Kafka是一个分布式的流媒体平台，因为能应用到的比较多，比如消息队列，日志收集

1.**高吞吐量**：能处理TB级的数据

消息持久化，将消息传入**硬盘**里，硬盘便宜，**空间大，处理海量数据**的前提（要先有海量数据）虽在硬盘中，但并不是说在硬盘的读取就一定慢了，**kafka的硬盘读取仍然很快**，他是**通过硬盘的顺序读取Segment 数据文件**，且消息消费后不会删除（删除策略是针对过期的 Segment 文件，而非随机读取，随机慢，但顺序读取甚至**能快过内存。**

2.**高扩展**：集群不够了，**加一台服务器**

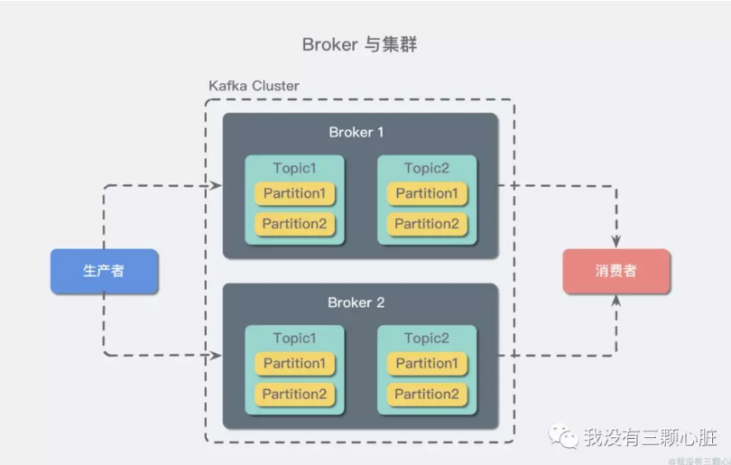
3.**高可靠**：通过**分布式**保证，可以**做集群部署**，一台服务器挂了还有别的服务器

**高可用：副本机制**。每个partition的数据都会同步到其他机器上，形成自己的副本。然后所有的副本会选一个leader出来，那么生产者消费者都和这个leader打交道，其他的副本就是follower。写的时候，leader会把数据同步到所有follower上面去，读的时候直接从leader上面读取即可

如果某个broker宕机了，那个broker的partition在其他机器上有副本，如果这上面有某个partition的leader，那么此时会重新选举出一个现代leader出来，继续读写这个新的leader即可

**写消息：**写数据的时候，生产者就写给leader，然后leader将数据落到磁盘上之后，然后其他follower自己从leader来pull数据。所有follower同步好数据就会发ack确认给leader，leader收到所有的follower的ack之后，就会返回 写成功 的消息给消息生产者。（这只是一种模式，可以调整）。

**读(消费)数据**:只会从leader进行消费。但是只有一个消息已经被所有follower都同步成功返回ack的时候，这个消息才会被消费者读到



**Broker**：消息中间件处理节点，一个Kafka服务器，把生产出的消息放到磁盘然后会让消费者pull(因为push目标是尽可能快地传输数据，会和消费者消费速度不适应造成来不及消费--网络阻塞，消费者pull允许消费者根据自己消费数据能力用适当的速率消费数据)。

**Topic**(只是逻辑上的一个东西)：主题，一类消息(代表发消息的位置和消息类别，比如点赞、关注的主题)，消息队列MQ的实现方式有两种：

1. 一种是**点对点**，**阻塞队列**就是点对点，每个数据被一个消费者消费；
2. **发布订阅模式**：生产者将消息放到每一个位置，可以有**很多消费者同时订阅消息**，这个消息可以被多个消费者读到（同时或先后）

**kafka**用的第二种发布订阅消息，**生产者将消息发送到的位置就是topic**

**Partition**：topic的物理上的分区，一个topic分多个partition，每个partition是一个有序队列。可以采用**多线程**的方式，同时向多个分区写数据，增强并发能力，每个分区按从前往后写入数据。比如加了A再加B，就会先消费A --> 这样也保证了数据的可靠

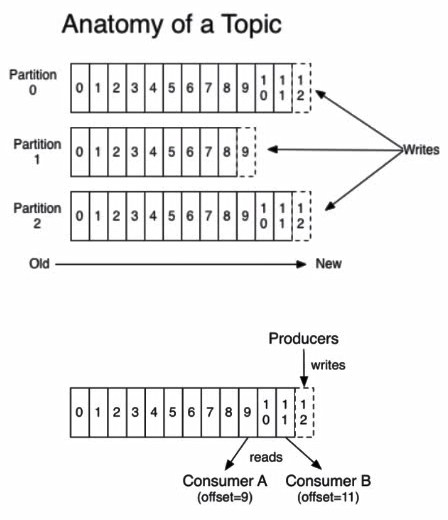
序号从0开始，序号最大值为partitions数量减1。

**Segment：**partition物理上由多个segment组成，每个partition(目录)相当于一个巨型文件被平均分配到多个大小相等segment(段)数据文件中。但每个段segment file消息数量不一定相等，这种特性方便old segment file快速被删除。默认保留7天的数据。

**Segment File = index file(索引文件".index") + data file(数据文件".log" )总是成对出现**

索引文件用**稀疏索引**存储的方式，每隔一定字节的数据建立一条索引，它减少了索引文件大小，使得能够把 index 映射到内存，降低了查询时的磁盘 IO 开销，同时也并没有给查询带来太多的时间消耗，所以index 文件并不是从0开始，也不是每次递增1的

**Offset**：**消息在分区内的索引，消息的序列号，用于分区里标识消息**



**Cluster**(集群)：多个Broker组成，某个Broker会被选为集群控制器来管理集群，监控故障等。一个分区也可以被复制到多个Broker，这样一个Broker有故障可以用其他Broker

**多集群：**用MirrorMaker(一个消费者和生产者)工具实现集群直接消息生产和消费。

好处：1.基于数据和安全的隔离 2.多数据中心(容灾)

**Zookeeper**(独立引用)：用于管理集群,比如选出分区的leader故障了,选个broker作为控制器

**Retention(日志保留)：**可以对某个主题单独设置日志保留时间/日志大小，老消息就被删除

**Replica副本：**因为**kafka是分布式**消息引擎，**将数据存放多份 -> 数据更可靠**

Leader Replica(主副本): 从分区**读数据**时，主副本可以做出响应

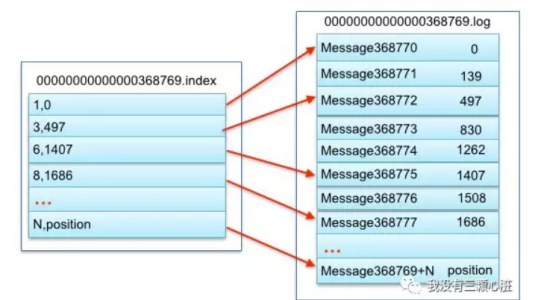
Follower Replica(从副本)只用来**存数据**，主副本挂掉时候集群会从副本中选一个作为leader

## **Kafka 中的底层存储设计**

假设我们现在 Kafka 集群只有一个 Broker，我们创建 2 个 Topic 名称分别为：「topic1」和「topic2」，Partition 数量分别为 1、2，那么我们的根目录下就会创建如下三个文件夹：

    | --topic1-0  
    | --topic2-0  
    | --topic2-1

现在假设我们设置每个 Segment 大小为 500 MB，并启动生产者向 topic1 中写入大量数据，topic1-0 文件夹中就会产生类似如下的一些文件：



索引文件中元数据 <3, 497>  = 数据文件中表示第 3 个 message（在全局 Partition 表示第 368769 + 3 = 368772 个 message）+ 该消息的物理偏移地址为 497。

**文件名 = 上一个 Segment 最后一条消息的 offset**

**找一个指定offset的message:**通过在所有 segment 的文件名中进行二分查找找到它归属的 segment，再在他 index索引文件种找到它对应到文件上的物理位置，就能拿出这个message

**Kafka如何知道 message的偏移**：因为在Kafka定义了标准的数据存储结构，在 Partition中的每一条message都包含了三个属性：

**1.offset：**表示 message 在当前 Partition 中的偏移量，是一个逻辑上的值，唯一确定了Partition 中的一条 message，可以简单的认为是一个**id**；

**2.MessageSize**：表示 message 内容 data 的大小；

**3.data：**message 的具体内容

## **Kafka 存储在文件系统上 --- 消息存储策略**

现代的操作系统针对磁盘的读写已经做了一些优化方案来加快磁盘的访问速度。比如，**预读**会提前将一个比较大的磁盘快读入内存。**后写**会将很多小的逻辑写操作合并起来组合成一个大的物理写操作。并且，操作系统还会将主内存剩余的所有空闲内存空间都用作**磁盘缓存**，所有的磁盘读写操作都会经过统一的磁盘缓存（除了直接 I/O 会绕过磁盘缓存）。综合这几点优化特点，**如果是针对磁盘的顺序访问，某些情况下它可能比随机的内存访问都要快，甚至可以和网络的速度相差无几。**

**上述的 Topic 其实是逻辑上的概念，面相消费者和生产者，物理上存储的其实是 Partition**，每一个 Partition 最终对应一个目录，里面存储所有的消息和索引文件。默认情况下，每一个Topic 在创建时如果不指定 Partition 数量时只会创建 1 个 Partition。比如，我创建了一个 Topic 名字为 test ，没有指定 Partition 的数量，那么会默认创建一个 test-0 的文件夹，这里的命名规则是：<topic\_name>-<partition\_id>

任何发布到 Partition 的消息都会被追加到 **Partition 数据文件的尾部**，这样的顺序写磁盘操作让 Kafka 的**效率非常高**（经验证，顺序写磁盘效率比随机写内存还要高，这是 Kafka 高吞吐率的一个很重要的保证）。

每一条消息被发送到 Broker 中，会根据 Partition 规则选择被存储到哪一个 Partition。如果 Partition 规则设置的合理，所有消息可以均匀分布到不同的 Partition中。

## **生产者设计概要**

每条消息都是很关键且不能容忍丢失么？偶尔重复消息可以么？我们关注的是消息延迟还是写入消息的吞吐量？

举个例子，有一个**信用卡交易处理系统**，当交易发生时会发送一条消息到 Kafka，另一个服务来读取消息并根据规则引擎来检查交易是否通过，将结果通过 Kafka 返回。对于这样的业务，**消息既不能丢失也不能重复**，由于交易量大因此吞吐量需要尽可能大，延迟可以稍微高一点。

再举个例子，假如我们需要收集用户在**网页上的点击数据**，对于这样的场景，**少量消息丢失或者重复是可以容忍的**，延迟多大都不重要只要不影响用户体验，吞吐则根据实时用户数来决定。

不同的业务需要使用不同的写入方式和配置。

**生产者写消息流程：**

1.创建一个ProducerRecord，这个对象需要包含消息的主题（topic）和值（value），可以选择性指定一个键值（key）或者分区（partition）。

2.发送消息时，生产者会对键值和值序列化成字节数组，然后发送到分配器（partitioner）。

3.如果我们指定了分区，那么分配器返回该分区即可；否则，分配器将会基于键值来选择一个分区并返回。

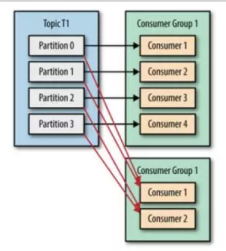
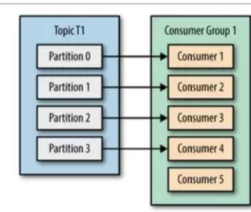
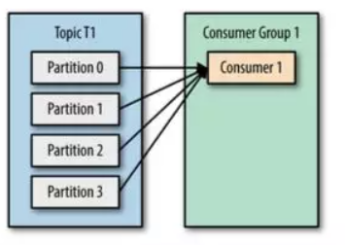
4.选择完分区后，生产者知道了消息所属的主题和分区，它将这条记录添加到相同主题和分区的批量消息中，另一个线程负责发送这些批量消息到对应的Kafka broker。

5.当broker接收到消息后，如果成功写入则返回一个包含消息的主题、分区及位移的RecordMetadata对象，否则返回异常。

6.生产者接收到结果后，对于异常可能会进行重试。

## **消费者设计概要**

一个消费者不够就在消费者组里加消费者



第三个图：Kafka很重要的特性：只要写入一次消息，可以支持任意多的应用读取这个消息

所以如果应用要读全量消息，就设置个消费组，如果消费能力不足就加消费者

**重平衡(rebalance)：**消费者离开后他负责的分区会分给其他分区

期间所有消费者都不能消费消息，所以会造成整个消费组不可用

消费者通过定期发送**心跳**（hearbeat）到一个作为组协调者（group coordinator）的 broker 来保持在消费组内存活

如果消费者超过一定时间没有发送心跳，那么它的会话（session）就会过期，组协调者会认为该消费者已经宕机，然后触发重平衡。可以看到，从消费者宕机到会话过期是有一定时间的，这段时间内该消费者的分区都不能进行消息消费；通常情况下，我们可以进行优雅关闭，这样消费者会发送离开的消息到组协调者，这样组协调者可以立即进行重平衡而不需要等待会话过期。

在 0.10.1 版本，Kafka 对心跳机制进行了修改，将发送心跳与拉取消息进行分离，这样使得发送心跳的频率不受拉取的频率影响。另外更高版本的 Kafka 支持配置一个消费者多长时间不拉取消息但仍然保持存活，这个配置可以避免活锁（livelock）。活锁，是指应用没有故障但是由于某些原因不能进一步消费。

## **Kafka 如何保证可靠性**

关系型数据库的可靠性保证是ACID，也就是原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）和持久性（Durability）。

Kafka 中的**可靠性**保证有如下四点：

1.对于一个分区来说，它的消息是有序的。如果一个生产者向一个分区先写入消息A，然后写入消息B，那么消费者会先读取消息A再读取消息B。

2.当消息写入所有in-sync状态的副本后，消息才会认为已提交（committed）。这里的写入有可能只是写入到文件系统的缓存，不一定刷新到磁盘。生产者可以等待不同时机的确认，比如等待分区主副本写入即返回，后者等待所有in-sync状态副本写入才返回。

3.一旦消息已提交，那么只要有一个副本存活，数据不会丢失。

4.消费者只能读取到已提交的消息。

面试题

**1. 为什么用Kafka不用别的**：数据传输块，吞吐量（每秒传的数据量）是百万级别的，其他的只有十万级或者万级。而且可靠性更高，比如有副本保证数据安全，有跳表思想设计log和index文件还有稀松索引来提升效率，零拷贝减少数据在不同环节来回复制，提升读写速度。但如果要保证高并发高性能高吞吐，那么Kafka有数据丢失风险，比如跟金额相关的要保证高可靠，丢失任何一个订单都不可以。

**2.Kafka快分为写入和读取**

**写入快**因为1.顺序写入 2.MMFile内存映射文件

1.顺序写入磁盘（因为每次读写都会寻址然后写入）读写速度大于内存随机读写，但不会删除数据，只会每个topic里每个消费者用offset表示读到第几条消息，解决是靠顺序写入基于时间和分区文件大小

2.顺序写入硬盘访问速度还是赶不上内存，所以用内存来非实时写入硬盘。MMF利用操作系统的page来实现文件到物理内存的直接映射，映射之后对物理内存的操作会被同步到硬盘上。但不可靠因为不会真正写到硬盘，要os的flush，但可以配置写入mmf就flush

**读取快**：1.基于sendfile实现零拷贝2.批量压缩

1.sendfile系统调用，文件数据被copy到内核中socket相关的缓冲区

再copy到协议引擎，比传统的减少了copy次数

2.把多个消息一起压缩

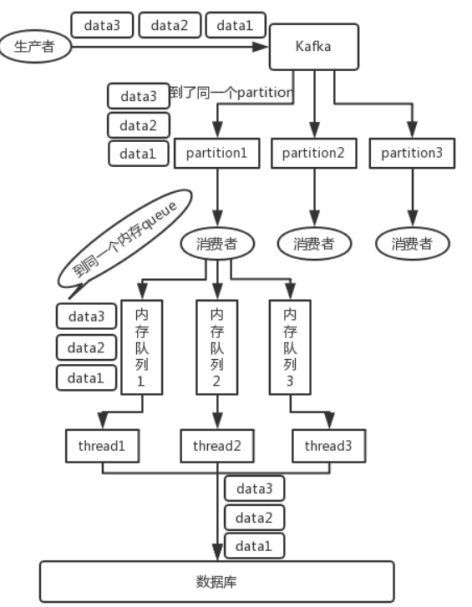
总结：它把所有的消息都变成一个批量的文件，并且进行合理的批量压缩，减少网络IO损耗，通过mmap提高I/O速度，写入数据的时候由于单个Partition是末尾添加所以速度最优；读取数据的时候配合sendfile直接暴力输出

1. **怎么保证顺序消费**：生产者在写的时候，其实可以指定一个 key，比如说我们指定了某个订单 id 作为 key，那么这个订单相关的数据，一定会被分发到同一个 partition 中去，而且这个 partition 中的数据一定是有顺序的。  
   消费者从 partition 中取出来数据的时候，也一定是有顺序的。到这里，顺序还是 ok 的，没有错乱。

**没顺序因为**：然后我们在消费者里可能会搞多个线程来并发处理消息。因为如果消费者是单线程消费处理，而处理比较耗时的话，比如处理一条消息耗时几十 ms，那么 1 秒钟只能处理几十条消息，这吞吐量太低了。而多个线程并发跑的话，顺序可能就乱掉了

**保证顺序消费：**1.一个 topic，一个 partition，一个 consumer，内部单线程消费，单线程吞吐量太低，一般不会用这个。

1. 写 N 个内存 queue，具有相同 key 的数据都到同一个内存 queue；然后对于 N 个线程，每个线程分别消费一个内存 queue 即可，这样就能保证顺序性



1. 如果一个发送失败了，后边的就不能继续发了，不然重发的那个肯定就乱序了
2. 生产者在收到发送成功的反馈之前，不能发下一条数据(但我感觉生产者是一个流，阻塞正产者感觉业务不可行，怎么会因为一条消息发出去没收到反馈，就阻塞生产者)

也就是用同步发送模式：发出消息后，必须阻塞等待收到通知后，才发送下一条消息

而异步发送模式：一直往缓冲区写，然后一把写到队列中去

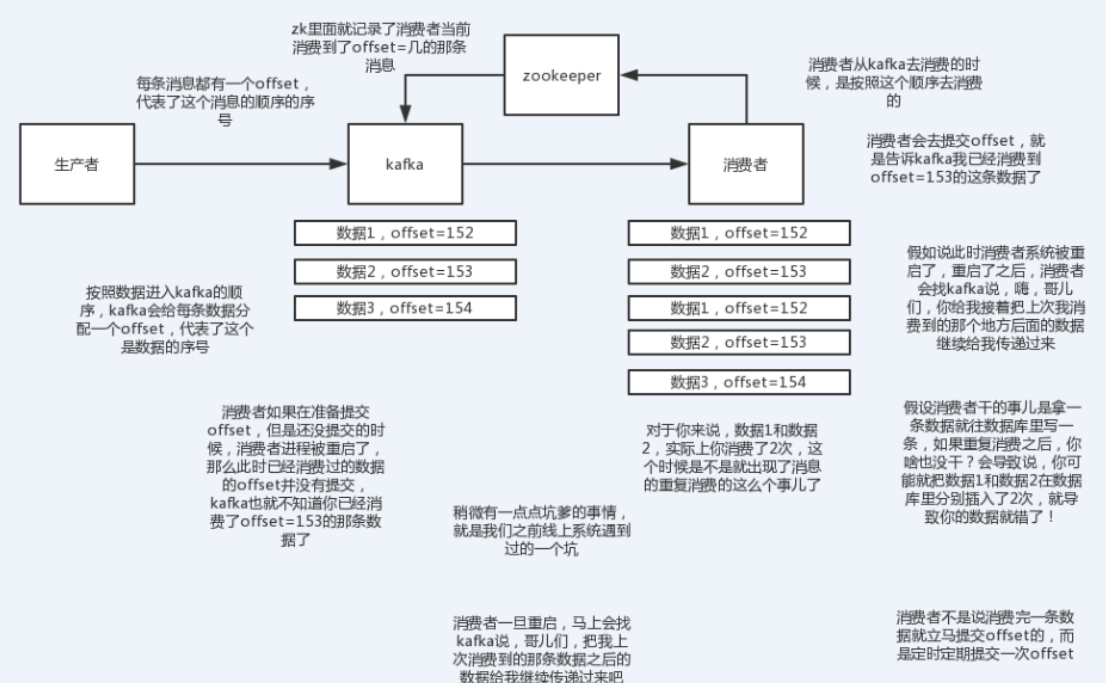
两者都是各有利弊：  
同步发送模式虽然吞吐量小，但是发一条收到确认后再发下一条，既能保证不丢失消息，又能保证顺序

**保证消息不丢失不重复**

**为什么会重复消费消息：**每个消息写进的时候都有个offset代表序号，consumer消费后，每隔一段时间会把自己消费过的消息的offset提交一下，代表已经消费过了，下次重启的话就从上次消费到的offset开始消费。但如果重启不当，比如直接kill进程再重启，会导致consumer把有些消息处理了但没提交offset，这就导致了有些消费被再消费一次。

重复消费不要紧，但幂等性(一个请求来多次，对应数据/结构也不变)不能保证就不行

如果消费消息是直接把消息插入到数据库，那么重复消费就会在数据库插入多条重复数据



**解决重复**：1.插入数据库的时候先拿主键查一下，数据已经有了就只update。如果是Redis 用set

2.情景复杂点的可以让生产者发消息的时候加个全局唯一的id，消费的时候去Redis查 有没有消费过这个id，没消费过才消费

3.还可以用数据库的唯一键，之后如果插入重复数据就会报错

**消息为什么会丢失**：1.使用同步模式的时候，有3种状态保证消息被安全生产，在配置为1（只保证写入leader成功）的话，如果刚好leader partition挂了，数据就会丢失。

2.使用异步模式的时候，当缓冲区满了，如果配置为0（还没有收到确认的情况下，缓冲池一满，就清空缓冲池里的消息），数据就会被立即丢弃掉。

**解决丢失**：1.同步模式的时候，确认机制设置为-1，也就是让消息写入leader和所有的副本。

2.在异步模式下，如果消息发出去了，但还没有收到确认的时候，缓冲池满了，在配置文件中设置成不限制阻塞超时的时间，也就说让生产端一直阻塞，这样也能保证数据不会丢失。

在数据消费时，如果使用了storm，要开启storm的ackfail机制；如果没有使用storm，确认数据被完成处理之后，再更新offset值。低级API中需要手动控制offset值。

总结：

消费端重复消费：建去重表

消费端丢失数据：关闭自动提交offset，处理完之后受到移位

生产端重复发送：这个不重要，消费端消费之前从去重表中判重就可以

生产端丢失数据：1.异步方式缓冲区满了，就阻塞在那，等着缓冲区可用，不能清空缓冲区

2.发送消息之后回调函数，发送成功就发送下一条，发送失败就记在日志 中，等着定时脚本来扫描

message delivery semantic(消息传递语义)：

1.最多一次（at most once）: 消息可能丢失也可能被处理，但最多只会被处理一次。

可能丢失 不会重复

2.至少一次（at least once）: 消息不会丢失，但可能被处理多次。

可能重复 不会丢失

3.精确传递一次（exactly once）: 消息被处理且只会被处理一次。

不丢失 不重复 就一次

而kafka其实有两次消息传递，一次生产者发送消息给kafka，一次消费者去kafka消费消息。

两次传递都会影响最终结果，两次都是精确一次，最终结果才是精确一次。

生产者里参数ack可以有三个值选择：

 0： producer完全不管broker的处理结果 回调也就没有用了 并不能保证消息成功发送 但是这种吞吐量最高

-1或者all： leader broker会等消息写入 并且ISR都写入后才会响应，这种只要ISR有副本存活就肯定不会丢失，但吞吐量最低。

1： 默认的值 leader broker自己写入后就响应，不会等待ISR其他的副本写入，只要leader broker存活就不会丢失，即保证了不丢失，也保证了吞吐量。

所以设置为0时，实现了at most once，而且从这边看只要保证集群稳定的情况下，不设置为0，消息不会丢失。

但是还有一种情况就是消息成功写入，而这个时候由于网络问题producer没有收到写入成功的响应，producer就会开启重试的操作，直到网络恢复，消息就发送了多次。这就是at least once了。

kafka producer 的参数acks 的默认值为1，所以默认的producer级别是at least once。并不能exactly once。消费者参数默认的也是at least once

实现exactly once：后面版本的Kafkai有idempotent(幂等 ai ding p tent) producer机制，在这个机制里同一消息可能被producer发送多次，但是在broker端只会写入一次，他为每一条消息编号去重，而且对kafka开销影响不大。在多分区的情况，用事务保证原子性的写入多个分区，也就是写入到多个分区的消息要么全部成功，要么全部回滚。

consumer端可能无法消费事务中所有消息，而且消息可能被删除，所以可能要自己管理offset的提交，不要自动提交

1. **消息队列的延时和过期失效问题, 消息队列满了, 很多数据挤压很久怎么办**

可能消费端出问题，消息队列集群的磁盘快写满了也没人消费，然后积压太久导致设置了过期时间的消息过期了(RabbitMQ里才能设置)怎么办？

常见场景是：消费端每次消费之后要写 mysql，结果 mysql 挂了，消费端卡那儿了，不动了；或者是消费端出了个什么岔子，导致消费速度极其慢

**解决消息队列满了：**临时紧急扩容

1.先修复 consumer 的问题，确保其恢复消费速度，然后将现有 consumer 都停掉。

2.新建一个 topic，partition 是原来的 10 倍，临时建立好原先 10 倍的 queue 数量。

3.然后写一个临时的分发数据的consumer 程序，这个程序部署上去消费积压的数据，消费 之后不做耗时的处理，直接均匀轮询写入临时建立好的 10 倍数量的 queue。

4.接着临时征用10 倍的机器来部署 consumer，每一批 consumer 消费一个临时 queue 的 数据。这种做法相当于是临时将 queue 资源和 consumer 资源扩大 10 倍，以正常的 10 倍 速度来消费数据。

5.等快速消费完积压数据之后，得恢复原先部署的架构，重新用原先的 consumer 机器来消 费消息。

**解决MQ中消息过期**：批量重导--在用户量少的时候，比如半夜，写临时程序把丢失的数据 重新灌入到MQ

1. **架构/设计一个MQ消息队列**

1.支持可伸缩性也就是能快速扩容，这样就可以增加吞吐量和容量，实现的话是设计个分布式的系统，比如按照kafka 的设计，broker -> topic -> partition，每个 partition 放一个机器，就存一部分数据。如果现在资源不够了，就给 topic 增加 partition，然后做数据迁移，增加机器，这样就可以存放更多数据，提供更高的吞吐量

2. 数据落磁盘，这样才能保证进程挂了数据还在。落磁盘是用顺序写，这样就没有磁盘随机读写的寻址开销，磁盘顺序读写的性能是很高的，也就是 kafka 的思路

3.要保证可用性，也就是要用多副本 -> leader & follower -> broker 挂了后重新选leader就可以继续对外服务。

4.支持数据不丢失

基本命令

先创建主题 发消息

