

Hive安装部署

www.huawei.com



目标

- 学完本课程后，您将能够：
 - 掌握**Hive**部署规划
 - 熟悉**Hive**常用配置参数
 - 掌握**Hive**备份容灾配置



目录

1. **Hive**部署规划
2. **Hive**关键参数配置
3. **Hive**监控介绍
4. **Hive**容灾配置

Hive内存要求和部署原则

服务名称	角色名称	内存最小要求	依赖关系	角色业务部署原则
Hive	HS (HiveServer)	4GB	依赖于HDFS、ZooKeeper、MapReduce、Yarn、DBservice	部署在控制节点上，集群模式，需要至少部署2个
	MS (MetaStore)	2GB		部署在控制节点上，集群模式，需要至少部署2个
	WH (WebHCat)	1GB		部署在控制节点上，集群模式，需要至少部署1个



目录

1. **Hive**部署规划
2. **Hive**关键参数配置
3. **Hive**监控介绍
4. **Hive**容灾配置

Hive关键参数配置路径

FusionInsight页面，单击 服务管理 > **Hive** > 服务配置，参数类别选择全部配置。

服务 > Hive 服务配置 >

服务状态 实例 **服务配置** 资源贡献排名

保存配置 导入服务配置 导出服务配置

参数类别: 全部配置

角色: HiveServer

主机:

HiveServer

- 集群
- 自定义
- JVM
- 日志
- MetaStore客户端
- 性能
- 安全
- 服务初始化
- 调优
- Hive数据仓库

参数	值
hive.tmp.file.cleanup.check.interval	1d
hive.tmp.file.cleanup.expire.time	3d
hive.zookeeper.session.timeout	180000

当前页: 1 总页数: 1 | < > | 跳转到: 确定

HiveServer关键参数配置

配置项	说明
HIVE_GC_OPTS	运行 HiveServer 进程的 JVM 参数，适当地调整改参数可以优化进程的性能。
hiveserver.auditlog.filesize	配置每个 HiveServer 审计日志文件大小的最大值。
hiveserver.auditlog.maxbackupindex	配置保留 HiveServer 审计日志文件的最大数量。
hiveserver.log.filesize	配置每个 HiveServer 运行日志文件大小的最大值。
hiveserver.log.level	配置 HiveServer 的日志级别。
hiveserver.log.maxbackupindex	配置保留 HiveServer 运行日志文件的最大数量。

MetaStore关键参数配置

配置项	说明
METASTORE_GC_OPTS	运行 MetaStore 进程的 JVM 参数，适当地调整该参数可以优化进程的性能。
Max ConnectionsPerPartition	配置 MetaStore 访问的底层数据库连接池大小，集群较大时可适当增大。
metastore.auditlog.filesize	配置每个 MetaStore 审计日志文件大小的最大值。
metastore.auditlog.maxbackupindex	配置保留 MetaStore 审计日志文件的最大数量。
metastore.log.filesize	配置每个 MetaStore 运行日志文件大小的最大值。
metastore.log.level	配置 MetaStore 的日志级别。
metastore.log.maxbackupindex	配置保留 MetaStore 运行日志文件的最大数量。

WebHCat 关键参数配置

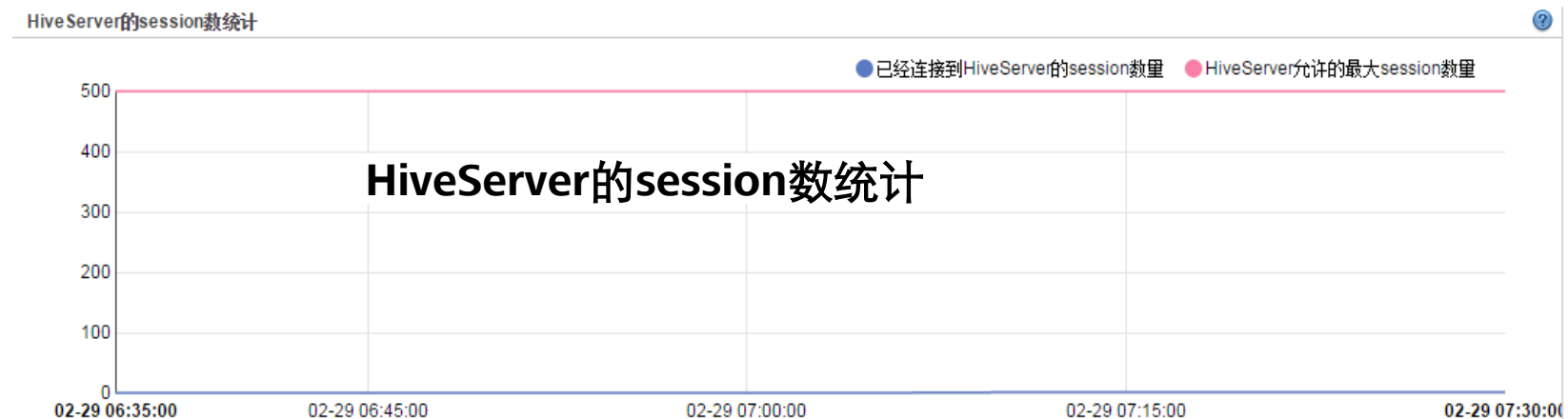
配置项	说明
WEBHCAT_GC_OPTS	运行 WebHCat 进程的 JVM 参数，适当地调整该参数可以优化进程的性能。
templeton.protocol.type	配置 WebHCat 的访问协议，可选 HTTPS 或 HTTP 。
webhcat.log.filesize	配置每个 WebHCat 运行日志文件大小的最大值。
webhcat.log.level	配置 WebHCat 的日志级别。
webhcat.log.maxbackupindex	配置保留 WebHCat 运行日志文件的最大数量。



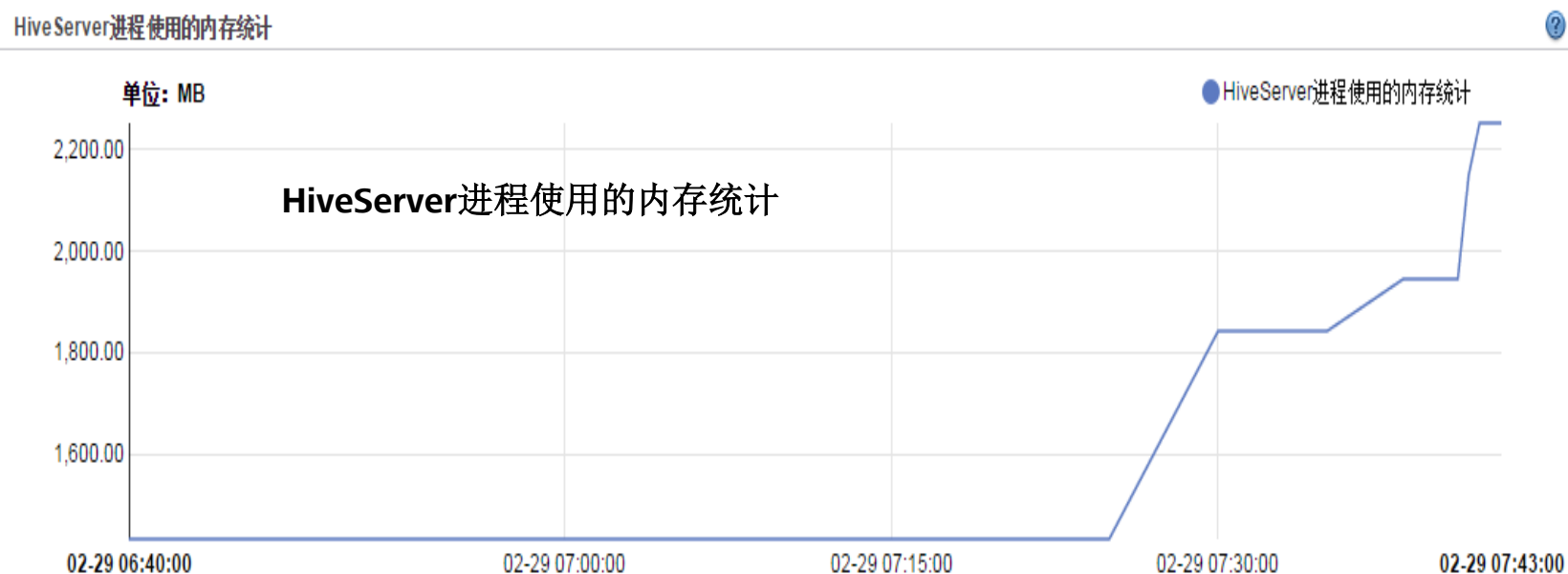
目录

1. **Hive**部署规划
2. **Hive**关键参数配置
3. **Hive**监控介绍
4. **Hive**容灾配置

HiveServer进程监控页面



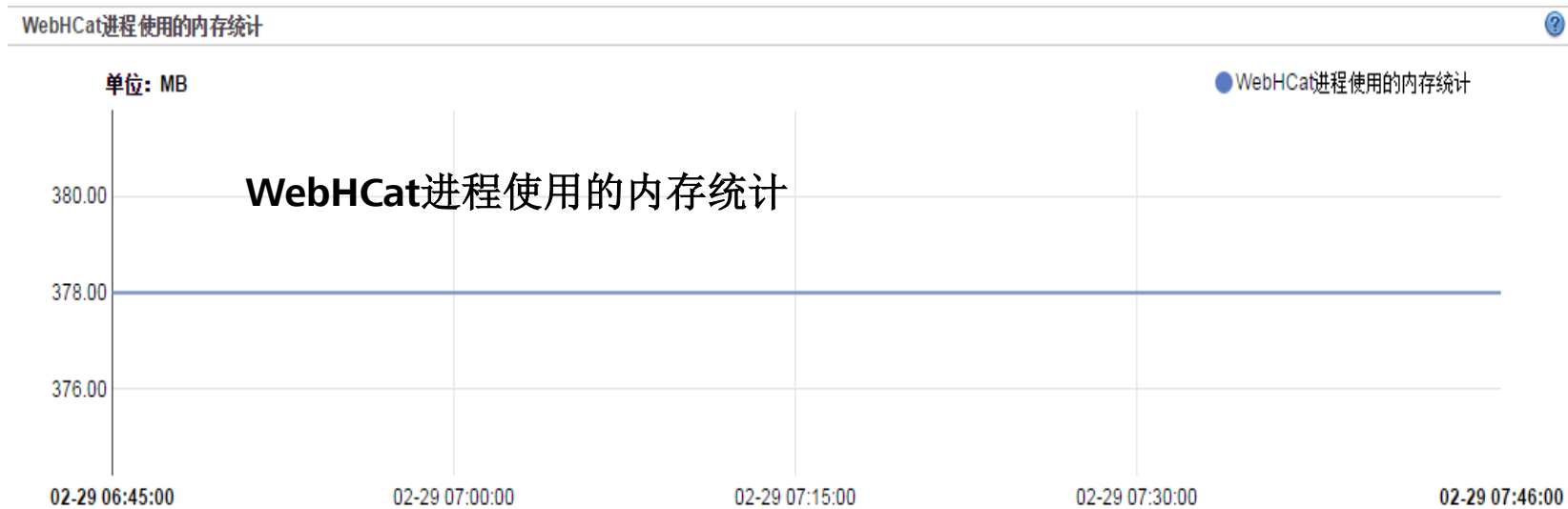
Hive Server进程监控页面



WebHCat进程监控页面



WebHCat进程监控页面





目录

1. **Hive**部署规划
2. **Hive**关键参数配置
3. **Hive**监控介绍
4. **Hive**容灾配置

配置Hive备份操作场景

为了确保**Hive**日常用户的业务数据安全，或者系统管理员需要对**Hive**进行重大操作（如升级或迁移等），需要对**Hive**数据进行备份，从而保证系统在出现异常或未达到预期结果时可以及时进行数据恢复，将对业务的影响降到最低。

该任务指导系统管理员通过**FusionInsight Manager**创建备份**Hive**任务。支持自动或手动备份数据。

Hive备份恢复功能不支持识别用户的**Hive**表、索引、视图等对象在业务和结构上存在的关联关系。用户需要根据业务场景管理统一的恢复点执行备份恢复任务，防止影响业务正常运行。

配置Hive备份前提条件

- 需要准备一个用于备份数据的备集群。
- 主备集群必须已配置跨集群互信。
- 主备集群必须已配置跨集群拷贝。
- 根据业务需要，规划备份任务的类型、周期、备份对象、备份目录和备份任务需要使用的**YARN**队列等策略规格。
- 检查备集群**HDFS**是否有充足的空间，备份文件保存的目录建议使用用户自定义的目录。
- 使用**HDFS**客户端，以“**hdfs**”用户执行**hdfs lsSnapshottableDir**检查当前集群中已创建**HDFS**快照的目录清单，确保待备份的数据文件所在**HDFS**路径的父目录或子目录不存在**HDFS**快照，否则无法创建备份任务。

配置Hive备份操作

创建备份任务

* 任务名称:

xxx

✓

备份类型:

☒ 周期备份

☐ 手动备份

* 开始时间:

✗ 开始时间不能为空。

* 周期:

☒ 按小时

☐ 按天

✗ 周期不能为空。

* 备份策略:

☒ 首次全量备份，后续增量备份

☐ 每次都全量备份

☐ 每n次进行一次全量备份

* 备份配置:

☐ DBService

☐ HBase

☐ HDFS

☐ NameNode

☒ Hive

目录配置

* 目录类型:

DistCp

* NameNode IP地址:

* 备份路径:

* 最大副本数:

* 队列名称:

数据配置

* 最大恢复点个数:

* 备份对象:

正则表达式输入

清除选中节点

☐ default

☐ tpcds_orc_hive_20000

配置Hive备份操作

- 创建备份任务
 - 在**FusionInsight Manager**，选择“系统设置 > 备份管理”。
 - 单击“创建备份任务”。
- 设置备份策略
 - 在“任务名称”填写备份任务的名称。
 - 在“备份类型”选择备份任务的运行类型：“周期备份”表示按周期自动执行备份，“手动备份”表示由手工执行备份。

创建周期备份任务，还需要填写以下参数：

“开始时间”：表示任务第一次启动的时间。

“周期”：表示任务下次启动与上一次运行的时间间隔，支持“按小时”或“按天”。

“备份策略”：表示任务每次启动时备份的数据量。支持“首次全量备份，后续增量备份”、“每次都全量备份”和“每n次进行一次全量备份”。选择“每n次进行一次全量备份”时，需要指定n的值。

配置Hive备份操作

- 选择备份源

- 在“备份配置”，勾选“**Hive**”。
- 在“目录类型”，选择“**DistCp**”。
- 在“**NameNode IP地址**”填写备集群**NameNode**业务平面**IP**地址，支持主节点或备节点。
- 在“备份路径”填写备集群保存备份数据的**HDFS**目录。不支持填写**HDFS**中的隐藏目录，例如快照或回收站的目录；也不支持默认的系统目录，例如“**/hbase**”或“**/user/hbase/backup**”。
- 在“最大副本数”填写备份目录中可保留的备份文件集数量。已备份的文件集数量大于“最大副本数”参数值时，默认保留最新的文件。
- 在“队列名称”填写备份任务执行时使用的**YARN**队列的名称。需和集群中已存在且状态正常的队列名称相同。

配置Hive恢复数据操作场景

操作场景

在用户意外修改删除、数据需要找回，系统管理员对**Hive**进行重大操作（如升级、重大数据调整等）后系统数据出现异常或未达到预期结果，模块全部故障完全无法使用，或者迁移数据到新集群的场景中，需要对**Hive**进行恢复数据操作。

该任务指导系统管理员通过**FusionInsight Manager**创建恢复**Hive**任务并恢复数据。只支持创建任务手动恢复数据。

Hive备份恢复功能不支持识别用户的**Hive**表、索引、视图等对象在业务和结构上存在的关联关系。用户在执行备份恢复任务时，需要根据业务场景管理统一的恢复点，防止影响业务正常运行。

配置Hive恢复数据前提条件和影响


前提条件








- 主备集群必须已配置跨集群互信。
- 主备集群必须已配置跨集群拷贝。
- 规划好恢复数据保存表的数据库，数据表在**HDFS**的保存位置，以及访问恢复数据的用户清单。
- 检查**Hive**备份文件保存路径。
- 停止**Hive**的上层应用。
- 登录**FusionInsight Manager**。

对系统的影响

- 恢复过程中会停止用户认证，用户无法开始新的连接。
- 数据恢复后，会丢失从备份时刻到恢复时刻之间的数据。
- 数据恢复后，需要重新启动**Hive**的上层应用。

配置Hive恢复数据操作

 FusionInsight Manager

 系统概览  服务管理  主机管理  告警管理  审计管理  租户管理  系统设置


系统设置 > 恢复管理 > 恢复配置 >


恢复配置


* 任务名称:

☐ DBService
☐ HBase
☐ HDFS
☐ NameNode
☒ Hive


目录配置

* 目录类型: 

* NameNode IP地址:  NameNode IP地址不能为空。

* 备份路径:  备份路径不能为空。

* 队列名称:

* 恢复点列表:  刷新

☐ LdapServer
☐ OMS



思考题

1. **Hive** 所依赖的服务有哪些?



习题

- 判断题

1、 **Hive**备份恢复功能不支持识别用户的**Hive**表、索引、视图等对象在业务和结构上存在的关联关系。（）



本章总结

本章介绍了**Hive**的安装部署规则，以及关键参数的配置，然后介绍了**Hive**健康状态所需监控项，最后介绍了**Hive**容灾的配置方法。

Thank you

www.huawei.com