**OceanBase 0.4.2旁路导入管理手册**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **文档**  **版本** | **修订章节** | **修订原因** | **修订日期** | **修订人** |
| 0.1 | 全部 | 新建 | 2013-10-14 | 永乐 |
| 0.2 |  | 修改了任务管理和附录 | 2013-11-26 | 永乐 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

[1 功能介绍 3](#_Toc23770)

[2 安装与部署 3](#_Toc7693)

[2.1 安装方式 3](#_Toc18496)

[2.2 目录结构 4](#_Toc9654)

[2.3 部署工作 4](#_Toc18645)

[3 启动importserver 6](#_Toc32233)

[4 启动proxyserver 6](#_Toc20615)

[5 客户端发起导入命令 6](#_Toc23436)

[6 任务管理 6](#_Toc26769)

[7 附录 7](#_Toc21296)

[7.1 配置文件说明 7](#_Toc19069)

[7.1.1 importserver.conf 7](#_Toc14987)

[7.1.2 data\_syntax.ini 8](#_Toc13866)

[7.1.3 configuration.xml 9](#_Toc14051)

[7.1.4 proxyserver.conf 12](#_Toc5419)

[7.2 导入数据的时间格式 12](#_Toc26480)

本文介绍了oceanbase的旁路导入管理员相关的信息，建议先阅读用户手册了解下大概怎么用，然后在阅读该管理员手册。

# 功能介绍

cs旁路导入的大致步骤如下

1. importserver将指定表的数据在hadoop上生成
2. Importserver发命令给rootserver发起加载
3. Rs通过proxyserver从云梯获取range table
4. Rs根据range table发命令给cs，cs从proxyserver拉去对应的range的tablet。

其中proxyserver的角色是隔离hadoop和ob，起到中转的作用。

Rs或优先让cs从集群内的cs拷贝tablet（如果已经从hadoop拉下来了，图上没有画），如果集群内没有，则会让cs从proxyserver拉取。

1. 全部的range的tablet都加载完毕后，rs切换改表的table id，使得新的数据生效，并且删掉旧的数据。

hadoop，hdfs

Rangetable

tablet

**chunkserver节点**

1. 调用mapreduce生成tablet

**rootserver节点**

importserver

proxyserver

2. 发导入命令给rs发起导入

Range table

tablet

3. Rs通过proxyserver从hadoop获取range table

chunkserver

rootserver

4. Rs发命令给cs通过proxyserver获取对应range的tablet

# 安装与部署

## 安装方式

推荐使用rpm包按照，然后用admin用户跑，这样可以尽可能使用默认配置，需要修改的配置相对比较少。具体rpm安装步骤请参考ob的部署安装文档。

sudo rpm –ivh oceanbaseXXX.rpm

## 目录结构

|-- bin(目录)

|-- etc目录

    |-- app\_name #这个需要手动重命名成对应的app\_name

    |   `-- table\_name #根据用户手册创建导入表的配置后生成

    |       |-- configuration.xml #根据用户手册创建导入表的配置后生成

    |       `-- data\_syntax.ini     #根据用户手册创建导入表的配置后生成

    |-- importcli.conf.template                #客户端的配置文件

|-- importserver.conf.template         #服务端的配置文件

`-- configuration.xml.template #这个schema文件在部署ob的时候应该已经配置了

## 部署工作

1. 检查importserver包依赖，Python版本大于等于2.6, jdk1.6，hadoop

which java #查找在的是那个java，然后用ls –l来查看是哪里的软连接，从而找到java home

ls -l /usr/java/ #用于查看java home

sudo yum install taobao-jdk-XXX

python –V #检测python版本，也可以用python2.6 –V 或者python2.7 –V，只要有大于2.6的，后面用这个允许就行。

1. 在importserver节点和proxyserver节点准备好连接hadoop的配置。

内部用户：sudo yum install yunti-hadoop-hdfs302814-mr303146-r302585-0.19.1-20130703 –b current

外部用户：用hadoop 0.19.1搭建，配置好客户端能跑mapreduce任务就可以了。

1. 将mrsstable\_lib（根据云梯所在机器的redhat决定，有5u和6u的区别）目录下的所有lib放到hdfs上的特定目录（后面configuration.xml中配置的）：

liblzo2.so, liblzo\_1.0.so, libmrsstable.so, libnone.so, libsnappy\_1.0.so, libsnappy.so。

其中liblzo2.so, libsnappy.so是第三方包，ob在mapreduce生成sstable的时候需要用到，这两个包也在oceanbase/mrsstable\_lib\_5u或者oceanbase/mrsstable\_lib\_6u目录下面可以找到。5u表示hadoop的环境是redhat5（centos5）或者兼容的linux环境，6u表示hadoop的环境是redhat6(centos6)或者兼容的linux环境。

1. 修改配置文件，具体参看配置文件的注释或者附录（importserver.conf，configuration.xml， proxyserver.conf）。

Importserver一般起在rs的一个节点上，需要配置importserver.conf，configuration.xml。

Proxyserver一般起在cs节点上，需要配置proxyserver.conf

1. 在ob上用mysql客户端创建需要的表，并且插入需要的信息

#创建需要的表

create table data\_source\_proxy(

cluster\_id int,

ip varchar(32),

port int,

supported\_data\_source\_name varchar(256),

primary key(cluster\_id, ip, port));

create table load\_data\_history (

start\_time timestamp,

end\_time timestamp,

table\_name varchar(256),

table\_id varchar(32),

status varchar(256),

uri varchar(1024),

primary key(table\_name, table\_id,start\_time));

#打开旁路导入的开关

alter system set enable\_load\_data=True server\_type=rootserver;

#创建importserver.conf中用到的监控帐号

CREATE USER 'load\_data\_monitor' IDENTIFIED BY 'import\_password';

GRANT select on load\_data\_history to 'load\_data\_monitor';

#插入起proxyserver的节点信息

insert into data\_source\_proxy values(0, '10.238.64.65', 2650, 'hadoop');

insert into data\_source\_proxy values(0, '10.238.64.66', 2650, 'hadoop');

insert into data\_source\_proxy values(0, '10.238.64.67', 2650, 'hadoop');

insert into data\_source\_proxy values(0, '10.238.64.68', 2650, 'hadoop');

insert into data\_source\_proxy values(0, '10.238.64.69', 2650, 'hadoop');

#和导入的相关的配置：并发和速度等参数，根据实际机器节点修改

alter system set migrate\_band\_limit\_per\_second='30M' server\_type=chunkserver;

alter system set balance\_max\_concurrent\_migrate\_num ='15' server\_type=rootserver;

alter system set balance\_max\_migrate\_out\_per\_cs ='3' server\_type= rootserver;

alter system set balance\_max\_migrate\_in\_per\_cs ='3' server\_type=rootserver;

alter system set max\_migrate\_task\_count ='3' server\_type=chunkserver;

alter system set datasource\_timeout ='30' server\_type=chunkserver;

alter system set load\_data\_max\_timeout\_per\_range ='5m' server\_type= rootserver;

## 注意点

* 磁盘空间警报建议设置在75%，需要预留一次导入的空间。
* 每日合并和旁路导入一定要错开，如果旁路导入过程中触发每日合并会导致改表导入失败。同时回复很复杂。

# 启动importserver

./bin/oceanbase.pl start\_is

# 启动proxyserver

建议每个cs上面部署一个proxyserver，这样cs会从本地的proxyserver取数据，性能较好。

./bin/oceanbase.pl start\_ps

# 客户端发起导入命令

 python2.6 bin/importcli.py -s import\_server\_ip -p import\_server\_port -t OVERWRITE 参看用户手册

# hadoop 任务管理

hadoop job -kill taskd 可以杀死正在跑得hadoop任务，taskid在/home/admin/oceanbase/log/appname/tablename.log里面可以找到。

# oceanbase导入任务管理

rs\_admin kill\_import table\_name table\_id

# importserver任务管理

Importserver提供了KILL和STATE两种任务管理的功能。语法如下

 python2.6 bin/importcli.py -s import\_server\_ip -p import\_server\_port -t KILL task\_id

 python2.6 bin/importcli.py -s import\_server\_ip -p import\_server\_port -t STATE task\_id （ALL,DOING,DONE,KILLED,NEED\_KILL）

其中task\_id在任务运行开始importcli.py第一行输出中可以找到，也可以通过STATE命令查看DOING状态的任务来查询。

当旁路导入错过dispatch\_time\_range这个设置的时间窗口的时候可能需要把该次旁路导入终止掉，这时候就需要KILL命令。

KILL命令是把对应的overwrite或者import的task杀死(集群复制不允许被杀死)，但是这个操作只是把任务状态改成NEED\_KILL，需要等待任务检测到自己状态被设置成NEED\_KILL才会退出来，这个操作有一定的延时和局限性。旁路导入中mapreduce过程不会被这个操作停止，拷贝和分发sstable的过程可以被终止。如果一个task在mapreduce阶段接收到KILL命令，只有在mapreduce结束进入拷贝和分发sstable的时候才会退出。如果想要立刻退出，可以用hadoop命令直接发kill命令给对应的mapreduce任务。

STATE命令后面的task\_id可以替换成（ALL,DOING,DONE,KILLED,NEED\_KILL）中的任意一个，分别表示显示对应状态的task：

* ALL 所有状态的任务
* DOING 正在导入的任务
* DONE 已经成功完成的任务
* KILLED 被杀死的任务
* NEED\_KILL 收到杀死命令但是还没退出的任务

保证的task状态数为32个。如果超过了这个限制，那么每次新增任务的时候就会清除DONE和KILLED状态的task的命令直到总task数目降到32个。

# 附录

## 配置文件说明

高亮的部分是必须需要改的。

### importserver.conf

## public conf for all server

[public]

app\_name = #OB的app\_name，必须和oceanbase的集群的相同

#sstable\_block\_size=65536

ob\_user = load\_data\_monitor #OB导入状态的监控帐号

ob\_password = import\_password #OB导入状态的监控帐号

ob\_port = 2828 #rs节点的lms对外服务端口

load\_table\_timeout = 3600

# 30s per retry, start\_load\_table\_retry for each table

start\_load\_table\_retry = 10

# all paths of importserver

# base\_dir must be defined manually

# default:

# conf\_dir = $base\_dir/etc/

# bin\_dir = $base\_dir/bin/

# log\_dir = $base\_dir/log/

# data\_local\_dir = $base\_dir/data/

base\_dir = /home/admin/oceanbase/ #这个是rpm包的默认位置，如果不是的话自己改这个就可以了，前面的四个目录如有没有特殊需求，只用默认值就行了

# hadoop env

# default:

# hadoop\_bin\_dir = $hadoop\_home/bin

java\_home = //home/yunti/java-current

hadoop\_home = /home/hadoop/hadoop-current

hadoop\_conf\_dir = /home/admin/config #hadoop的配置目录

keep\_hadoop\_data\_days = 3 #hadoop上生成的数据保留的天数

# hdfs path

# hdfs\_name is the name of hdfs server

# hdfs\_data\_dir is the path on hdfs to store work data

# hdfs\_conf\_dir is the path on hdfs to store the tmp conf of the tool mrsstable to generate sstable

hdfs\_name = hdfs://hdpnn:9000

hdfs\_data\_dir = /group/oceanbase/app\_name/output/

hdfs\_conf\_dir = /group/oceanbase/app\_name/config/ #这个目录是用来存旁路导入生成sstable的mapreduce的配置的，需要在hadoop上手动建立该目录

# dispatch\_time\_range format is: start\_time,hours

#dispatch\_time\_range=01:00,8 #这个是用于延迟加载的选项，前一个时间点是可以开始加载的时间点，后者是可以加载点的持续时间。不过一旦开始加载操作，就算过了这个时间点也不会自动停止。

[import\_server]

port = 3000

log\_level = debug

# default pid\_file = $log\_dir/import\_server.pid

# default log\_file = $log\_dir/import\_server.log

# keep\_task\_state\_count = 128

## list the master rs of each cluster

# gateway format is ip1 ip2 ip3, sample: gateway=10.232.36.100 10.232.36.101

[ob\_instances] #这里是配置集群RS用的，如果是单集群，那么就只要配置obi\_count为1

obi\_count=0

obi0\_rs\_vip=

obi0\_rs\_port=

obi1\_rs\_vip=

obi1\_rs\_port=

### data\_syntax.ini

不用手动配。

[public]

table\_name=table\_name #这个是导入的表名，需要改成schema中表的名字。

delim=1 # 输入数据的分隔符，建议为\1。如果是空格，这里设置成32.

null\_flag=2 # 输入数据的null符，建议为\2

raw\_data\_field\_cnt=5 # 数据文件的字段个数，如果小于这个值，生成sstable的时候会直接报错

sstable\_block\_size=65536

#table\_name

[table\_name] #这个是导入的表名，需要改成schema中表的名字。

#rowkey\_desc='column index in raw data','column type','len' （只有varchar才有len）

#column type:

#1:INT

#5:DATETIME

#6:VARCHAR

#the column define order is the same as the order in rowkey

#MUST specify this option

rowkey\_desc=0-1,1-5,2-6-8

#column\_info='the column id in schema','the column index in raw data' #column id必须从小到大的顺序排列

#if some column in the schema but not in raw\_data,just set column\_index to -1

#'the column id in schema' must be ascending order

column\_info=16,3

column\_info=17,5

column\_info=18,4

### configuration.xml

<?xml version="1.0"?>

<configuration>

<property>

<name>mrsstable.native.lib.path</name>#这个请根据hadoop的lib位置自行配置

<value>hdfs://hdpnn:9000/group/oceanbase/app\_name/lib</value>

<description>native library path in hdfs.</description>

</property>

<property>

<name>mapred.reduce.tasks</name> #这个请根据数据量自行配置，这个就是生成sstable的个数,允许的最小值为2.

<value>10</value>

<description>reduce tasks count</description>

</property>

<property>

<name>mrsstable.presort.sampler</name> #这个需要根据数据特征配置,建议咨询ob的dba。

随机采样，适合于原始数据的每个文件都是随机的,

Interval采样，适合于每个数据文件都是有序的，但全局是无序的,

Split采样，获取mrsstable.max.sample.split个输入文件的前mrsstable.sample.number个作为采样

<value>random</value>

<description>

sampler type, default user specify sampler file,

user can slect 4 types: specify, random, interval, split

</description>

</property>

<!--

<property>

<name>mrsstable.partition.file</name> #除非mrsstable.presort.sampler配置为specify，否则不需要自己生成分区文件

<value>hdfs://OceanBase036029.sqa.cm4.tbsite.net:54310/home/partition/\_partition.lst</value>

<description>

partition file path in hdfs, the file name must be '\_partition.lst'.

if the sampler is specify, it's not necessaray to set this parameter.

</description>

</property>

-->

<property>

<name>mrsstable.max.sample.split</name>

<value>1</value>

<description>

the maximum sample split count, for split and random sampler,

the default value is reduceNum / 5, for interval sampler, the default

value is max(reduceNum \* 0.05, 1)

</description>

</property>

<property>

<name>mrsstable.sample.number</name> #这个参数建议调整到10000-50000

<value>10000</value>

<description>

the total sample number, for split and random sampler,

the default value is reduceNum \* 10, for interval sampler, it isn't necessary

</description>

</property>

<property>

<name>mrsstable.sample.choose.percent</name>

<value>0.001</value>

<description>

the choose percent, for interval sampler, it means the frequency with which records will be emitted.

the random and split sampler needn't this parameter

</description>

</property>

<property>

<name>mrsstable.input.format</name> #这个必须和输入文件格式相同

<value>text</value>

<description>

the input file format, default format is text, it supports text input format

and sequence input format

</description>

</property>

<property>

<name>mrsstable.skip.invalid.row</name> #如果是1，那么有错误的行的时候还能继续执行，单行的错误只能看mapreduce的日志才能发现；如果是0，那么一但有错误就会导致整个旁路导入过程失败。

<value>0</value>

<description>whether skip invalid row, default is false</description>

</property>

<!--

<property>

<name>mrsstable.rowkey.desc</name> 这个参数旁路导入的时候会从命令行传过来，不需要配置

<value>0-0-1,1-1-1,2-2-1</value>

<description>

rowkey description, it defines the rowkey formation

format = 'column index in raw data','column type','len'

</description>

</property>

-->

<!--

<property>

<name>mrsstable.primary.delimeter</name>

<value>\u0001</value>这个参数旁路导入的时候会从命令行传过来，不需要配置

<description>primary delimeter in hdfs data file, default is \1.

this will be past as parameter, no use to set this param.

</description>

</property>

-->

<!--

<property>

<name>mrsstable.table.id</name>这个参数旁路导入的时候会从命令行传过来，不需要配置

<value>1001</value>

<description>table id, this will be past as parameter, no use to set this param.</description>

</property>

-->

<property>

<name>mapred.child.java.opts</name>

<value>-Xmx1024m</value>

<description>memory limit of java child process</description>

</property>

<property>

<name>io.sort.mb</name>

<value>512</value>

<description>max memory used by sore operation</description>

</property>

<property>

<name>io.sort.factor</name>

<value>100</value>

<description>io sort factor</description>

</property>

<property>

<name>mapred.compress.map.output</name>

<value>true</value>

<description>whether compress the output of map</description>

</property>

</configuration>

### proxyserver.conf

#[public]

task\_queue\_size=10000

task\_thread\_count=8

io\_thread\_count=8

#second

network\_timeout=30

#millisecond

task\_left\_time=300

#java home

java\_home=/home/yunti/java-current/

[yunti]

#hdfs config

hdfs\_conf\_dir=/home/admin/config

#hadoop home

hadoop\_home=/home/yunti/hadoop-current/

#mem use to cache hadoop meta (bytes)

mem\_limit=10000000

#buffer use to read file from yunti in one work thread, so total memory is buf\_size \* task\_thread\_count (bytes)

buf\_size=64000000

## 导入数据的时间格式

导入的数据中支持的时间格式如下：

%4d-%2d-%2d %2d:%2d:%2d //YYYY-MM-DD hh:mm:ss

%4d-%2d-%2d //YYYYMMDD

%4d%2d%2d %2d:%2d:%2d //YYYYMMDD hh:mm:ss

%4d%2d%2d%2d%2d%2d //YYYYMMDDhhmmss

%4d%2d%2d //YYYYMMDD

## 常见错误处理

### importserver日志中发现rs\_admin发送加载命令失败

importserver中的错误日志如下：

2013-10-21 10:17:18,999] ERROR overwrite\_load\_table (importserver.py:949) [139948619069184] ERROR: failed to call rs\_admin to import table 0: Shell.popen(/home/admin/oceanbase/bin//rs\_admin -r 182.119.166.239 -p 2500 import test 3087 proxy://yunti://oceanbase/output/obtest/20131021101651-096739/test/sstables)=>255 Output=>"server[182.119.166.239:2500], timeout=20000000  
import table\_name=test table\_id=3087 uri=proxy://yunti://oceanbase/output/obtest/20131021101651-096739/test/sstables  
failed to import table, err=-3012

rootserver的错误日志如下：

[2013-10-21 10:20:20.122635] ERROR add\_load\_table (ob\_root\_balancer.cpp:1972) [139945159145216] load table could only start when at least one replica num reach the last frozen mem version

[2013-10-21 10:20:20.122647] ERROR import (ob\_root\_server2.cpp:6355) [139945159145216] fail to add import table=test with new table\_id=3087, uri=proxy://yunti://oceanbase/output/obtest/20131021101651-096739/test/sstables, ret=-3012

处理方法：用rs\_admin dump\_unusual\_tablets确认哪些tablet不全，处理这些不全的tablet。

### 旁路导入在切换schema成功后设置DONE状态的时候失败

这个现象只有在rsvip切换且旁路导入刚好在结束点的时候才可能发生。

这个时候新的数据已经可以查询了，但是同一个表新的导入会由于内部还记录了当前正在导入的状态而失败。

这个时候需要删掉rs的commitlog重启。