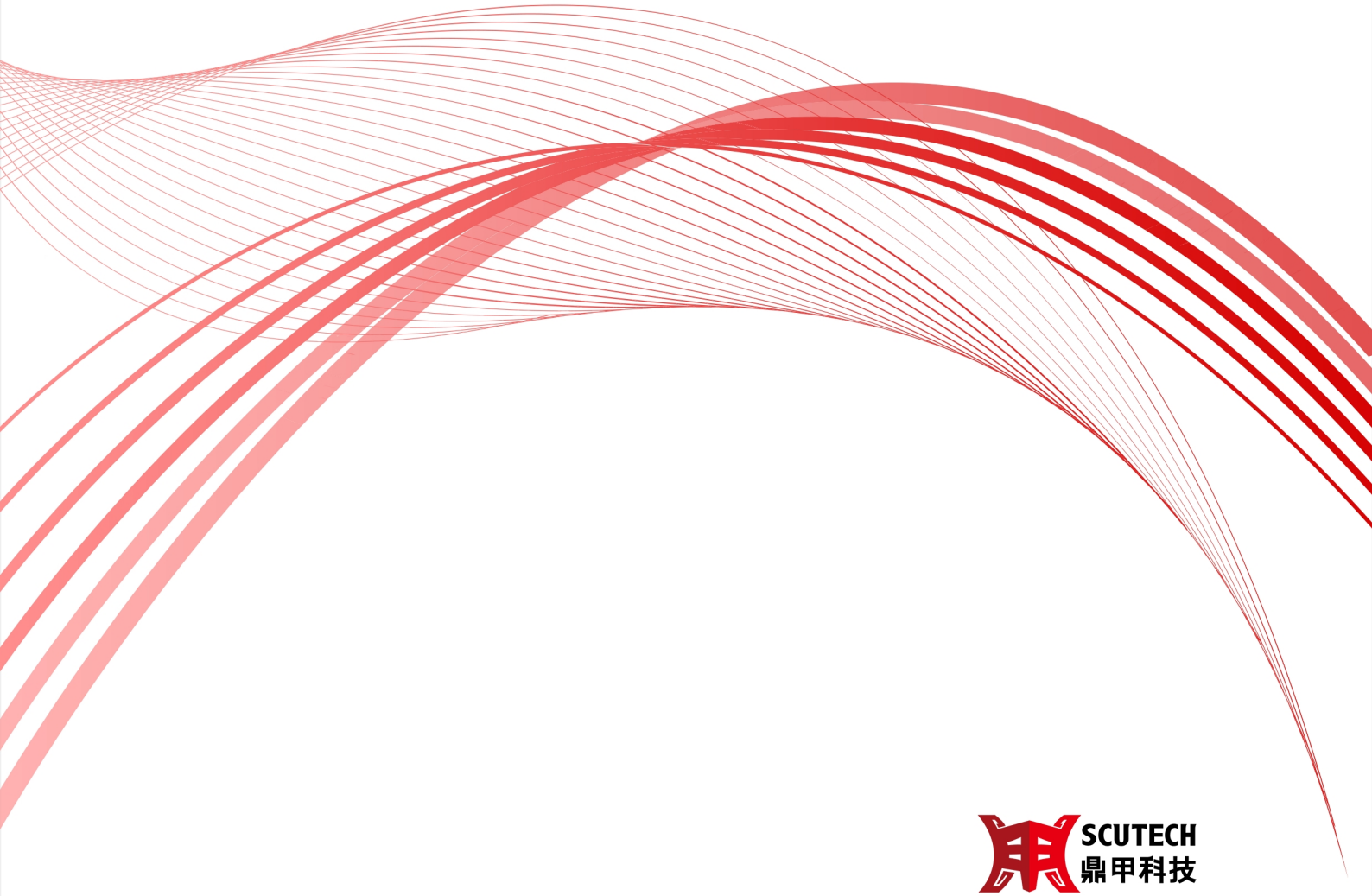
鼎甲迪备

Hive 备份恢复用户指南

Release V8.0-9

January, 2025



SCUTECH 鼎甲科技

目录

1 概述

该文档主要描述如何安装配置迪备代理端以及如何正确使用迪备备份和恢复 Hive。

迪备支持 Hive 备份恢复主要特性包括：

- 备份内容

数据库、用户

- 备份类型

完全备份、增量备份

- 备份目标

标准存储池、重删存储池、本地存储池、磁带库池、对象存储池、LAN-free 池等

- 备份策略

迪备提供 7 种备份计划，立即、一次、手动、每小时、每天、每周、每月

- 数据处理

数据压缩、数据加密、多通道、断点续传、限速、复制

- 恢复类型

数据库时间点恢复

- 恢复目标

原机恢复、异机恢复

- 恢复选项

备份主机、通道数、同名数据库处理方式、临时目录、后置重建表分区

2 计划和准备

在安装迪备代理端之前，需确保满足以下要求：

1. 确保所有备份组件都已安装和部署，包括备份服务器、存储服务器。
2. 迪备控制台创建一个至少具备操作员和管理员角色的用户，使用此用户登录迪备控制台并对资源进行备份恢复。

备注：管理员角色用于代理端安装和配置、激活许可证和授权用户。操作员角色用于创建备份和恢复作业、副本管理。

要实现 Hive 备份及恢复，需要在能与 Hadoop 和 Hive 通信的主机安装迪备代理端。

3.1 验证兼容性

在安装代理端之前，需先确保 Hive 的环境已在鼎甲迪备的适配列表中。

迪备支持多种版本 Hive 备份恢复。支持的版本主要有：

- Hive 3.1.2
- 华为 MRS(Hive) 3.1.0

3.2 安装迪备代理端

HBase 客户端安装方式与 Hadoop 文件系统一致，参考《Hadoop 备份恢复用户指南》的“安装迪备代理端”章节。

3.2.1 Apache Hadoop

安装 Hadoop 备份代理端前需要在代理端主机安装 Hadoop 运行环境。

备注：如果同时存在 2.x 和 3.x 版本的 Hadoop 集群，而且需要支持 Hive 3.x 备份与恢复，则应根据具体需求部署两个代理端。一个代理端部署 Hadoop 2.10.0 组件版本，该代理端支持 Hadoop 2.x 和 3.x，不支持 Hive 的备份与恢复功能；另一个代理端部署 Hadoop 3.2.1 组件版本，该代理端支持 Hadoop 3.x 和 Hive 3.x 的备份与恢复功能。

- 解压 Hadoop 和 Hive 运行环境离线包：

```
$ sudo tar -axf hadoop-3.2.1.tar.xz -C <dir>
$ sudo tar -axf apache-hive-3.1.2-bin.tar.gz -C <dir>
```

  解压出目录为：hadoop-3.2.1、apache-hive-3.1.2-bin
- 安装 OpenJDK：

```
$ sudo tar -axf Ubuntu20.04-OpenJDK11-AMD64.tar.gz
```

  解压出目录为：openjdk11 只需要安装 openjdk-11-jre-headless 环境即可

```
$ sudo dpkg -i openjdk11/*.deb
```

  备注：jre 的目录以及版本根据实际的安装情况，默认在 `/usr/lib/jvm/` 目录下

- 主机配置 Hadoop 客户端时，需执行 config 配置环境变量，以压缩包解压后目录‘/opt/hadoopclient’为例，

```
$ /etc/init.d/dbackup3-agent config hadoop
$ Configure Huawei MRS? [y/N] n
$ Please input JRE home []: /usr/lib/jvm/java-8-openjdk-amd64
$ Please input Hadoop home []:/opt/hadoopclient/hadoop-3.2.1
$ Configure Hive? [y/N] y
$ Please input Hive home []:/opt/hadoopclient/apache-hive-3.1.2-bin
$ Restarting dbackup3-agent (via systemctl):                  [  OK  ]
$ [ ok ] Restarting dbackup3-agent (via systemctl): dbackup3-agent.service.
```

3.2.2 FusionInsight MRS 集群环境

参考《Hadoop 备份恢复用户指南》的“*FusionInsight Manager* 集群环境”章节。

4 激活代理端许可证和分配授权

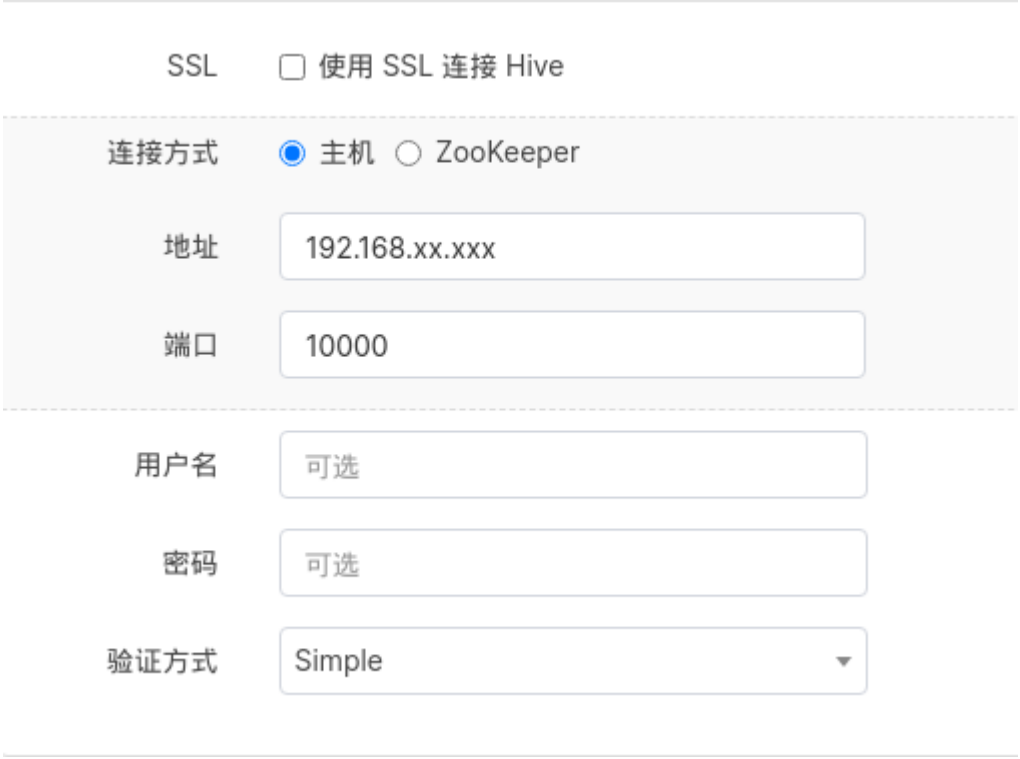
参考《Hadoop 备份恢复用户指南》的“激活代理端许可证和分配授权”章节。

5.1 添加 Hadoop 集群

参考《Hadoop 备份恢复用户指南》的“添加 *Hadoop* 集群”章节。

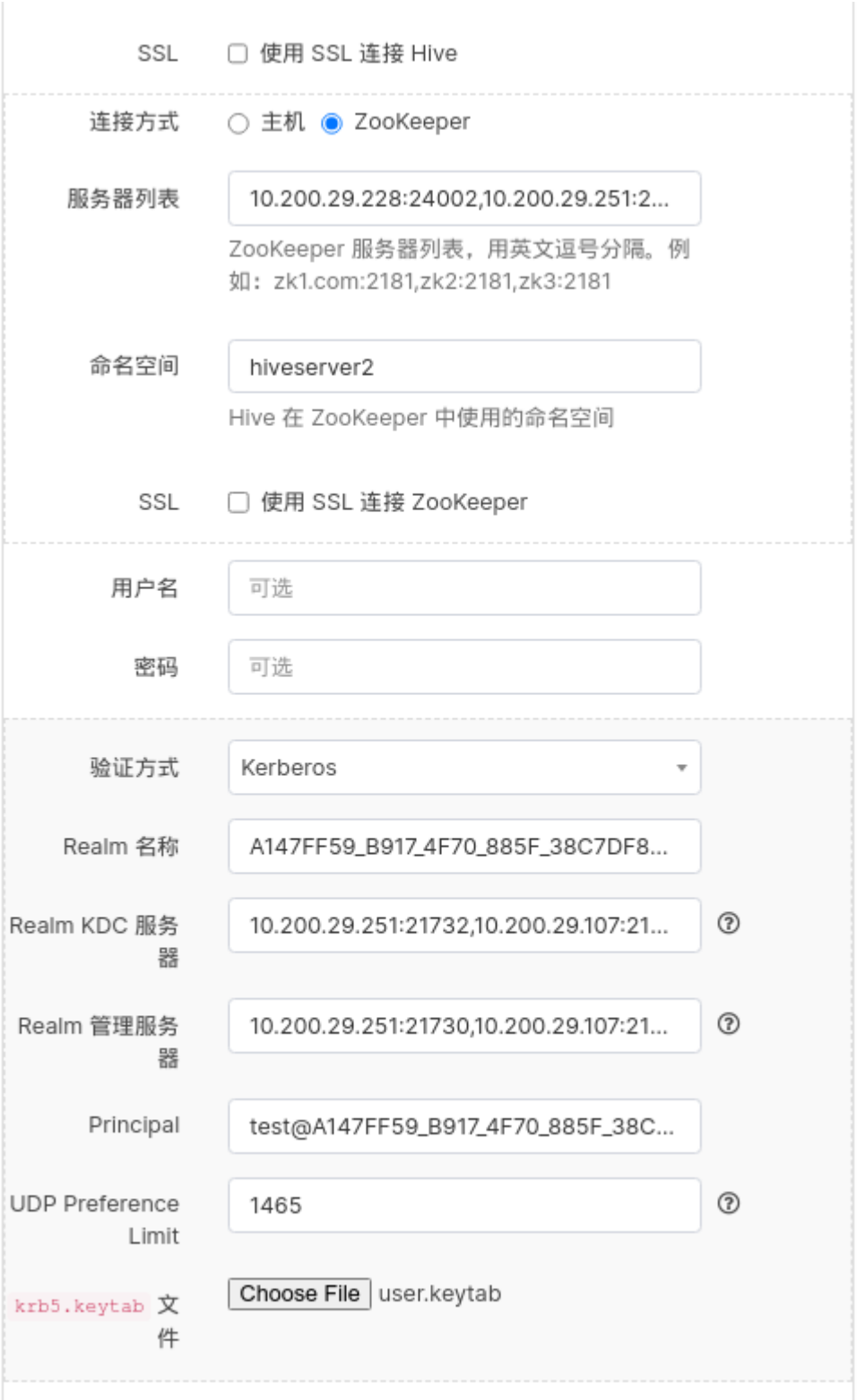
5.2 添加 Hive

（1）主机连接方式



- 安全连接：使用 SSL 安全连接。该选项需要 Hive 配置和启用 HTTPS 服务才能使用，否则不需要勾选该选项。
- 验证方式：选择主机。
- 名称：自定义资源名称。
- 主机：Hiveserver 所在的主机 IP 或主机名。如果配置 Kerberos 认证时 Principal 使用主机名进行创建，那么该字段需要填写主机名，并且代理端所在的机器的 hosts 文件也要添加该主机的 IP 及其对应的主机名解析。
- 安全连接：使用 SSL 安全连接。该选项需要 Hive 配置和启用 HTTPS 服务才能使用，否则不需要勾选该选项。
- 端口：默认为 10000，如果集群配置为其他端口那么该选项需要根据实际端口进行修改。
- 用户名：默认为空，用户如若有设置数据库用户则按实际填写。
- 密码：用户设置数据库用户时填写。

(2) ZooKeeper 连接方式



- 安全连接：使用 SSL 安全连接。该选项需要 Hive 配置和启用 HTTPS 服务才能使用，否则不需要勾选该选项。
- 验证方式：选择 ZooKeeper。
- 服务列表：ZooKeeper 服务器的列表，用英文逗号分隔。在 FusionInsight MRS 客户端主机上执行 echo $CLIENT_HIVE_URI 获取相关信息。
- 命名空间：Hive 在 ZooKeeper 中使用的命名空间。在 FusionInsight MRS 客户端主机上执行 echo $CLIENT_HIVE_URI 获取相关信息。
- 端口：默认为 10000，如果集群配置为其他端口那么该选项需要根据实际端口进行修改。
- 用户名：默认为空，用户如若有设置数据库用户则按实际填写。
- 密码：用户设置数据库用户时填写。

(3) Simple 验证方式和 Kerberos 验证方式与添加 Hadoop 集群一致。

5.3 激活 HDFS 和 Hive

1. 添加 Hadoop 集群和 Hive 成功后，会弹出【Hadoop 许可证】激活窗口，点击 HDFS、Hive 资源的**【激活】**按钮。
2. 激活后，点击 HDFS、Hive 资源**【授权】**按钮，进行授权。
3. 在**【授权】**窗口，可对 HDFS、Hive 资源进行授权用户操作。
   - 用户组：授权该资源给用户组。
   - 受保护：被标记为受保护的资源将无法用于恢复或数据复制的目标，除非管理员移除该标记。

备注：
1. 若提示“许可证不足”，需联系迪备管理员增加许可证。
2. 若已添加的集群参数发生变更，包括主机 IP、端口、验证方式等参数，用户可以通过点击**【设置】**对已添加的 Hadoop 集群和 Hive 进行修改。

6.1 备份类型

迪备为 Hive 备份提供完全备份、增量备份两种常规的备份类型。

- 完全备份

备份 Hive 实例下的所有数据。

- 增量备份

增量备份基于完全备份创建，备份相较于上次完全/增量备份以来改变的数据文件。若第一次增量备份前未做完全备份，则会做完全备份。

6.2 备份策略

迪备提供 7 种备份计划，立即、一次、手动、每小时、每天、每周、每月。

- 立即：作业创建后就执行。
- 一次：作业在指定时间执行一次。
- 手动：作业创建后可手动启动作业执行。
- 每小时：作业每天在设置的时间范围内以特定的小时/分钟间隔重复运行。
- 每天：作业以特定的天数间隔在特定时间重复运行。
- 每周：作业以特定的周数间隔在特定时间重复运行。
- 每月：作业在特定月份和时间重复运行。

针对用户的实际情况和需求，设置合理的备份策略。通常，推荐用户使用常规的备份策略：

1. 完全备份：每周在应用访问量较小的时间（例如周末）进行一次完全备份，以确保每周至少有一个可恢复的时间点。
2. 增量备份：每天在业务低峰期（例如凌晨 02:00）进行一次增量备份，可以更好地节省存储空间和备份时间，保证每天至少有一个可恢复的时间点。

6.3 开始之前

在备份恢复 Hive 之前，需保证已完成如下操作：

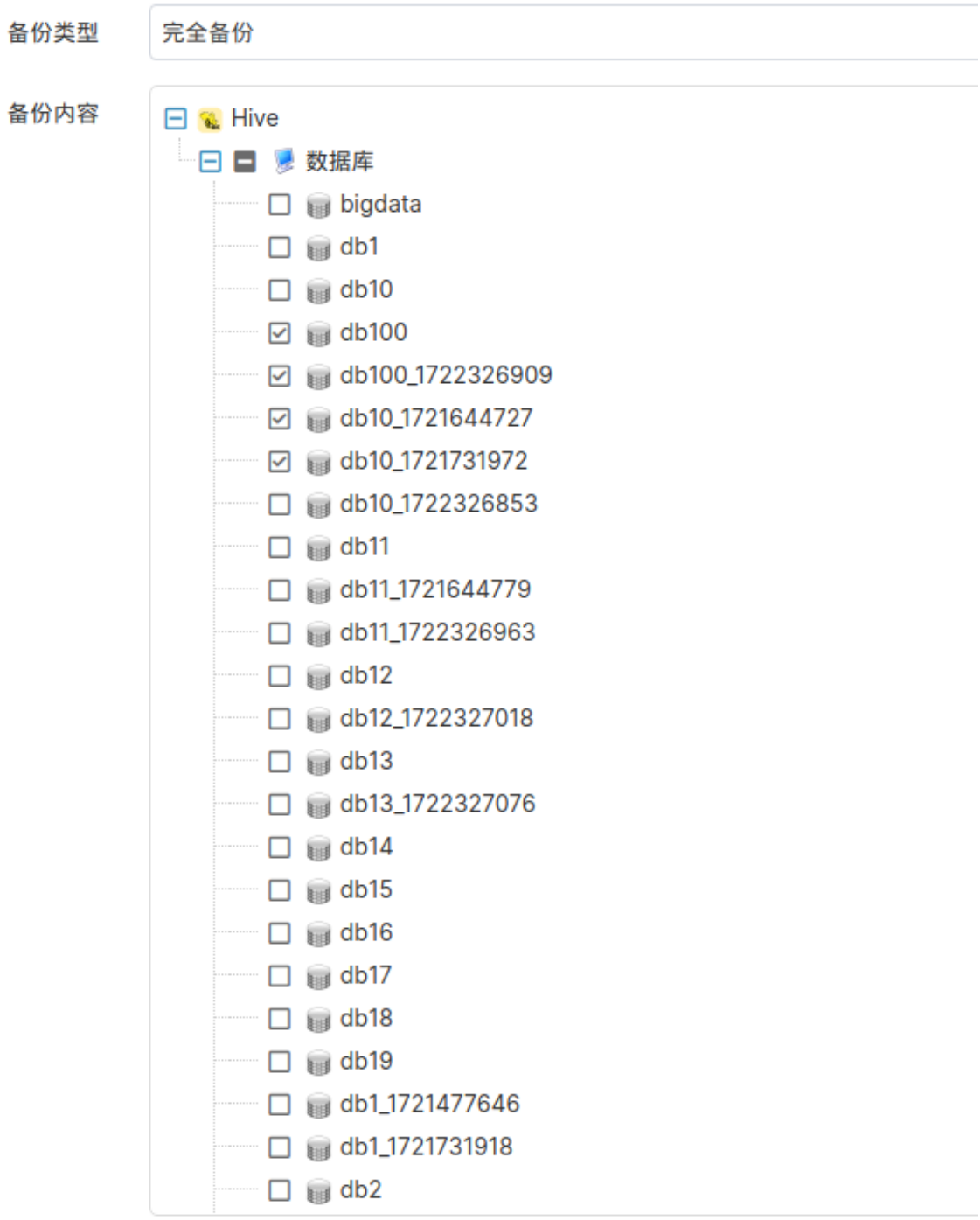
1. 检查存储池

（1）在迪备菜单栏中，点击**【存储池】**，进入**【存储池】**页面。

（2）检查展示区是否存在存储池。如果没有，需参考《管理员用户指南》的“创建存储池”，创建存储池并授权给当前控制台用户。

6.4 创建备份作业

1. 在菜单栏中，点击【**备份**】，进入【**备份**】页面。
2. 在【**主机和资源**】页面，选择 Hadoop 主机和 Hive 实例，自动跳转【**下一步**】。
3. 在【**备份内容**】页面，选择一个【**备份类型**】，勾选您希望备份的数据库，点击【**下一步**】。



（1）【**备份类型**】选择完全备份、增量备份。

> **备注：** 对于增量备份，【**备份内容**】步骤只需要选择完全备份作为基准，无需再次选择数据库和表。

（2）点击 + 可以展开数据库的表，勾选需要备份的表。

4. 在【**备份目标**】页面，选择一个备份主机和一个存储池，点击【**下一步**】。

备注: 增量备份没有【**备份目标**】步骤，由于的它们的【**备份目标**】与【**备份内容**】步骤中选择的基准完全备份相同。

5. 在【**备份计划**】页面，选择一个计划类型，参考备份策略。点击【**下一步**】。
   - 选择“立即”，作业创建后就执行。
   - 选择“一次”，设置作业的开始时间。
   - 选择“手动”，作业创建后可手动启动作业执行。
   - 选择“每小时”，设置开始时间和结束时间，用于指定作业一天内执行的时间范围。输入作业执行的时间间隔，单位可选择小时或分钟。
   - 选择“每天”，设置作业的开始时间。输入作业执行的时间间隔，单位为天。
   - 选择“每周”，设置作业的开始时间。输入作业执行的时间间隔，单位为周，并选择一周内具体执行的日期。
   - 选择“每月”，设置作业的开始时间。选择作业执行的月份。按每月的自然日，或每月的周选择具体日期。
6. 在【**备份选项**】页面，根据需要设置常规选项和高级选项，参考备份选项。点击【**下一步**】。

7. 在【**完成**】页面，设置【**作业名**】，并检查作业信息是否有误。点击【**提交**】。
8. 提交成功后，自动跳转到作业页面。您还可以对作业进行开始、修改、删除等管理操作。

6.5 备份选项

迪备为 Hive 提供以下备份选项:

- 常规选项

表 1：备份常规选项

功能	描述	限制性说明
压缩	默认启用快速压缩。 - 不压缩：备份过程中不压缩。 - 可调节：自定义压缩级别，需激活高级功能。 - 快速压缩：备份过程中压缩，使用快速压缩算法。	

续下页

表 1 – 接上页

功能	描述	限制性说明
通道数	开启该选项可提高备份效率。通道数默认为 1，选择范围为 1~255，单位为个。 一般建议跟 CPU 核心数一致，超过 CPU 核心数之后效率提高不明显。	该选项仅完全备份支持，增量备份和完全备份保持一致。
快照	开启该选项可进行 Hive 快照备份，默认不开启。	该选项仅完全备份支持，增量备份和完全备份保持一致。
数据库未配置复制策略时	开启该选项可选择发送警报信息并取消作业或者添加复制策略，Hive 3.x 以上不需配置选项即可进行备份	增量备份和完全备份保持一致。
重删模式	可选择代理端重删或服务端重删。选择代理端重删时，备份数据在代理端进行重删，仅传输唯一数据块至存储服务器；选择服务端重删时，备份数据先传输至存储服务器，再进行重删。为避免在处理重复数据块时（例如代理端压缩或加密）消耗代理端的计算资源，建议仅在首次备份或增量备份等重复数据较少的场景下使用服务端重删。	备份目标中选择存储池为重删池时出现该选项。

- 高级选项：

表 2：备份高级选项

功能	描述	限制性说明
断线重连时间	支持 1~60，单位为分钟。在设置时间内网络发生异常复位后作业继续进行。	
断点续传缓冲区	设置断点续传缓冲区大小，默认为 10 MiB。加大缓冲区将消耗更多物理内存，但在高吞吐量场景下加大缓冲区可避免断点续传失效。	
速度限制	可分时段限制数据传输速度或磁盘读写速度。单位为 KiB/s、MiB/s 或 GiB/s。	
前置条件	作业开始前调用，当前置条件不成立时中止作业执行，作业变成空闲状态。	
前置/后置脚本	前置脚本在作业开始后资源进行备份前调用，后置脚本在资源进行备份后调用。	

针对不同需求，迪备提供多种 Hive 的恢复方式，包括：

- 时间点恢复

当 Hive 的数据库数据丢失时，可以通过时间点恢复功能将 Hive 恢复到指定的时间点状态。Hive 时间点恢复支持本机和异机恢复，可以覆盖恢复、重命名恢复或跳过重名数据库恢复。

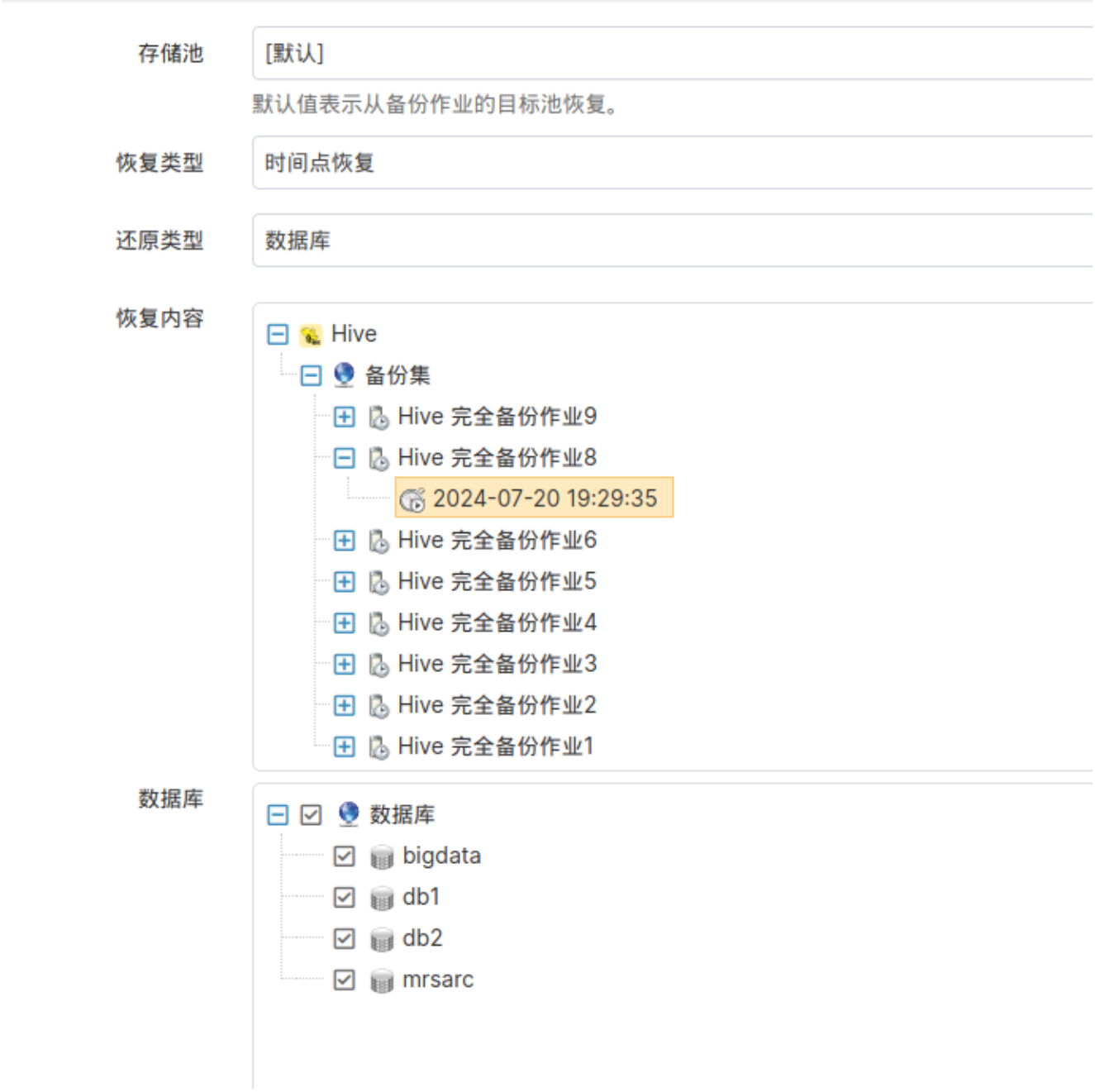
7.1 开始之前

如果要恢复 Hive 到其他主机，需先注册其他主机的 Hive 资源，激活许可证，并将 Hive 资源授权给当前迪备控制台用户。

7.2 创建数据库时间点恢复作业

创建数据库时间点恢复作业的步骤如下：

1. 在菜单栏中，点击**【恢复】**，进入**【恢复】**页面。
2. 在**【主机和资源】**页面，选择 Hive 所在主机和实例，自动跳转**【下一步】**。
3. 在**【备份集】**页面中，完成以下操作：



（1）**【存储池】**默认值表示从备份作业的目标池恢复，可选择任意已产生备份集的存储池。包括做池复制的目的池。

（2）**【恢复类型】**选择**时间点恢复**。

（3）**【还原类型】**选择**数据库**。

（4）在**【恢复内容】**列表中，选择需要恢复的备份集时间点。

（5）选择**【数据库】**。默认恢复备份集中的所有数据库，也可以手动“取消/勾选”选择恢复部分数据库。

4. 在**【恢复目标】**页面，支持恢复到本机异实例或异机实例。自动跳转**【下一步】**。
5. 在**【恢复计划】**页面，选择“立即”、“一次”或“手动”，点击**【下一步】**。
   - 选择“立即”，作业创建后就执行。
   - 选择“一次”，设置作业的开始时间。
   - 选择“手动”，作业创建后可手动启动作业执行。
6. 在**【恢复选项】**页面，参考恢复选项，根据所需进行设置。点击**【下一步】**。
7. 在**【完成】**页面，设置作业名称，并确认恢复内容。点击**【提交】**，等待作业执行。

8. 提交成功后，自动跳转到作业页面。您还可以对作业进行开始、修改、删除等管理操作。

7.3 恢复选项

迪备为 Hive 提供以下恢复选项：

- 常规选项：

表 3：恢复常规选项

功能	描述	限制性说明
备份主机	可以修改备份主机。默认为 Hive 实例设置的主机。	
通道数	开启该选项可提高恢复效率。通道数默认为 1，选择范围的最大值不能超过备份集最大的通道数，单位为个。	
同名数据库处理方式	可选择覆盖恢复、跳过重名数据库的恢复和新数据库添加时间戳后缀	
临时目录	可自主选择临时目录，用于存放恢复过程中产生的临时文件。恢复完成后，将自动清理此作业产生的临时文件。	
后置重建表分区	默认开启，在恢复 Hive 元数据时先忽略分区结构，待元数据恢复完成后再重建表分区。可避免 Hive 3.1.0 执行 REPL LOAD 时内存占用过高的问题。	
数据文件副本数	恢复时 Hive 数据文件在 HDFS 集群中存储的副本数量。auto：默认值 auto 代表恢复个数使用 HDFS 配置文件中的参数值。自定义：提供自定义恢复副本数范围为 1~3，减少副本个数可提升恢复速度并减少空间占用，但也增加了数据丢失的风险，请根据实际需求选择。	当 DataNode 的节点数少于 3 时，恢复出来的数据文件副本数只能小于等于 DataNode 的节点数

- 高级选项：

表 4：恢复高级选项

功能	描述	限制性说明
断线重连时间	支持 1~60，单位为分钟。在设置时间内网络发生异常复位后作业继续进行。	
断点续传缓冲区	默认为 10 MiB。设置断点续传缓冲区大小。加大缓冲区将消耗更多物理内存，但在高吞吐量场景下加大缓冲区可避免断点续传失效。	
速度限制	可分时段限制数据传输速度或磁盘读写速度。单位为 KiB/s、MiB/s 或 GiB/s。	

续下页

表 4 – 接上页

功能	描述	限制性说明
前置条件	作业开始前调用，当前置条件不成立时中止作业执行，作业变成空闲状态。	
前置/后置脚本	前置脚本在作业开始后资源进行恢复前调用，后置脚本在资源进行恢复后调用。	

8 限制性

表 5：限制性

功能	限制描述
数据文件副本数	不适用华为 MRS 环境，因为该环境默认禁止自定义数据文件副本数，并且无法解除此限制。

9 术语表

表 6：术语表

术语	说明
快速压缩	备份过程中压缩，使用快速压缩算法。

SCUTECH
鼎甲科技

全国销售热线：400-650-0081　　全国服务热线：400-003-3191

电话：+86 20 32053160　　网址：www.scutech.com

总部地址：广州市科学城科学大道243号总部经济区A5栋9楼