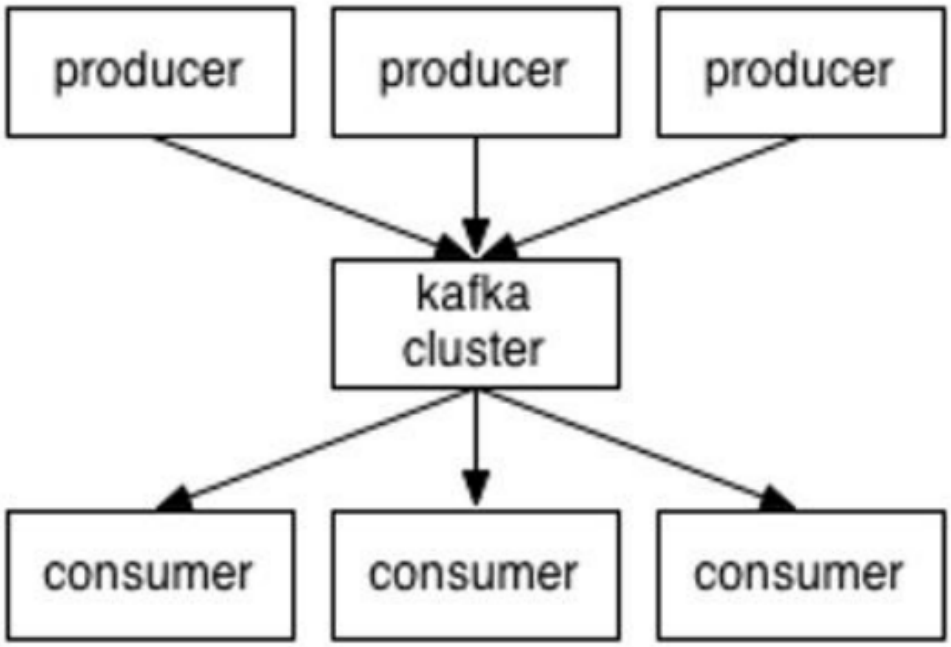
Kafka2 基础架构

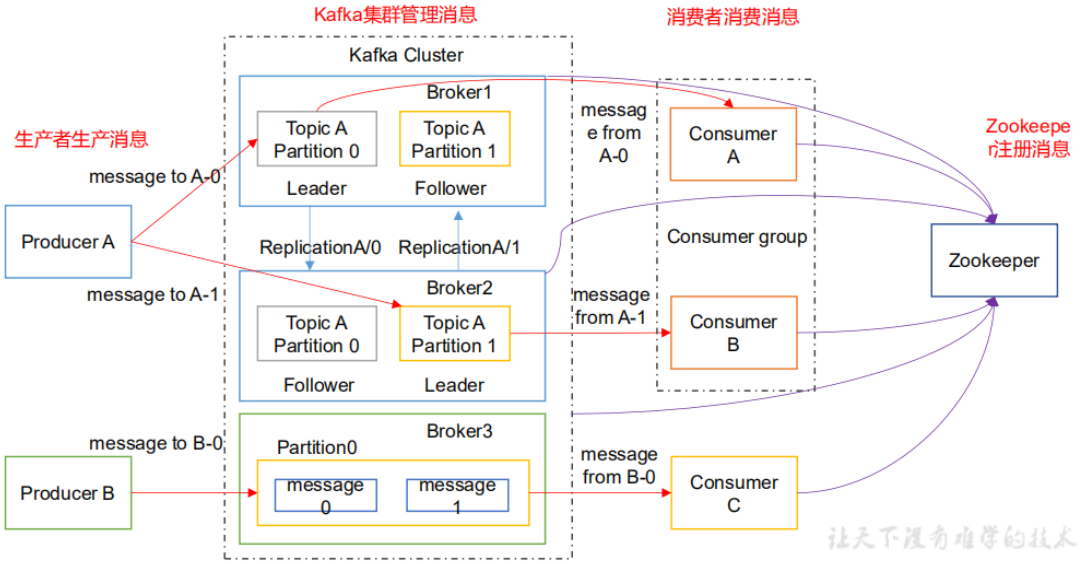
什么是Kafka

在流式计算中，Kafka 一般用来缓存数据，Storm 通过消费 Kafka 的数据进行计算。**Kafka 是一个分布式消息队列。**Kafka 对消息保存时根据 Topic 进行归类，发送消息者称为 Producer，消息接受者称为 Consumer，此外 kafka 集群有多个 kafka 实例组成，每个实例(server)称为 broker。

整体架构



详细架构



- Producer

> 消费生产者，就是向 Kafka broker 发消息的客户端

- Consumer

> 消息消费者，向 Kafka broker 取消息的客户端

- Topic

> 可以理解为一个队列

- Consumer Group（CG）

> 这是Kafka用来实现一个topic消息的广播（发给所有的consumer）和单播（发给任意一个consumer）的手段。一个topic可以有多个CG。topic的消息会复制（并不是真的复制，只是概念上的）到所有的CG，但每个partion只会把消息发给该CG中的一个consumer。如果需要实现广播，只要每个consumer有一个独立的CG就赢了。要实现单播只要所有的 consumer 在同一个 CG。用 CG 还可以将 consumer 进行自由的分组而不需要多次发送消息到不同的 topic；

- broker

> 一台 kafka 服务器就是一个 broker。一个集群由多个 broker 组成。一个 broker可以容纳多个 topic；

- Partition

> 为了实现扩展性，一个非常大的 topic 可以分布到多个 broker（即服务器）上，一个 topic 可以分为多个 partition，每个 partition 是一个有序的队列。partition 中的每条消息都会被分配一个有序的 id（offset）。kafka 只保证按一个 partition 中的顺序将消息发给consumer，不保证一个 topic 的整体（多个 partition 间）的顺序；

- Offset

kafka 的存储文件都是按照 offset.kafka 来命名，用 offset 做名字的好处是方便查找。例如你想找位于 2049 的位置，只要找到 2048.kafka 的文件即可。当然 the first offset 就 是 00000000000.kafka。