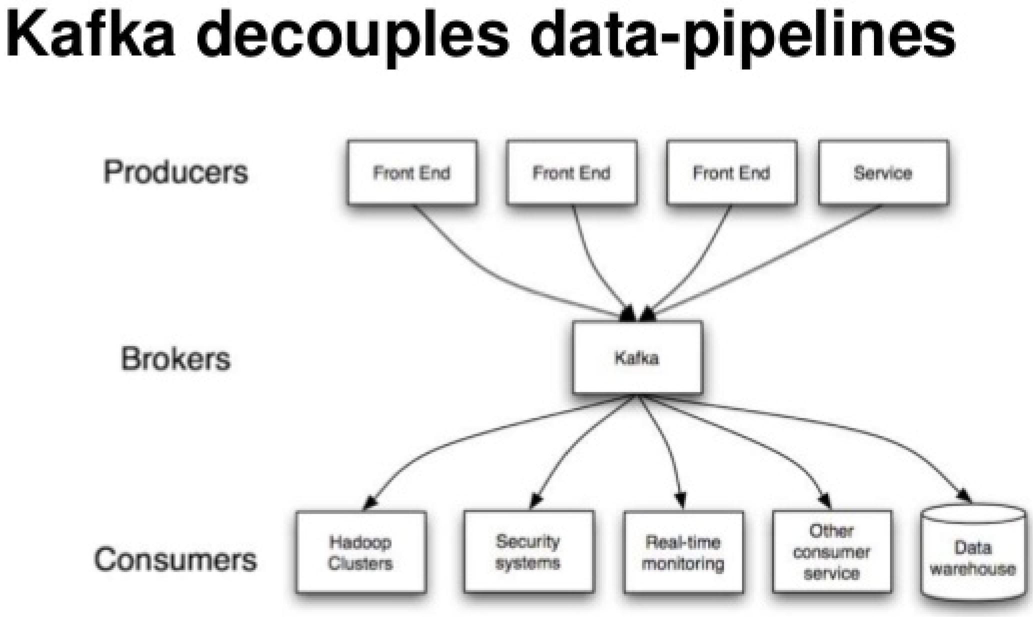
LAKSHAYSURI

Apache Kafka – Real time data processing

Posted on October 2, 2016October 8, 2016 by lakshaysuri



Kafka is a messaging, storage and stream processing system developed by Apache. It enables us to produce and consume a set/stream of records. It works well for real time stream of data which when sent from producer could be analysed at the consumer end.

How it works?

a) Kafka runs as a cluster on one or more servers.

b) This Kafka cluster stores set of records in topics.

c) Each record consist of a key, value and a timestamp.

Types of API present in kafka ?

Producer API: allows an application to publish stream of records to a particular topic.

Consumer API: allows an application to consume the stream of records from the topic where the producer has sent the data.

Connector API: allows to build producers or consumers that could connect topics easily to the existing systems such as database etc.

Streams API: allows an application to consume stream of records from one or more topics , do some processing on this input data and output it to one or more topics.

Communication in Kafka ?

Communication between client and server is done with a simple, high performance TCP protocol.

Usage of Kafka ?

Kafka as a messaging system

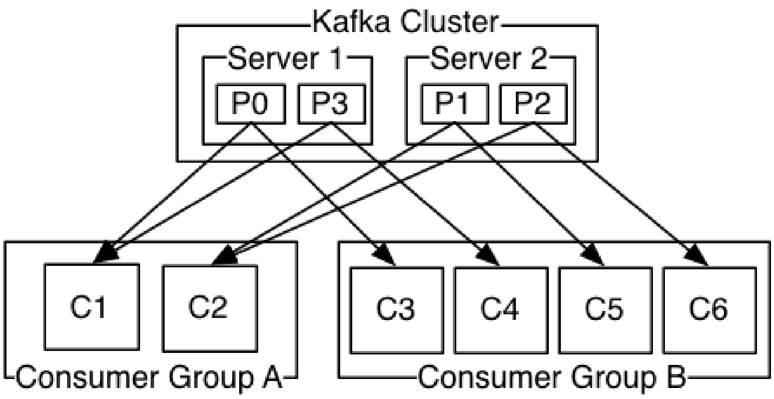


Image Source: Kafka Website: http://kafka.apache.org/intro

a) Messages are published to topic(s). There exist a partition for each topic and records in the partition are identified using a unique number called the offset. These records are maintained in the partition for a configurable amount of time. Each partition has one server which act as a leader and the rest of the servers act as followers.

b) Consumer label themselves within a consumer group. A single consumer group can have multiple consumers and there could be several consumer groups. Do note that each record published to a topic is delivered to only one consumer instance in a consumer group.

c) Kafka has an advantage of traditional messaging system which is: that it allows for order of message guarantee. This is obtained by assigning each partition(say P0) to exactly one consumer (C1) in the consumer group (A). In other words, P0 can not be consumer by consumer c2 in consumer group A. This would ensure that consumer C1 is the only consumer (in that group) of messages arriving from partition P0 and hence would consume the data in order. Also, note that, kafka provides ordered records only per partition and not between different partitions in a topic. Traditional messaging has a flaw that the messages sent out from the topic to the consumer are asynchronous and hence may arrive out of order on different consumers.

Kafka as a storage system

a) A message queue is essentially a storage system but however in kafka data is also written to the disk and replicated for fault tolerance. Kafka allows producers to wait for acknowledgement and a write is not complete until is replicated.

b) Kafka can be considered as a distributed filesystem with low latency log storage and replication.

Kafka for stream processing

a) In kafka, a stream processor would take continuous input from the topic, perform some processing or analysis and produce continuous output stream of data to output topics.

b) For instance, an application sending input data (sales, quantity and price per quantity) in json format and then this data is processed to form another set of data which would indicate the new price per quantity and sent to output topics.

Posted in Uncategorized Leave a comment

This site uses Akismet to reduce spam. Learn how your comment data is processed.

Create a free website or blog at WordPress.com.