kuu2101)

Paper Review, 16 October 2016

1 Motivation

Today's business needs require not only offline processing of log data for analysis, but also real-time processing for applications such as making recommendations or doing targeted advertising.

2 Goal

Developing a simple API that can store large volumes of log data while enabling efficient real-time consumption by applications.

3 Key Idea

Storing logs in a distributed fashion and making optimizations at several levels of the system for performance.

4 Approach

All logs are categorized into topics and stored in a distributed fashion on nodes called 'brokers'. Producers write to these topics, while consumers can create multiple streams that subscribe to these topics and receive messages in an equitably across streams. To prevent unnecessary seeks, the logical offset of the log in file system is used as the ID instead of maintaining separate IDs. An important optimization is to make the consumption process stateless. The broker does not maintain the state of the consumer; it deletes messages after an expiry period has elapsed. The architecture relies consumers coordinating among themselves in a decentralized way rather than having a master server. The Zookeeper service is used for this purpose.

5 Results

The production as well as consumption steps consistently show linear performance scaling with increase in data size, even upto terabytes.

6 Conclusion

At the time of writing of the paper, Kafka had been in production use for 6 months at LinkedIn. This talks about the success of the design process and validates the compromises made during it. Kafka is also an example of how a specialized system can lead to high performance gains for certain use cases.

7 Comments

Messages written by producers are flushed to the queue after an interval of time. This trades availability of message for efficiency. Moreover, the design choices offer several side benefits such as rewinding of messages that can be leveraged by consumers.