链滴

Zookeeper 集群配置

作者：flowaters

原文链接：https://ld246.com/article/1514967012054

来源网站：链滴

基本使用

下载

wget http://mirror.bit.edu.cn/apache/zookeeper/zookeeper-3.4.11/zookeeper-3.4.11.tar.gz

修改配置文件 zoo.cfg

加入机器列表，如

```
server.1=10.147.0.1:2185:2188
server.2=10.147.0.2:2185:2188
server.3=10.147.0.3:2185:2188
```

server.id=host:port:port的配置说明：

- server.id：在 zoo.cfg 文件中配置的配置的 dataDir 路径下（即：/var/lib/zookeeper）创建 myid文件，里面写入 id 值
- host：构建 zookeeper 集群的服务器的 ip 地址
- port：第一个 port 用于连接 leader 服务器；第二个 port 用于 leader election

创建myid文件

在 zoo.cfg 文件中配置的配置的 dataDir 路径下（即：/var/lib/zookeeper）创建myid文件，里面入 id 值

参考

- ZooKeeper 单机模式和集群模式的环境搭建

日志清理

默认是不清理snapshot日志的，会导致磁盘空间越来越少。。。

手动清理

保留100个snapshot

bin/zkCleanup.sh /path/to/zookeeper/data/ -n 100

自动清理

修改zoo.cfg配置文件

```
# 保留50个snap
autopurge.snapRetainCount=50

# 每3小时清理一次. 0为不清理.
```

autopurge.purgeInterval=3

四字命令

翻译自官方文档ZooKeeper Commands: The Four Letter Words

conf

new in 3.3.0

Print details about serving configuration

输出相关服务配置的详细信息

cons

new in 3.3.0

List full connection/session details for all clients connected to this server

列出所有连接到服务器的客户端的完全的连接 / 会话的详细信息
包括“接受 / 发送”的包数量、会话 id 、操作延迟、最后的操作执行等等信息。

crst

new in 3.3.0

Reset connection/session statistics for all connections
重置所有连接的连接/会话统计信息

dump

Lists the outstanding sessions and ephemeral nodes. This only works on the leader.

列出未经处理的会话和临时节点

envi

Print details about serving environment

ruok

Tests if server is running in a non-error state. The server will respond with imok if it is running. Otherwise it will not respond at all.

A response of "imok" does not necessarily indicate that the server has joined the quorum, just that the server process is active and bound to the specified client port. Use "stat" for details o state wrt quorum and client connection information.

测试服务是否处于正确状态。如果确实如此，那么服务返回“imok”，否则不做任何相应。

srst

Reset server statistics.

重置服务端统计信息

srvr

New in 3.3.0

Lists full details for the server.

显示服务端详细信息

stat

Lists brief details for the server and connected clients.

显示服务端概要信息

wchs

New in 3.3.0

Lists brief information on watches for the server.

列出服务器watch的详细信息

wchc

New in 3.3.0

Lists detailed information on watches for the server, by session. This outputs a list of sessions(onnections) with associated watches (paths). Note, depending on the number of watches this operation may be expensive (ie impact server performance), use it carefully.

通过 session 列出服务器 watch 的详细信息，它的输出是一个与watch 相关的会话的列表。

wchp

New in 3.3.0

Lists detailed information on watches for the server, by path. This outputs a list of paths (zno es) with associated sessions. Note, depending on the number of watches this operation may e expensive (ie impact server performance), use it carefully.

通过路径列出服务器 watch 的详细信息。它输出一个与 session相关的路径。

mntr

New in 3.4.0

Outputs a list of variables that could be used for monitoring the health of the cluster.

显示和运行健康度相关的监控信息

监控

srvr：
Zookeeper version:版本
Latency min/avg/max: 延时
Received: 收包
Sent: 发包
Connections: 连接数
Outstanding: 堆积数
Zxid: 操作id
Mode: leader/follower
Node count: 节点数

mntr:
zk_version=版本
zk_avg_latency=平均延时
zk_max_latency=最大延时
zk_min_latency=最小延时
zk_packets_received=收包数
zk_packets_sent=发包数
zk_num_alive_connections=连接数
zk_outstanding_requests=堆积请求数
zk_server_state=leader/follower 状态
zk_znode_count=znode数量
zk_watch_count=watch数量
zk_ephemerals_count=临时节点（znode）
zk_approximate_data_size=数据大小
zk_open_file_descriptor_count=打开的文件描述符数量
zk_max_file_descriptor_count=最大文件描述符数量
zk_followers=follower数量
zk_synced_followers=同步的follower数量
zk_pending_syncs=准备同步数

参考

- ZooKeeper Commands: The Four Letter Words
- ZooKeeper 使用
- zookeeper-03-常用配置和四字命令
- zookeeper四字命令
- Zookeeper 监控原型开发
- zookeeper-four-letter-words monitor: github项目，可以将zookeeper指标转化为metrics

扩容

基础知识

ZooKeeper的机制中

- myid只会向更小的myid发起连接

大的会向小的发起连接，而小的不会向大的发起连接。
所以重启时，要最先重启最小的机器，

扩容

- 在 conf/zoo.cfg末尾增加上新增的机器
- 在 data/myid加上对应的id
- 启动

注意：增加的机器个数，不能>=集群原有的机器数目.

ClickHouse中Zookeeper的配置

见ClickHouse官方建议

具体如下：

- 版本：> 3.5

zoo.cfg

```
# http://hadoop.apache.org/zookeeper/docs/current/zookeeperAdmin.html

# The number of milliseconds of each tick
tickTime=2000
# The number of ticks that the initial
# synchronization phase can take
initLimit=30000
# The number of ticks that can pass between
# sending a request and getting an acknowledgement
syncLimit=10

maxClientCnxns=2000

maxSessionTimeout=60000000
# the directory where the snapshot is stored.
dataDir=/opt/zookeeper/{{ cluster['name'] }}/data
# Place the dataLogDir to a separate physical disc for better performance
dataLogDir=/opt/zookeeper/{{ cluster['name'] }}/logs

autopurge.snapRetainCount=10
autopurge.purgeInterval=1

# To avoid seeks ZooKeeper allocates space in the transaction log file in
# blocks of preAllocSize kilobytes. The default block size is 64M. One reason
```

```
# for changing the size of the blocks is to reduce the block size if snapshots
# are taken more often. (Also, see snapCount).
preAllocSize=131072

# Clients can submit requests faster than ZooKeeper can process them,
# especially if there are a lot of clients. To prevent ZooKeeper from running
# out of memory due to queued requests, ZooKeeper will throttle clients so that
# there is no more than globalOutstandingLimit outstanding requests in the
# system. The default limit is 1,000.ZooKeeper logs transactions to a
# transaction log. After snapCount transactions are written to a log file a
# snapshot is started and a new transaction log file is started. The default
# snapCount is 10,000.
snapCount=3000000

# If this option is defined, requests will be will logged to a trace file named
# traceFile.year.month.day.
#traceFile=

# Leader accepts client connections. Default value is "yes". The leader machine
# coordinates updates. For higher update throughput at thes slight expense of
# read throughput the leader can be configured to not accept clients and focus
# on coordination.
leaderServes=yes

standaloneEnabled=false
dynamicConfigFile=/etc/zookeeper-{{ cluster['name'] }}/conf/zoo.cfg.dynamic
```

Java version

```
Java(TM) SE Runtime Environment (build 1.8.0_25-b17)
Java HotSpot(TM) 64-Bit Server VM (build 25.25-b02, mixed mode)
```

JVM parameters

```
NAME=zookeeper-{{ cluster['name'] }}
ZOOCFGDIR=/etc/$NAME/conf

# TODO this is really ugly
# How to find out, which jars are needed?
# seems, that log4j requires the log4j.properties file to be in the classpath
CLASSPATH="$ZOOCFGDIR:/usr/build/classes:/usr/build/lib/*.jar:/usr/share/zookeeper/zook
eper-3.5.1-metrika.jar:/usr/share/zookeeper/slf4j-log4j12-1.7.5.jar:/usr/share/zookeeper/slf4j-
pi-1.7.5.jar:/usr/share/zookeeper/servlet-api-2.5-20081211.jar:/usr/share/zookeeper/netty-3.7
0.Final.jar:/usr/share/zookeeper/log4j-1.2.16.jar:/usr/share/zookeeper/jline-2.11.jar:/usr/share
zookeeper/jetty-util-6.1.26.jar:/usr/share/zookeeper/jetty-6.1.26.jar:/usr/share/zookeeper/jav
cc.jar:/usr/share/zookeeper/jackson-mapper-asl-1.9.11.jar:/usr/share/zookeeper/jackson-core
asl-1.9.11.jar:/usr/share/zookeeper/commons-cli-1.2.jar:/usr/src/java/lib/*.jar:/usr/etc/zookee
er"

ZOOCFG="$ZOOCFGDIR/zoo.cfg"
ZOO_LOG_DIR=/var/log/$NAME
USER=zookeeper
GROUP=zookeeper
PIDDIR=/var/run/$NAME
```

```
PIDFILE=$PIDDIR/$NAME.pid
SCRIPTNAME=/etc/init.d/$NAME
JAVA=/usr/bin/java
ZOOMAIN="org.apache.zookeeper.server.quorum.QuorumPeerMain"
ZOO_LOG4J_PROP="INFO,ROLLINGFILE"
JMXLOCALONLY=false
JAVA_OPTS="-Xms{{ cluster.get('xms','128M') }} \
 -Xmx{{ cluster.get('xmx','1G') }} \
 -Xloggc:/var/log/$NAME/zookeeper-gc.log \
 -XX:+UseGCLogFileRotation \
 -XX:NumberOfGCLogFiles=16 \
 -XX:GCLogFileSize=16M \
 -verbose:gc \
 -XX:+PrintGCTimeStamps \
 -XX:+PrintGCDateStamps \
 -XX:+PrintGCDetails
 -XX:+PrintTenuringDistribution \
 -XX:+PrintGCApplicationStoppedTime \
 -XX:+PrintGCApplicationConcurrentTime \
 -XX:+PrintSafepointStatistics \
 -XX:+UseParNewGC \
 -XX:+UseConcMarkSweepGC \
-XX:+CMSParallelRemarkEnabled"
```

Salt init

```
description "zookeeper-{{ cluster['name'] }} centralized coordination service"

start on runlevel [2345]
stop on runlevel [!2345]

respawn

limit nofile 8192 8192

pre-start script
    [ -r "/etc/zookeeper-{{ cluster['name'] }}/conf/environment" ] || exit 0
    . /etc/zookeeper-{{ cluster['name'] }}/conf/environment
    [ -d $ZOO_LOG_DIR ] || mkdir -p $ZOO_LOG_DIR
    chown $USER:$GROUP $ZOO_LOG_DIR
end script

script
    . /etc/zookeeper-{{ cluster['name'] }}/conf/environment
    [ -r /etc/default/zookeeper ] && . /etc/default/zookeeper
    if [ -z "$JMXDISABLE" ]; then
        JAVA_OPTS="$JAVA_OPTS -Dcom.sun.management.jmxremote -Dcom.sun.management.
mxremote.local.only=$JMXLOCALONLY"
    fi
    exec start-stop-daemon --start -c $USER --exec $JAVA --name zookeeper-{{ cluster['name']
}} \
        -- -cp $CLASSPATH $JAVA_OPTS -Dzookeeper.log.dir=${ZOO_LOG_DIR} \
        -Dzookeeper.root.logger=${ZOO_LOG4J_PROP} $ZOOMAIN $ZOOCFG
end script
```

参考

- zookeeper集群扩容/下线节点实践
- ZooKeeper在线迁移