**Zookeeper**

**什么是Zookeeper**

zookeeper是一个开源的分布式协调服务，由雅虎公司创建，是google chubby的开源实现。zookeeper的设计目标是将哪些复杂且容易出错的分布式一致性服务封装起来，构成一个高效可靠的原语集（由若干条指令组成的，完成一定功能的一个过程），并且以一些列简单一用的接口提供给用户使用

**应用场景**

1. 数据发布与订阅（配置中心）

发布与订阅模型，即所谓的配置中心，顾名思义就是发布者将数据发布到ZK节点上，供订阅者动态获取数据，实现配置信息的集中式管理和动态更新。例如全局的配置信息，服务式服务框架的服务地址列表等就非常适合使用。

1. 负载均衡

这里说的负载均衡是指软负载均衡。在分布式环境中，为了保证高可用性，通常同一个应用或同一个服务的提供方都会部署多份，达到对等服务。而消费者就须要在这些对等的服务器中选择一个来执行相关的业务逻辑，其中比较典型的是消息中间件中的生产者，消费者负载均衡。

1. 命名服务(Naming Service)（注册中心）

命名服务也是分布式系统中比较常见的一类场景。在分布式系统中，通过使用命名服务，客户端应用能够根据指定名字来获取资源或服务的地址，提供者等信息。被命名的实体通常可以是集群中的机器，提供的服务地址，远程对象等等——这些我们都可以统称他们为名字（Name）。其中较为常见的就是一些分布式服务框架中的服务地址列表。通过调用ZK提供的创建节点的API，能够很容易创建一个全局唯一的path，这个path就可以作为一个名称。

服务提供者在启动的时候，向ZK上的指定节点/dubbo/${serviceName}/providers目录下写入自己的URL地址，这个操作就完成了服务的发布。

服务消费者启动的时候，订阅/dubbo/${serviceName}/providers目录下的提供者URL地址， 并向/dubbo/${serviceName} /consumers目录下写入自己的URL地址。

注意，所有向ZK上注册的地址都是临时节点，这样就能够保证服务提供者和消费者能够自动感应资源的变化。

1. 分布式通知/协调

ZooKeeper中特有watcher注册与异步通知机制，能够很好的实现分布式环境下不同系统之间的通知与协调，实现对数据变更的实时处理。使用方法通常是不同系统都对ZK上同一个znode进行注册，监听znode的变化（包括znode本身内容及子节点的），其中一个系统update了znode，那么另一个系统能够收到通知，并作出相应处理

1. 集群管理与Master选举
2. 分布式锁
3. 分布式队列

**配置**

http://apache.fayea.com/zookeeper/下载完成，通过tar -zxvf解压

单机安装

一般情况下，在开发测试环境，没有这么多资源的情况下，而且也不需要特别好的稳定性的前提下，我们可以使用单机部署；初 次 使 用zookeeper， 需要将conf目 录 下 的zoo\_sample.cfg文件copy一份重命名为zoo.cfg修改dataDir目录，dataDir表示日志文件存放的路径

集群环境安装

zookeeper集群中，各个节点总共有三种角色，分别是：leader，follower，observer

模拟3台机器来搭建zookeeper集群。分别复制安装包到三台机器上并解压，同时copy一份zoo.cfg。

修改配置文件：

server.A=B：C：D

A 是一个数字，表示这个是第几号服务器

B 是这个服务器的ip地址

C 表示的是这个服务器与集群中的Leader 服务器交换信息的端口

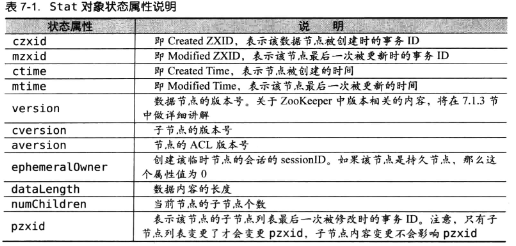
D 表示的是万一集群中的Leader 服务器挂了，需要一个端口来重新进行选举，选出一个新的Leader，而这个端口就是用来执行选举时服务器相互通信的端口。

新建datadir目录，设置myid

在每台zookeeper机器上，我们都需要在数据目录(dataDir)下创建一个myid文件，该文件只有一行内容，对应每台机器的Server ID数字；比如server.1的myid文件内容就是1。【必须确保每个服务器的myid文件中的数字不同，并且和自己所在机器的zoo.cfg中server.id的id值一致，id的范围是1~255】

**节点特性**

1. 同级节点的唯一性
2. 临时节点（会话结束会自动删除）和持久化节点
3. 有序节点
4. 临时节点下不能存在节点



**指令**

1. 创建

create [-s] [-e] path acl

s: 顺序节点

e: 临时节点

acl: 权限控制

create /zk-book 123

1. 读取

ls：ls path [watch] 例子：ls / 只能查看当前节点的下一层节点

get: get path [watch] 例子：get /zk-book 可以得到节点的value值（123）

1. 更新

set path data [version](节点有版本概念dataVersion)

set /zk-book 456

1. 删除

delete path [version]

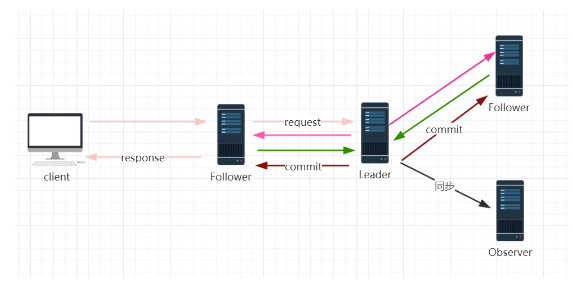
delete /zk-book (如果删除的节点下面有子节点，不能删除)

**常用命令**

1. 启动ZK服务: bin/zkServer.sh start
2. 查看ZK服务状态: bin/zkServer.sh status
3. 停止ZK服务: bin/zkServer.sh stop
4. 重启ZK服务: bin/zkServer.sh restart
5. 连接服务器：zkCli.sh -timeout 0 -r -server ip:port
6. 登陆客户端：sh zkCli.sh

**zookeeper的集群**

在zookeeper中，客户端会随机连接到zookeeper集群中的一个节点，如果是读请求，就直接从当前节点中读取数据，如果是写请求，那么请求会被转发给leader提交事务，然后leader会广播事务，只要有超过半数节点写入成功，那么写请求就会被提交（类2PC事务）



所有事务请求必须由一个全局唯一的服务器来协调处理，这个服务器就是Leader服务器，其他的服务器就是follower。leader服务器把客户端的失去请求转化成一个事务Proposal（提议），并把这个Proposal分发给集群中的所有Follower服务器。之后Leader服务器需要等待所有Follower服务器的反馈，一旦超过半数的Follower服务器进行了正确的反馈，那么Leader就会再次向所有的Follower服务器发送Commit消息，要求各个follower节点对前面的一个Proposal进行提交

**集群角色**

1. leader角色

Leader服务器是整个zookeeper集群的核心，主要的工作任务有两项

1. .事物请求的唯一调度和处理者，保证集群事物处理的顺序性
2. .集群内部各服务器的调度者
3. Follower角色
4. .处理客户端非事物请求、转发事物请求给leader服务器
5. .参与事物请求Proposal的投票（需要半数以上服务器通过才能通知leader commit数据; Leader发起的提案，要求Follower投票）
6. .参与Leader选举的投票
7. Observer角色

Observer是zookeeper3.3开始引入的一个全新的服务器角色，从字面来理解，该角色充当了观察者的角色。观察zookeeper集群中的最新状态变化并将这些状态变化同步到observer服务器上。Observer的工作原理与follower角色基本一致，而它和follower角色唯一的不同在于observer不参与任何形式的投票，包括事物请求Proposal的投票和leader选举的投票。简单来说，observer服务器只提供非事物请求服务，通常在于不影响集群事物处理能力的前提下提升集群非事物处理的能力（在不影响写的性能下提升了读的性能）

**集群组成**

通常zookeeper是由2n+1台server组成，每个server都知道彼此的存在。对于2n+1台server，只要有n+1台（大多数）server可用，整个系统保持可用。我们已经了解到，一个zookeeper集群如果要对外提供可用的服务，那么集群中必须要有过半的机器正常工作并且彼此之间能够正常通信，基于这个特性，如果向搭建一个能够允许F台机器down掉的集群，那么就要部署2\*F+1台服务器构成的zookeeper集群。因此3台机器构成的zookeeper集群，能够在挂掉一台机器后依然正常工作。一个5台机器集群的服务，能够对2台机器怪调的情况下进行容灾。如果一台由6台服务构成的集群，同样只能挂掉2台机器。因此，5台和6台在容灾能力上并没有明显优势，反而增加了网络通信负担。系统启动时，集群中的server会选举出一台server为Leader，其它的就作为follower（这里先不考虑observer角色）。之所以要满足这样一个等式，是因为一个节点要成为集群中的leader，需要有超过及群众过半数的节点支持，这个涉及到leader选举算法。同时也涉及到事务请求的提交投票

**ZAB协议**

概念：ZAB(Zookeeper Atomic Broadcast) 协议是为分布式协调服务 ZooKeeper 专门设计的一种支持崩溃恢复的原子广播协议。在 ZooKeeper 中，主要依赖 ZAB 协议来实现分布式数据一致性，基于该协议ZooKeeper 实现了一种主备模式的系统架构来保持集群中各个副本之间的数据一致性。

模式：ZAB 协议包含两种基本模式，分别是崩溃恢复和原子广播。

1. 崩溃恢复

当整个集群在启动或者当 leader 节点出现网络中断、崩溃等情况时，ZAB 协议就会进入恢复模式并选举产生新的 leader，当 leader 服务器选举出来后并且集群中有过半的机器和该 leader 节点完成数据同步(用来保证集群中过半的机器能够和 leader 服务器的数据状态保持一致)后，ZAB 协议就会退出恢复模式。简单地说，就是 leader 挂了，然后需要选举出新的 leader（ZK是有中心化节点的中间件），完成新 leader 和 follower 和 observer 的数据同步。

原理：

1. 、选举出新的leader

基于下面的数据恢复的要点，必须要设计一个leader算法，能够确保已经被 leader 提交的事务 Proposal 能够提交、同时丢弃已经被跳过的事务 Proposal

选举阶段，集群间互传的消息称为投票，投票Vote主要包括二个维度的信息：ID、ZXID

1. 、被推举的leader的服务器ID，集群中的每个zk节点启动前就要配置好这个全局唯一的ID
2. 、被推举的leader的事务ID ，该值是从机器DataTree内存中取的，即事务已经在机器上被commit过了。

节点进入选举阶段后的大体执行逻辑如下：

1. 、设置状态为LOOKING，初始化内部投票Vote (id,zxid) 数据至内存，并将其广播到集群其它节点。节点首次投票都是选举自己作为leader，将自身的服务ID、处理的最近一个事务请求的ZXID（ZXID是从内存数据库里取的，即该节点最近一个完成commit的事务id）及当前状态广播出去。然后进入循环等待及处理其它节点的投票信息的流程中。
2. 、循环等待流程中，节点每收到一个外部的Vote信息，都需要将其与自己内存Vote数据进行PK，规则为取ZXID大的，若ZXID相等，则取ID大的那个投票。若外部投票胜选，节点需要将该选票覆盖之前的内存Vote数据，并再次广播出去；同时还要统计是否有过半的赞同者与新的内存投票数据一致，无则继续循环等待新的投票，有则需要判断leader是否在赞同者之中，在则退出循环，选举结束后根据选举结果及各自角色切换状态，leader切换成LEADING、follower切换到FOLLOWING、observer切换到OBSERVING状态。
3. 、数据恢复
4. 、已经被处理的消息不能丢：当 leader收到合法数量 follower的acks后，就向各个follower广播COMMIT命令，同时也会在本地执行COMMIT并向连接的客户端返回acks。但是如果在各个follower在收到COMMIT命令前leader就挂了，导致剩下的服务器并没有执行都这条消息。即leader在发commit命令前挂了（此时follower已经将该消息存入本地，只是没有commit），导致 follower 没有收到这条已经被 leader 处理的消息。
5. 、被丢弃的消息不能再次出现：当leader接收到消息请求生成 proposal后就挂了，其他follower并没有收到此proposal，因此经过恢复模式重新选了leader后，这条消息是被跳过的。 此时之前挂了的leader重新启动并注册成了follower，他保留了被跳过消息的proposal状态，与整个系统的状态是不一致的，需要将其删除。
6. 原子广播

当集群中已经有过半的 follower 节点完成了和 leader 状态同步以后，那么整个集群就进入了消息广播模式。这个时候 leader 节点正常工作，如果启动一台新的服务器加入到集群，那这个服务器会直接进入数据恢复模式，和 leader 节点进行数据同步。同步完成后即可正常对外提供非事务请求的处理。

消息广播的原理（zxid）：消息广播的过程实际上是一个简化版本的二阶段（2PC）提交过程。

1). leader 接收到消息请求后，将消息赋予一个全局唯一的64 位自增 id（叫:zxid），通过 zxid 的大小比较既可以实现因果有序这个特征。

2). leader 为每个 follower 准备了一个 FIFO 队列(通过 TCP协议来实现，以实现了全局有序这一个特点)将带有 zxid的消息作为一个提案(proposal)分发给所有的 follower。

3). 当 follower 接收到 proposal，先把 proposal 写到磁盘，写入成功以后再向 leader 回复一个 ack。

4). 当 leader 接收到合法数量(超过半数节点)的 ACK 后，leader 就会向这些 follower 发送 commit 命令，同时会在本地执行该消息。

5). 当 follower 收到消息的 commit 命令以后会提交该消息。

**数据存储**

事务日志:zoo.cfg的配置文件下面的dataDir配置项指定路径生成

快照日志

运行时日志: bin/zookeeper.out

**Java应用**

见com.jyh.middleware.zookeeper.ZookeeperTest

**Watcher事件机制**

绑定事件的方法：getData、Exists、getChildren

触发事件：事务操作增（create）、删（delete）、改（setData）

事件类型：

None (-1)：客户端链接状态发生变化的时候，会收到none的事件

NodeCreated (1)：创建节点的事件。比如zk-persis-micNode

Deleted (2): 删除节点的事件NodeData

Changed (3): 节点数据发生变更

NodeChildrenChanged (4): 子节点被创建、被删除、会发生事件触发

**Curator客户端使用**

Curator是Netflix公司开源的一个ZooKeeper client library，用于简化ZooKeeper客户端编程