# 安装部署

部署数据同步工具的客户端服务器必须要是大数据平台的Hadoop客户端节点，可以正确执行并输出hadoop classpath命令。该客户端必须可以识别所有集群的所有主机名hostname。

将压缩包hive-sync.zip上传到服务器并解压。

|  |
| --- |
| unzip hive-sync.zip |

解压后所有文件都在hive-sync文件夹下，该文件夹下有5个文件夹

* conf 存储数据同步配置文件目录，cluster.properties存储在该目录下
* krb 存储kerberos的配置文件krb5.conf和认证文件keytab
* lib 存储数据同步运行依赖jar包目录
* logs 存储数据同步运行日志目录和数据同步运行结果文件
* sbin 存储数据同步运行脚本目录

# 配置说明

## 配置系统环境变量

在运行数据同步工具用户的家目录的.bash\_profile中配置HIVE\_SYNC\_HOME环境变量/home/user/.bash\_profile。如果是root用户则直接在/root/.bash\_profile中配置。

|  |
| --- |
| vi /root/.bash\_profile  export HIVE\_SYNC\_HOME=/root/zxh/hive-sync  source /root/.bash\_profile |

## 修改运行配置参数

数据同步工具所产生的jar包中包含了默认的cluster.properties配置文件，当用户不再额外指定配置文件时使用的是jar包中默认配置，用户可以指定外部cluster.properties配置文件覆盖jar包中默认配置项。

修改$HIVE\_SYNC\_HOME/conf/目录下的cluster.properties，修改kerberos配置文件存储目录和并发同步线程数。

|  |
| --- |
| *# kerberos配置，该目录下存储了对应用户的keytab文件和krb5.conf文件* **krb.path**=**/root/hive-sync/krb krb.user**=**admin@MC.COM** *# 洞察集群配置，每个集群包含3个配置项 # name.hive.conf 指定集群名称和集群对应的hive-site.xml路径* **dongcha.hive.conf**=**dongcha/hive-site.xml** *# name.hdfs.fs 指定集群名称和对应集群的hdfs nameservice名称* **dongcha.hdfs.fs**=**nameservice1** *# name.version 指定集群名称和对应集群的版本* **dongcha.version**=**cdh5** *# 演练集群配置* **yanlian.hive.conf**=**yanlian/hive-site.xml yanlian.hdfs.fs**=**nameservice1 yanlian.version**=**cdh6** *# cdp7集群配置* **cdp7.hive.conf**=**cdp7/hive-site.xml cdp7.hdfs.fs**=**bdpdiylnameservice1 cdp7.version**=**cdp7** *# distcp nameservices配置文件路径，包含所有集群的nameservice解析* **distcp.nameservices.conf**=**distcp/hdfs-site.xml** *# distcp所提交的yarn集群配置文件路径，即yarn-site.xml配置文件路径 # 该文件决定了distcp具体提交到哪个yarn集群运行，该集群必须能够识别所有集群的主机hostname* **distcp.yarn.conf**=**distcp/yarn-site.xml** *# 并发数配置* **sync.concurrent**=1*# distcp重试次数配置* **distcp.retry.num**=1*# mysql配置* **mysql.driver**=**com.mysql.jdbc.Driver mysql.url**=**jdbc:mysql://10.10.10.1355:3306 mysql.db**=**hive mysql.user**=**zh mysql.passwd**=**2ws** |

用户可以在配置文件中指定参数项对应的值，同时可以在数据同步启动命令行中覆盖配置文件中的同名参数，命令行参数优先级最高，外部cluster.properties中配置参数优先级其次，再到jar包中cluster.properties配置参数，最后是硬编码中的配置项。目前命令和cluster.properties中支持覆盖的同名参数有sync.concurrent和distcp.retry.num。

命令行启动时可以追加参数对本次运行生效，对于cluster.properties中的同名参数会进行覆盖，命令行当前支持的参数如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 参数名 | 可选值 | 默认值 | 参数说明 |
| -s | single  multi | 无 | single为单表模式  multi为多表模式 |
| -D sync.concurrent | 数字类型 | 3 | 数据表处理并发数配置，这里可以理解成提交到yarn上同时运行的application个数，即distcp并发数 |
| -D distcp.retry.num | 数字类型 | 3 | distcp重试次数配置 |
| -update | 无需给值 | 无 | distcp 以update模式运行，-update和-overwrite参数不能同时出现，如果都不指定则默认-update |
| -overwrite | 无需给值 | 无 | distcp 以overwrite模式运行，-update和-overwrite参数不能同时出现，如果都不指定则默认-update |
| -skipcrccheck | 无需给值 | 无 | distcp跳过crc校验 |
| -skipchecksum | 无需给值 | 无 | 跳过数据同步之后的文件一致性校验 |
| -maps | 数字类型 | 3 | 每个distcp map数配置 |
| -bandwidth | 数字类型 | 20 | distcp 时每个map的带宽配置，单位是MB/second，实际本次数据同步同时占用的带宽是sync.concurrent \* maps \* bandwidth MB/second |
| -snapshot | true  false | true | 是否基于快照进行distcp |
| -speed | 无需给值 | 无 | 以极速模式运行 |
| -resume | 主任务ID | 无 | 以断点续传模式运行 |
| -kill | 主任务ID | 无 | 将主任务手动kill |

# 使用说明

数据同步分为新任务模式、极速模式、断点续传模式、kill主任务模式，新任务模式和极速模式以及断点续传模式默认都使用基于快照的方式进行distcp，可以在运行数据同步脚本时追加-snapshot false强制以不基于快照的方式进行distcp。具体每种模式的用法下文有详细说明。

## 新任务模式

默认运行模式即为新任务模式，新任务模式运行数据同步首先会生成一条主任务记录，每张数据表的数据比对发现不一致后生成一条子任务，主任务和子任务是一对多的关系。

进入$HIVE\_SYNC\_HOME/sbin目录，运行hive-sync.sh脚本，传入参数，第一个参数是源集群名称，第二个参数是目标集群名称，第三个参数是模式-s，第四个参数为第三个参数-s指定具体的值为single或multi，第五个参数是源集群的库表，第六个参数是目标集群的库表，这6个参数是必须要给且位置不能错乱的。

例如将yanlian集群zxh\_warehouse数据库下的order\_info表同步到cdp7集群的zxh\_warehouse\_bak数据库下，并且表增加pe\_前缀和\_sf后缀（表的重命名目前只支持增加前缀后后缀，前缀使用prefix来指定，后缀使用suffix指定，命令行传递时需要对{he}进行转移），启动命令

|  |
| --- |
| sh hive-sync.sh yanlian cdp7 -s single zxh\_warehouse.order\_info zxh\_warehouse\_bak\{prefix=pe\_,suffix=\_sf\} |

例如将yanlian集群zxh\_warehouse数据库下的所有表同步到cdp7集群的zxh\_warehouse\_bak数据库下，并且表增加pe\_前缀和\_sf后缀（表的重命名目前只支持增加前缀后后缀，前缀使用prefix来指定，后缀使用suffix指定，命令行传递时需要对{he}进行转移），启动命令

|  |
| --- |
| sh hive-sync.sh yanlian cdp7 -s multi zxh\_warehouse.\* zxh\_warehouse\_bak\{prefix=pe\_,suffix=\_sf\} |

可以继续追加新的参数对本次数据同步生效，追加的参数位置是在固定必须的6个参数之后，例如将yanlian集群zxh\_warehouse数据库下的所有表同步到cdp7集群的zxh\_warehouse\_bak数据库下，数据表的处理并发数为2，distcp失败重试次数为2，distcp以-update方式进行，distcp的每个map的带宽为10 MB/second，每个distcp 3个map，启动命令

|  |
| --- |
| sh hive-sync.sh yanlian cdp7 -s multi zxh\_warehouse.\* zxh\_warehouse\_bak -D sync.concurrent=2 -D distcp.retry.num=2 -update -bandwidth 10 -maps 3 |

数据同步过程会输出详细的日志，日志存储在$HIVE\_SYNC\_HOME/logs目录下的hive-sync.log中。数据同步主任务执行成功后会输出本次主任务所包含的子任务个数，同步成功的子任务个数，主任务运行结果日志，如下所示

|  |
| --- |
| 数据同步任务运行结束，主任务ID：165，匹配到  的数据表总数 4，元数据发生变化的表数量 4，元数据未发生变化的表数量 0，数据同步成功的表数量 4，数据同步失败的表数量 0，主任务执行结果 s  uccess，耗时：39秒  运行结果明细信  息请查看/root/zxh/hive-sync/logs/result/165-1625210085506.result |

/root/zxh/hive-sync/logs/result/165-1625210085506.result中的内容是每张表的元数据同步状态，主数据同步状态和同步耗时

|  |
| --- |
| 数据同步任务运行结束，主任务ID：165，匹配到的数据表总数 4，元数据发生变化的表数量 4，元数据未发生变化的表数量 0，数据同步成功的表数量  4，数据同步失败的表数量 0，主任务执行结果 success，耗时：39秒  数据表zxh\_warehouse.user\_info元数据同步成功，主数据同步成功，耗时：16秒  数据表zxh\_warehouse.spe\_loc元数据同步成功，主数据同步成功，耗时：16秒  数据表zxh\_warehouse.order\_info元数据同步成功，主数据同步成功，耗时：16秒  数据表zxh\_warehouse.def\_loc元数据同步成功，主数据同步成功，耗时：16秒 |

## 极速模式

极速模式和新任务模式的区别在于进行数据表的处理时不再进行元数据差异化比对，在明确知道目标集群不含所需同步的数据库时在启动数据同步时追加-speed参数声明本次数据同步使用极速模式运行，如果程序检测出目标集群包含该数据库则给出错误提示并退出，启动命令

|  |
| --- |
| sh hive-sync.sh yanlian cdp7 -s single zxh\_warehouse.order\_info zxh\_warehouse\_bak -speed |

## 断点续传模式

断点续传模式是针对之前历史传输失败的数据同步任务进行断点续传，比如之前的一次传输的是10张表，其中5张表传输成功另外5张表传输失败，使用断点续传模式可以继续处理这5张传输失败的表。断点续传需要传入2个参数，第一个参数是固定值-resume表示使用断点续传模式，第二个参数是需要续传的主任务ID（每次执行数据同步时都会有详细的日志数据，日志中可以看到主任务的ID值）。例如断点续传主任务ID为13的任务，启动命令

|  |
| --- |
| sh hive-sync.sh -resume 13 |

## Kill主任务模式

在数据同步主任务运行期间可以手动将数据同步kill，kill后正在处理的数据表会继续处理直至处理完成，还未处理的数据表不再继续处理。Kill主任务模式需要传入2个参数，第一个参数是固定值-kill表示使用kill模式，第二个参数是需要续传的主任务ID（每次执行数据同步时都会有详细的日志数据，日志中可以看到主任务的ID值）。例如kill主任务ID为13的任务，启动命令

|  |
| --- |
| sh hive-sync.sh -kill 13 |