想了解全部详情的同学请从上往下看：

只关心运维的同学请直接跳到第3章：

# RootServer自举文档

# 1综述

## 1.1背景

为了使oceanbase的运维更加可视化，oceanbase系统采用了内部表的方式来存储schema和root\_table，这样就使得需要为系统自举设置一个特殊的流程；

## 1.2 名词解释

* \_\_first\_tablet\_entry：存储表的第一个tablet的位置信息；
* \_\_all\_all\_column：存储所有表的列名、column\_id及其类型
* \_\_all\_join\_info：存储内部join关系
* \_\_all\_sys\_param：存储系统配置参数及其含义
* \_\_all\_sys\_stat：存储系统状态值及其含义
* 核心表：指的是\_\_first\_tablet\_entry、\_\_all\_all\_column、\_\_all\_join\_info三张表的统称；
* RS自举：集群启动到正常提供服务的一个引导过程；
* Rs\_admin：oceanbase运维使用的一个工具，可以用来读取系统的某些状态和值；
* 建表：包括产生表的schema，以及对应的空tablet;
* root\_table：用来存储oceanbase静态数据的位置信息的元数据结构；

# 2.自举流程

## 2.1 触发时机

RS端持久化存储了\_\_first\_tablet\_entry表的储存位置在first\_meta.bin文件中，该文件名可配置，如果这个文件不存在，则需要启动自举流程。在没有自举之前，RS只提供机器MS/CS/UPS的管理服务；

## 2.2 自举流程

自举流程的启动，使用rs\_admin向rootserver发送一个boot\_strap的命令，自举的过程如下：

1. RS等待至少有一个CS已经注册；
2. RS选定若干CS，发送命令在上面新建\_\_first\_tablet\_entry, \_\_all\_all\_column, \_\_all\_all\_join三个表的空表；
3. RS给UPS发送mutate消息，新建\_\_all\_sys\_param, \_\_all\_sys\_stat两个表的schema，这两张表的Schema会写入到核心表中；
4. 如果还有schema.ini形式的用户表，则会将用户表的schema写入到核心表中，这部分只在过渡时期内会存在，以后都会全部采用create table的命令来生成用户表的schema;
5. 以上步骤都成功，RS把\_\_first\_tablet\_entry的存储位置等信息持久化，至此，自举过程成功；
6. 强制要求CS重新汇报；
7. 通知UPS switch\_schema；
8. RS给UPS发送写入消息，写入\_\_all\_sys\_param, \_\_all\_sys\_stat两张表的初始内容；

## 2.3 建表流程

为了支持自举过程和内部表的实现，需要实现新的建表流程：

（1）RS向UPS取的last\_frozen\_version；

（2）解析新表的schena，由RS向UPS发送修改命令，通过修改first\_tablet\_entry, all\_all\_column, all\_all\_join三张表来完成新建表格；新建的table的schema中有一个特殊属性表示新建表格时刻的frozen\_version；

这样RS就完成了建表的过程，可以开始提供新表的读写过程，但是此时还没有为该表建立对应的空tablet，RS可以在适当的时候，为该表建立空tablet，直到该表在root\_table中记录的tablet个数不为0。

MS在读取该表的数据时，会判断schema中显示的该表新建时frozen\_version等于当前系统的frozen\_version，则不需要去查找和读取CS端的SStable上数据；

# 3.RS运维

（1）使用rs\_admin工具发送boot\_strap命令，启动RS自举流程；

（2）RS重启过程中不建议删除first\_meta.bin文件；

（3）重启RS的过程中不建议删除commit\_log；

（4）万一，不幸，将first\_meta.bin文件删除了，请使用rs\_admin发送boot\_recover的命令，启动恢复过程，恢复过程的目的只要是为了产生first\_meta.bin文件；

（5）RS没有检查root\_table中的表的tablet是否完整的过程，所以，请人工确认所有的cs都已经上线；

详细原因说明如下：

RS自举过程成功后，会在本地生成一个first\_meta.bin的文件，以表示自举完成：

（1）如果RS启动的时候，本地没有生成first\_meta.bin文件，则表明没有自举成功过，请用rs\_admin发送boot\_strap命令；

（2）如果本地存在first\_meta.bin文件，则表明RS已经自举过，则RS重启后可直接进入服务阶段；RS会通过回放日志以及接受cs的汇报来恢复root\_table；

（3）rootserver重启时如果没有first\_meta.bin文件，那RS是无法检测是否自举成功过，所以你发送boot\_recover命令的结果有两个，一，自举成功过，recover成功；二，没有成功自举过，recover失败，会出现取schema错误，这时，则需要将已经产生的集群数据清除掉，重新开始；

（4）如果rootserver重启后将commit\_log文件删除了，这种情况有一点风险，RS可以将汇报上来的表的tablet重新加入到root\_table中，但是没有一个检查的过程来判断root\_table是否完整合理；

这里的表主要分为两类：

一类是核心表，如果rootserver已经自举成功，那么可以直接将这些tablet加入到root\_table中；

另一类是用户表，RS重启后，只有在核心表的root\_table信息已经恢复以后，才能通过核心表获取用户表的schema,如果用户汇报上来的tablet没有对应的schema则拒绝将其加入到root\_table中；

（5）RS不再有init\_root\_table\_by\_report的过程，RS自举成功后，会要求CS重新汇报tablet，RS接受到汇报的tablet的时候，会检查对应的schema是否存在，如果存在则直接将其加入到root\_table中，并开始提供该tablet的服务，不会去检查单个表的tablet是否全部汇报上来；

# 4.版本兼容

## 4.1 schema的兼容

现在oceanbase还没有实现由用户手动发起create table的功能，所以还是需要使用schema.ini的方式将用户表的schema加载到集群中；create table功能将会在随后的版本中发布；

## 4.2 root\_table和first\_tablet\_entry的并行

First\_tablet\_entry表格的作用主要有两个，用来记录表的schema以及表的第一个tablet的位置信息，但是在现在的使用过程中，first\_tablet\_entry只用来记录schema，tablet的位置信息依旧有root\_table来记录；以后会逐步启用root\_table\_service的服务来提供tablet位置的储存功能；

# 附录

自举完成后，各个系统表的初始状态如下(#开头的行注释了行的schema)。

## \_\_first\_tablet\_entry

# schema := table\_name(16),table\_id,table\_type,meta\_tname,meta\_tid,load\_type,table\_def\_type,rowkey\_column\_num,column\_num,max\_used\_column\_id,replica\_num(26),...

\_\_all\_sys\_param,4,3,\_\_first\_tablet\_entry,1,1,1,9,12,27,replica1\_ipv4,replica1\_ipv6\_high,replica1\_ipv6\_low,replica1\_port,1,replica1\_row\_count,replica1\_size,replica1\_checksum,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

\_\_all\_sys\_stat,5,3,\_\_first\_tablet\_entry,1,1,1,10,14,29,replica1\_ipv4,replica1\_ipv6\_high,replica1\_ipv6\_low,replica1\_port,1,replica1\_row\_count,replica1\_size,replica1\_checksum,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

用户表的schema内容

## \_\_all\_all\_column

# schema := table\_name(16),table\_id,column\_name,column\_id,column\_group\_id,rowkey\_id,join\_table\_id,join\_column\_id,data\_type,data\_length,data\_precision,data\_scale,nullable

\_\_all\_sys\_param,4,name,16,0,1,0,0,6,256,0,0,0

\_\_all\_sys\_param,4,cluster\_role,17,0,2,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,cluster\_id,18,0,3,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,server\_type,19,0,4,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,server\_role,20,0,5,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,server\_ipv4,21,0,6,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,server\_ipv6\_high,22,0,7,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,server\_ipv6\_low,23,0,8,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,server\_ip\_port,24,0,9,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,data\_type,25,0,0,0,0,1,8,0,0,0

\_\_all\_sys\_param,4,value,26,0,0,0,0,6,256,0,0,0

\_\_all\_sys\_param,4,info,27,0,0,0,0,6,256,0,0,0

\_\_all\_sys\_stat,5,name,16,0,1,0,0,6,256,0,0,0

\_\_all\_sys\_stat,5,cluster\_role,17,0,2,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,cluster\_id,18,0,3,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,server\_type,19,0,4,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,server\_role,20,0,5,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,server\_ipv4,21,0,6,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,server\_ipv6\_high,22,0,7,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,server\_ipv6\_low,23,0,8,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,server\_ip\_port,24,0,9,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,table\_id,25,0,10,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,data\_type,26,0,0,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,value1,27,0,0,0,0,1,8,0,0,0

\_\_all\_sys\_stat,5,value2,28,0,0,0,0,6,256,0,0,0

\_\_all\_sys\_stat,5,info,29,0,0,0,0,6,256,0,0,0

用户表的列schema

## \_\_all\_all\_join（空）

# schema := left\_table\_name,left\_table\_id,left\_column\_name,left\_column\_id,right\_table\_name,right\_table\_id,right\_column\_name,right\_column\_id

用户表的join\_info

## \_\_all\_sys\_param

# schema := name,cluster\_role,cluster\_id,server\_type,server\_role,server\_ipv4,server\_ipv6\_high,server\_ipv6\_low,server\_ip\_port,data\_type,value,info

ob\_app\_name,0,0,0,0,0,0,0,0,6,my app name,oceanbase app name

## \_\_all\_sys\_stat

# schema := name,cluster\_role,cluster\_id,server\_type,server\_role,server\_ipv4,server\_ipv6\_high,server\_ipv6\_low,server\_ip\_port,table\_id,data\_type,value1,value2,info

max\_used\_table\_id,0,0,0,0,0,0,0,0,0,1,1000,,max used table id