

云数据库 HBase 版

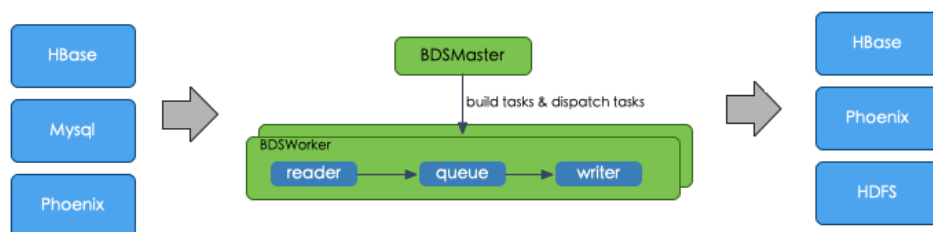
HBase数据导入导出

HBase数据导入导出

BDS服务介绍

BDS 是阿里云针对HBase自主研发的一套迁移同步服务，主要帮助云上的客户进行自建HBase、云HBase集群的数据导入和导出。

概况



BDS可以用于HBase集群的无缝迁移、主备容灾、异地多活、在线离线业务分离、HBase数据归档、对接RDS实时增量数据等等，方便云上客户围绕HBase构建高可用、灵活的业务系统。

对BDS实现原理及优势感兴趣的朋友可以阅读《BDS - HBase数据迁移同步方案的设计与实践》

功能

功能	场景	使用链接
HBase <-> HBase	新旧集群无缝迁移、集群升配、在线离线业务分离、主备容灾、异地多活	HBase集群迁移 HBase多集群实时数据同步
RDS -> HBase、Phoenix	在线离线业务分离、历史库	RDS实时数据同步HBase、Phoenix
HBase 数据导出 ODPS(MaxCompute)	支持历史数据和增量数据	HBase历史数据导出ODPS HBase实时数据归档ODPS
HBase 数据导出 RocketMQ	支持HBase实时增量数据同步到MQ	HBase实时数据同步MQ

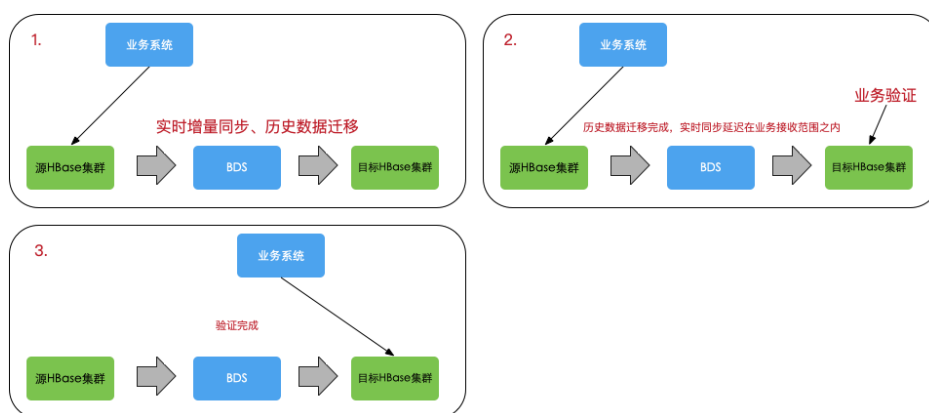
典型场景

新旧集群无缝迁移

使用范围

- HBase大版本升级，1.x 升级 2.x
- 集群配置升级
- 集群网络的切换，经典切换VPC
- 异地跨机房迁移
- 业务拆分

迁移步骤

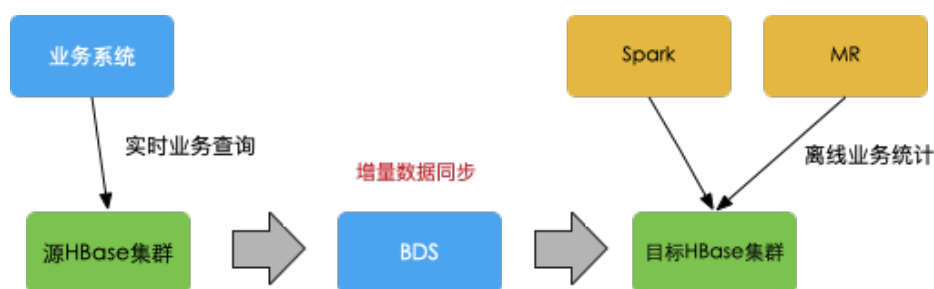


特点

1. 不停服数据迁移，同时搞定历史数据迁移、实时增量数据的同步。
2. 迁移过程中不会和源集群的HBase交互，只读取源集群的HDFS，尽可能减少对源集群在线业务的影响
3. 文件层的数据拷贝比通常API层的数据迁移通常能节省50%以上的流量
4. 高效性，单个节点迁移速度可达到100MB/s，节点数支持水平扩展，能够满足支持TB、PB级别的数据迁移
5. 稳定性，有完善的错误重试机制，实时监控任务速度和进度，支持任务失败报警
6. 正确性，能够进行数据校验
7. 支持自动同步Schema，保证分区一致

在线离线业务分离

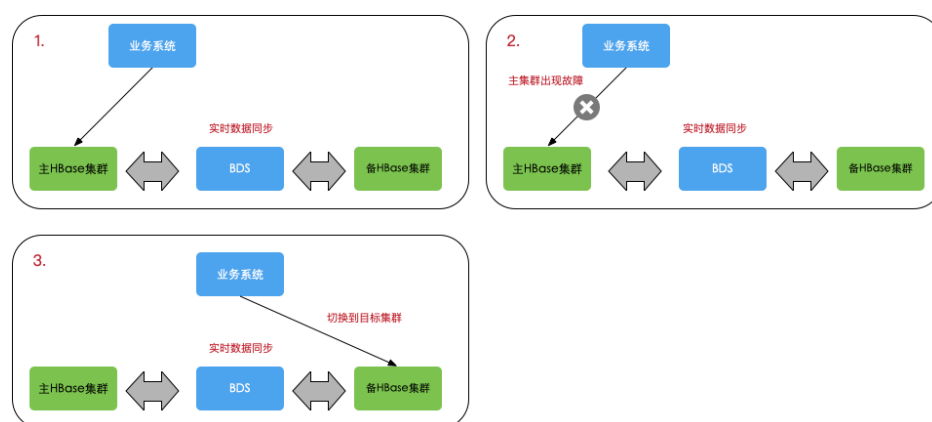
业务架构



通过BDS，将在线业务数据实时同步到离线集群，结合Spark、MR等大数据组件进行离线的数据分析，从而不影响在线业务的查询

主备容灾

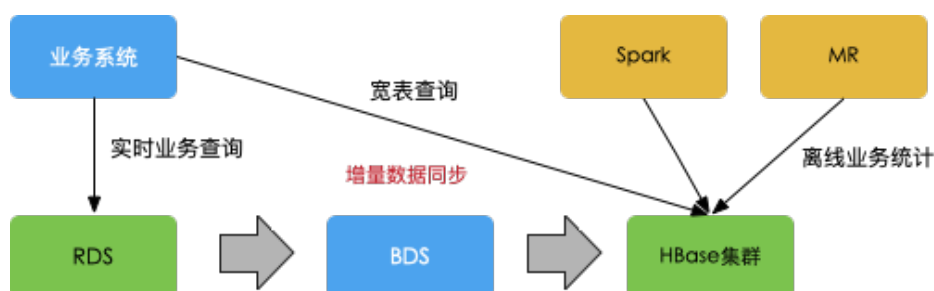
业务流程



通过BDS在主备集群之间实现实时数据的双向同步，当主集群出现问题，可以切换到备集群，尽可能减少对业务产生的影响，当主集群服务恢复之后，通过BDS同步服务补齐主集群遗漏的数据

RDS实时数据同步到HBase、Phoenix

业务架构



通过BDS将RDS数据实时同步到HBase集群

- 结合Spark、MR等大数据组件进行离线的数据分析，从而不影响在线业务的查询
- 利用HBase支持动态列的功能，将mysql中多张表构建一张HBase的大宽表，避免在RDS中进行join查询
- 做RDS的历史库

BDS服务开通

BDS同步服务购买标准

节点迁移能力

单节点(4cores 8G)的迁移能力

1. 单机迁移最大速度 50MB/s ~ 100MB/s
2. 历史数据迁移的线程数默认为5，表示同时有5个线程在拷贝HBase文件
3. 增量同步任务单机能同时处理5个日志, 日志迁移速度最大50 ~ 60MB/s

BDS服务规模推荐

迁移数据量	同步服务节点配置	同步服务节点数
$n < 1T$	4 cores 8G	1
$1T < n < 10T$	4 cores 8G	2 ~ 5
$10T < n$	4 cores 8G	5 以上

BDS开通步骤

服务购买

云数据库 HBase 版

集群版预付费

集群版后付费

单节点预付费

单节点后付费

云HBase正式支持Spark，满足复杂分析及流式处理需求。

服务

HBase

Spark(分析)

OpenTSDB(时序)

Ganos(时空)

GraphDB(图)

BDS(数据同步)

地域

华东1 (杭州)

华东2 (上海)

华北2 (北京)

华南1 (深圳)

华北1 (青岛)

不同地域之间的产品内网不互通；订购后不支持更换地域。请谨慎选择

可用区

华东2 可用区 D

版本

1.0

网络类型

专有网络

请确保与需要连接ECS在同一网络类型，创建后不能切换网络类型！

vpcId

solr-sh-vpc

vswitchId

solr-sh-vswitch

绑定vpc，vpc下的HBase集群可以相互迁移

集群名称

长度为2-128个字符，以大小写字母或中文开头，可包含数字、"."、"_"或"-"

当前配置

服务

BDS(数据同步)

地域

华东2 (上海)

可用区

华东2 可用区 D

版本

1.0

网络类型

专有网络

vpcId

solr-sh-vpc

vswitchId

solr-sh-vswitch

集群名称

-

Core规格

4 核 8 GB (独享规格)

core节点数量

1

core节点数量: 1

配置费用：
¥0.000 /小时

省¥2.070 /小时

内部结算用户0元付

立即购买

Core规格

4 核 8 GB (独享规格)

core节点数量

1

core节点数量: 1

单次最多可开通20个节点，开通后可再次添加节点。

购买步骤

- 在云HBase的购买页面有BDS数据同步服务的购买
- 选择vpc所在的区域，绑定指定的vpc
- 选择购买节点的规格，以及节点的数量

购买成功后

在控制台集群列表就能看到开通的BDS迁移服务的集群

HBase控制台									
集群列表									
全部 云HBase集群 云HBase实例 云HBase节点									
ID / 名称	服务	版本	类型	状态	付费类型	网络类型	创建时间		
hbase-uf64016z0ycp1419p(BDS)	BDS	1.0	集群	运行中	按量付费	专有网络	2019年4月10日 17:50:53		
hbase-uf64016z0ycp1419p(BDS)	BDS	1.0	集群	运行中	包年包月	专有网络	2019年4月10日 11:36:29		

页面登陆

添加白名单

将需要访问的机器ip添加到集群的白名单中

管理控制台

搜索

消息 10 费用 工单

<

实例 hb-uf64016z0ycp1419p(BDS) 运行中

基本信息

修改网络白名单

数据库连接

IP地址

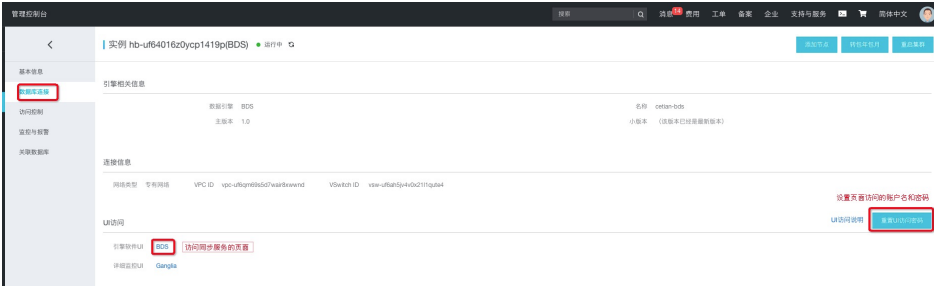
访问控制

0.0.0.0/0

监控与报警

关联数据库

设置登陆账户名密码并登陆页面



BDS网络环境打通

网络环境

源集群网络环境	目标集群网络环境	BDS服务是否支持	网络如何打通
VPC	相同VPC	是	1. BDS控制台关联HBase集群
VPC	不同的VPC	是	1. VPC之间需要建立一个高速通道 2. HBase集群添加BDS集群访问白名单 3. 开通集群HDFS（仅云HBase需要）
经典网络	相同的经典网络	否（如有需求请联系侧田）	1. HBase集群添加BDS集群访问白名单 2. 开放HDFS、HBASE的端口
经典网络	VPC	是	1. 需要创建 VPC 的 ClassicLink，经典网络的 esc 需要绑定对应的 ClassicLink 2. HBase集群添加BDS集群访问白名单 3. 开通集群的 HDFS（仅云HBase需要）
经典网络	云HBase经典网络	否（如有需求请联系侧田）	1. 需要临时购买几台经典网络ECS 2. HBase集群添加BDS集群访问白名单 3. 开通集群HDFS（仅云HBase需要）
云HBase经典网络	VPC	否（如有需求请联系	1. 需要临时购买几台

		侧田)	经典网络ECS做中转，迁移完可以释放，ECS和VPC HBase属于一个地域（比如华东2、华东1） 2. ECS和VPC打通走Classiclink 3. HBase集群添加BDS集群访问白名单 4. 开通集群HDFS（仅云HBase需要）
--	--	------	---

注意事项

- 高速通道和ClassicLink是创建是有条件约束的，请认真阅读相关文档

迁移涉及的端口

迁移过程访问HDFS和HBase涉及的端口

类型	默认端口	说明
fs.default.name	8020	Namenode RPC 交互端口
dfs.http.address	50070	NameNode web 管理端口
dfs.datanode.ipc.address	50020	Datanode的RPC服务器地址和端口
dfs.datanode.address	50010	datanode控制端口
zookeeper.property.clientPort	2181	zk的地址
hbase.regionserver.port	16020	HBase RegionServer绑定的端口
hbase.master.port	16000	HBase Master绑定的端口

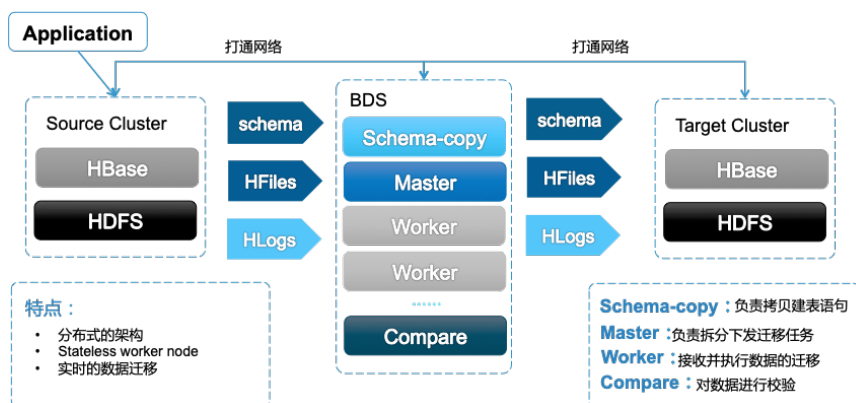
HBase集群迁移

使用场景

1. 大版本升级，HBase1.x 升级 HBase2.x

2. 跨地域迁移，例如从青岛机房迁移到北京机房
3. HBase1.x 升级 HBase2.x
4. 集群升配，例如从4核8G迁移到8核16G集群
5. 业务拆分，将部分业务迁移到新的集群

BDS迁移架构



- BDS采用分布式的架构，由Master节点生成、调度任务给各个worker节点
- worker节点负责具体的任务执行
- BDS支持表meta信息、历史数据、增量的日志数据从源端迁移到目标端

BDS迁移优势

1. 不停服数据迁移，同时搞定历史数据迁移、实时增量数据的同步。
2. 迁移过程中不会和源集群的HBASE交互，只读取源集群的HDFS，尽可能减少对源集群在线业务的影响
3. 文件层的数据拷贝比通常API层的数据迁移通常能节省50%以上的流量
4. 高效性，单个节点迁移速度可达到100MB/s，节点数支持水平扩展，能够满足支持TB、PB级别的数据迁移
5. 稳定性，有完善的错误重试机制，实时监控任务速度和进度，支持任务失败报警
6. 正确性，能够进行数据校验
7. 支持自动同步Schema，保证分区一致

版本支持

- Hadoop 版本：2.x
- HBase：1.x <-> 1.x | 1.x -> 2.x | 2.x <-> 2.x（不支持开启Kerberos集群和单节点云HBase实例）
- Phoenix：4.x <-> 4.x | 5.x <-> 5.x（跨版本可能存在兼容问题）

迁移步骤

- 购买及环境准备

- 客户 - 购买BDS迁移服务，并绑定指定VPC（VPC下的HBase集群可以相互迁移）
- 协商 - 环境打通（详见网络环境打通）

- 迁移

- 客户 - 登陆BDS操作页面
- 客户 - 添加HBase数据源
- 客户 - 创建增量同步通道，同步表的实时增量写入数据到目标集群
- 客户 - 提交历史数据迁移任务，迁移历史存量数据

验证

- 客户 - 验证
- 客户 - 业务切换

数据迁移过程中为了保证数据不丢失，优先开启增量数据的同步，增量同步通道稳定后提交历史数据迁移

注意事项

1. 增量同步不支持bulkload的数据
2. 迁移前请确认目标集群的HDFS容量，防止迁移过程中出现容量写满的情况
3. 增量同步提交前，推荐修改一下源集群的日志保留时间，给增量同步出错预留一些处理的时间（hbase-site.xml 的 hbase.master.logcleaner.ttl 调大为12小时以上，重启HMaster）
4. 关于业务切换，当增量同步的时延比较小（几秒或者几百毫秒），业务不敏感的话，源集群可以不停写，直接切换业务，同步服务会快速补齐数据。
5. 客户无需在目标集群创建表，BDS同步服务会自动创建和源集群一样的表，包括分区的信息。客户自建的目标表可能会和源表分区很不一致，这可能会导致迁移完成之后，目标表会进行频繁的split、compaction，如果表的数据量十分庞大，可能会导致这个过程非常耗时
6. 如果源表带有 coprocessor，在创建目标表的时候可能会报错，因为目标集群不包含对应的 coprocessor，遇到此类问题请钉钉联系侧田
7. 对于大表的数据校验，抽样比例不宜过大，不然会对目标集群的业务产生一定的影响

HBase多集群实时数据同步

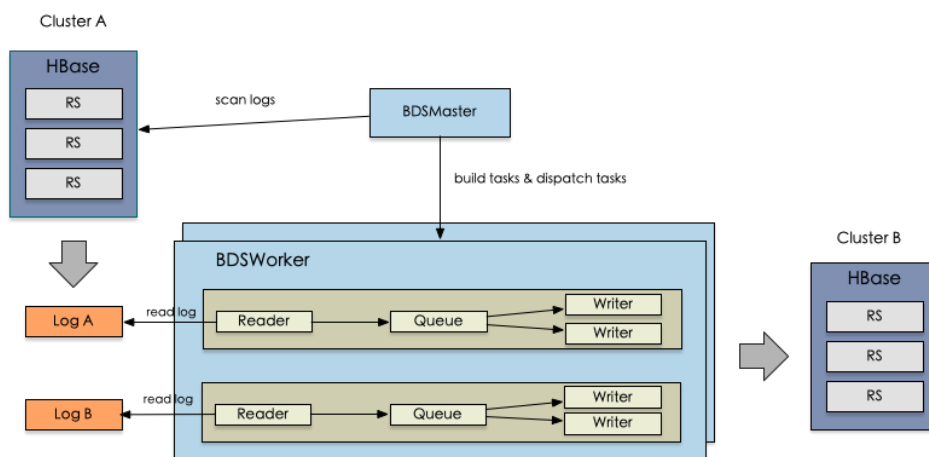
本文主要介绍HBase实时数据同步的使用场景、BDS实时同步的原理以及与HBase Replication相比的优势，帮助云上HBase客户打通多集群之间长期、稳定的实时数据同步的通道

使用场景

1. 主备容灾，主集群有故障，可以快速切换到备集群，混合云部署

2. 异地多活，跨地域机房的数据同步
3. 在线离线业务分离
4. 新旧集群不停服迁移

BDS实时数据同步架构



- BDS采用分布式的架构，由BDSMaster节点扫描源集群的日志并生成、调度任务给各个BDSWorker节点
- BDSWorker节点接收到对应的Task会创建TaskTunnel用于同步对应的Log
- TaskTunnel 包含 Reader、Queue、Writer，Reader 负责读取 log并将读取的log数据丢到queue中，writer异步消费queue中的数据发往目标集群

BDS数据同步 VS HBase Replication

HBase Replication 劣势

- 数据同步相关逻辑集成在 Region Server 中
 - bug fix，系统更新困难
 - Region Server 读写和数据同步相互影响
 - 同步积压，不太容易通过扩展节点来解决
- 不同Region Server同步压力不一样
 - Region Server宕机，日志会被另外一台Region Server接管
 - 写热点，写热点的RegionServer会出现同步不过来的现象
- 不支持跨版本的数据同步
- 没有一套完善的UI监控页面，难定位同步问题

BDS实时同步的优势

- 功能性
 - 支持跨版本的同步

- 支持双向同步、环状同步
- 稳定性
 - 抗热点，同步的任务平均分配到各个BDSWorker节点
 - 对业务影响更小
 - 独立于HBase，不受RegionServer读写影响
 - 同步过程基本不会访问源HBase集群，只读取源HDFS上的日志
 - 可以控制同步的线程数和节点的并发数
 - 更加完善的failover机制
- 易运维
 - BDSWorker stateless，便于扩展、系统升级
 - 完善的监控、报警机制
 - 监控同步时延，上下游流量，读写耗时等指标
 - 任务失败，延迟过大报警
- 性能
 - 可以通过扩展节点数量来提升迁移同步的速度
 - 支持动态配置线程数
 - 读写异步

版本支持

- Hadoop 版本：2.x
- HBase：1.x <-> 1.x | 1.x <-> 2.x | 2.x <-> 2.x（不支持开启kerberos的集群、单节点云HBase）
- Phoenix：4.x <-> 4.x | 5.x <-> 5.x（暂不支持跨版本）

购买推荐规格

一个BDSWorker（4cores 8G）默认能同时处理5个日志，同步节点个数 $\approx$ 源集群RS节点 / 5 + 1

同步步骤

- 购买及环境准备
 - 客户 - 购买BDS迁移服务，并绑定指定VPC（VPC下的HBase集群可以相互迁移）
 - 协商 - 环境打通（详见网络环境打通）
- 迁移
 - 客户 - 登陆BDS操作页面
 - 客户 - 添加HBase集群
 - 客户 - 修改源集群日志保留策略
 - 自建HBase、EMR HBase - hbase-site.xml配置的hbase.master.logcleaner.ttl调大为12小时以上
 - 云HBase - 开启日志订阅功能
 - 客户 - 创建增量同步通道，同步表的实时增量写入数据到目标集群

注意事项

- 暂不支持开启Kerberos的集群
- 增量数据的迁移是基于读取HBase WAL并在目标集群进行回放来实现的，不记录到WAL的数据是不能同步到目标集群的，Bulkload加载到表的数据是不会同步到目标集群的
- 自建HBase、EMR HBase集群增量同步提交前推荐修改一下源集群的日志保留时间，给增量同步出错预留一些处理的时间（hbase-site.xml 的 hbase.master.logcleaner.ttl 调大为12小时以上，重启HMaster ）

Apsara HBase 开启日志订阅功能

< 实例 hb-44n57sq2kg5ai4048(hbase) ● 湖广中 Q 立即开通 了解版本 购买实例 修改实例							
可修改参数 修改历史							
参数名	参数默认值	运行参数值	单位	是否需要重启	可修改参数值	参数描述	操作
hbase.dynamic.jars.enabled	false	false	STR	是	[true/false]	Enable dynamic class lib...	修改
hbase.exporter.enabled	false	True	STR	是	[*]		修改
hbase.region.majorcompact...	604800000	604800000	INT	是	[0-604800000]	The time (in milliseconds) ...	修改
hbase.zoo.server.callqueue.re...	0.0	0.0	FLLOAT	是	[0.0-0.7]	Split the call queue into ...	修改
hbase.zoo.server.callqueue.s...	0.0	0.0	FLLOAT	是	[0.0-0.7]	Given the number of mast...	修改
hbase.regionserver.global.me...	0.0	0.0	FLLOAT	是	[0.0-479]	Maximum size of all mem...	修改
hbase.regionserver.global.me...	0.35	0.35	FLLOAT	是	[0.0-0.5]	Maximum size of all mem...	修改

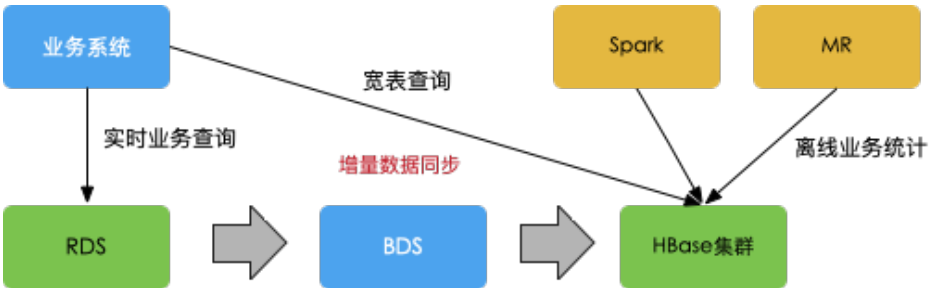
- 设置 hbase.exporter.enabled 为 true
- 滚动重启源集群

RDS实时数据同步HBase、Phoenix

本文主要介绍RDS实时数据如何同步到自建HBase或者云HBase

使用场景

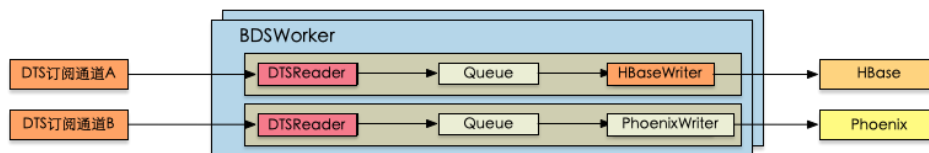
业务架构



通过BDS将RDS数据实时同步到HBase集群

- 结合Spark、MR等大数据组件进行离线的数据分析，从而不影响在线业务的查询
- 利用HBase支持动态列的功能，将mysql中多张表构建成一张HBase的大宽表，避免在RDS中进行join查询
- 做RDS的历史库

BDS实时数据同步架构



- DTS服务提供对RDS binlog收集，实时抓取源RDS实例的binlog，并提供实时数据的订阅
- BDS同步服务订阅DTS所收集的数据，将实时的数据写入到HBase或者Phoenix

RDS -> HBase 实时数据同步

版本支持

Mysql版本：

- 支持公共云、金融云RDS For MySQL实例、DRDS实例的数据订阅。
- 支持经典网络、VPC网络下RDS For MySQL实例、DRDS实例的数据订阅。

HBase版本：1.x、2.x（不开启Kerberos自建HBase、云HBase）

优势

- 一个同步通道可以支持多张表
- 支持多个RDS上的多张表同步到一张HBase表
- 单节点(4 cores 8G)，同步的吞吐最高可达20 ~ 30MB/s

同步步骤

- 购买及环境准备
 - 客户 - 开通DTS订阅通道
 - 客户 - 创建DTS消费通道
 - 客户 - 设置DTS通道延迟报警
 - 客户 - 购买BDS同步服务，并绑定指定VPC（VPC下的HBase集群可以相互迁移）
- 开启同步
 - 在目标HBase集群建表
 - 登陆BDS操作页面
 - 添加DTS订阅通道

- 添加HBase数据源
- 创建实时同步通道

注意事项

- RDS原表的数据类型（除了字节数组）都被转成了String写入到了HBase

RDS -> Phoenix 实时数据同步

版本支持

Mysql版本：

- 支持公共云、金融云RDS For MySQL实例、DRDS实例的数据订阅。
- 支持经典网络、VPC网络下RDS For MySQL实例、DRDS实例的数据订阅。

Phoenix版本：5.x (云2.xHBase)

同步步骤

- 购买及环境准备
 - 客户 - 开通DTS订阅通道
 - 客户 - 创建DTS消费通道
 - 客户 - 设置DTS通道延迟报警
 - 客户 - 购买BDS同步服务，并绑定指定VPC（VPC下的HBase集群可以相互迁移）
- 开启同步
 - 开通目标集群HBase SQL服务
 - 在目标集群建表，Phoenix表列类型和Mysql列类型对应
 - 登陆BDS操作页面
 - 添加Phoenix数据源
 - 创建实时同步通道

注意事项

- RDS原表的数据类型和Phoenix表类型要对应
- 目标集群是阿里云HBase2.x，开启query server的集群

BDS OPEN API 使用说明

前置条件

- 购买BDS数据迁移同步服务
- 确保客户端和BDS网络互通
 - 添加BDS白名单
 - telnet {{ BDSMaster }} 12311

HBase全量数据迁移接口说明

创建任务

接口（POST）：http://{BDSMaster}:12311/hbase/fulldata/create

参数：

@RequestParam("src") String srcName -> 源集群名

@RequestParam("dst") String sinkName -> 目标集群名

@RequestParam("tableNames") String tableNames -> 迁移的表（支持多张表，","号分割）

@RequestParam("bulkload") boolean bulkload -> 是否需要进行bulkload

@RequestParam("createTable") boolean createTable -> 是否要创建目标表

@RequestParam("extraAttrs") String extraAttrs -> 高级配置，可不填

例子：

```
curl -d "src=hbase1&dst=hbase2&tableNames=t1,t2,t3&bulkload=true&createTable=true" -H "Content-Type: application/x-www-form-urlencoded" -X POST http://{BDSMaster}:12311/hbase/fulldata/create
```

返回值：

```
{
  "success": "true", // 请求是否成功
  "message": "workflow-36c61d1240f84ecfbb472751a29b6deb" //任务id
}
```

查看所有任务

接口（GET）：http://{BDSMaster}:12311/hbase/fulldata

例子：curl http://{BDSMaster}:12311/hbase/fulldata

返回值：

```
{
  "tunnels": [ // 罗列出所有的任务信息
  {
    "jobName": "workflow-779298e682a44d5f9c226cea0e6f4dfd",
    "tableNames": [
```



```
"t1"
],
"dst": "100.81.140.116",
"fullJobs": [
{
"jobName": "fulldata-104f42d9bbf341099ae02c0b61a70f4c",
"state": "SUCCEEDED",
"table": "t1"
}
],
"bulkloadJobs": [
{
"TaskName": "bulkload-73bf97f03f6140da91c3183e3d4fc1d8",
"state": "SUCCEEDED"
}
],
"src": "11.164.87.151",
"mergeTasks": [],
"state": "SUCCEEDED",
"inrcJobs": [],
"schemeTasks": [
{
"TaskName": "360338ff-abfb-4f78-9b73-580cc9fd48be",
"state": "SKIPPED",
"table": "t1"
}
]
},
{
"jobName": "workflow-36c61d1240f84ecfbb472751a29b6deb",
"tableNames": [
"t1",
"t2"
],
"dst": "100.81.140.116",
"fullJobs": [
{
"jobName": "fulldata-9014e168fccd433dbd56bcc61ba4ccbb",
"state": "SUCCEEDED",
"table": "t2"
},
{
"jobName": "fulldata-114cc535693344238572583268ae4c67",
"state": "SUCCEEDED",
"table": "t1"
}
],
"bulkloadJobs": [
{
"TaskName": "bulkload-92402af7ef544f6182af892fe4b35c65",
"state": "SUCCEEDED"
},
{
"TaskName": "bulkload-f7580e9ebf734c999656c1aaad41f67b",
"state": "SUCCEEDED"
}
]
```

```
],
"src": "11.164.87.151",
"mergeTasks": [],
"state": "SUCCEEDED",
"inrcJobs": [],
"schemeTasks": [
{
"TaskName": "b31b8f7c-9a74-4840-814a-6cd62d2b08c9",
"state": "SKIPPED",
"table": "t1"
},
{
"TaskName": "b1968a67-6684-4dd5-ba07-dbd7a52af24f",
"state": "SKIPPED",
"table": "t2"
}
]
},
"success": "true",
"message": "ok"
}
```

查看指定任务id的任务

接口（GET）：<http://{BDSCMaster}:12311/hbase/fulldata/{jobId}/detail>

暂停任务

接口（GET）：<http://{BDSCMaster}:12311/hbase/fulldata/{jobId}/abort>

返回值：

```
{
"success": "true",
"message": "ok"
}
```

删除任务

接口（GET）：<http://{BDSCMaster}:12311/hbase/fulldata/{jobId}/del>

返回值：

```
{
"success": "true",
"message": "ok"
}
```

重试任务

接口（GET）：http://{BDSCMaster}:12311/hbase/fulldata/{jobId}/retry

返回值：

```
{
  "success": "true",
  "message": "ok"
}
```

HBase增量任务同步接口

创建任务

接口（POST）：http://{BDSCMaster}:12311/hbase/incr/create

参数：

@RequestParam("name") String name -> 同步通道的名称

@RequestParam("src") String srcName -> 源集群名

@RequestParam("dst") String sinkName -> 目标集群名

@RequestParam("tableNames") String tableNames -> 迁移的表（支持多张表，","号分割）

@RequestParam("createTable") boolean createTable -> 是否要创建目标表

@RequestParam("extraAttrs") String extraAttrs -> 高级配置，可不填

例子：

```
curl -d "name=test1&src=hbase1&dst=hbase2&tableNames=t1,t2,t3&createTable=true" -H "Content-Type: application/x-www-form-urlencoded" -X POST http://{BDSCMaster}:12311/hbase/incr/create
```

返回值：

```
{
  "success": "true",
  "message": "wal-08a087da488349969ff4a4b78efbceba" // 同步通道的id
}
```

更新任务

接口（POST）：http://{BDSCMaster}:12311/hbase/incr/update

参数：

@RequestParam("jobId") final String jobId

@RequestParam("tableNames") final String tableNames

例子：

```
curl -d "jobId=xxx&tableNames=t1,t2,t3" -H "Content-Type: application/x-www-form-urlencoded" -X POST
http://{BDSMaster}:12311/hbase/incr/update
```

查看所有任务

接口（GET）：<http://{BDSMaster}:12311/hbase/incr>

返回值：

```
{
  "tunnels": [
    {
      "jobId": "wal-2be4c2c89b7140a7b42c050916b3f5bb",
      "tableNames": [
        "default:t1"
      ],
      "dst": "100.81.140.116",
      "offset": 1573560272951,
      "src": "11.164.87.151",
      "name": "wal-2be4c2c89b7140a7b42c050916b3f5bb",
      "syncDate": "2019-11-12 20:04:32",
      "state": "KILLED",
      "synctime": 1573560272951
    },
    {
      "jobId": "wal-08a087da488349969ff4a4b78efbceba",
      "tableNames": [
        "default:t1",
        "default:t2",
        "default:t3"
      ],
      "dst": "100.81.140.116",
      "offset": 1573651478405,
      "src": "11.164.87.151",
      "name": "test1",
      "syncDate": "2019-11-13 21:31:46",
      "state": "RUNNING",
      "synctime": 1573651906518
    }
  ],
  "success": "true",
  "message": "ok"
}
```

查看指定id的任务

接口（GET）：<http://{BDSMaster}:12311/hbase/incr/{jobId}/detail>

暂停任务

接口（GET）：<http://{BDSMaster}:12311/hbase/incr/{jobId}/abort>

返回值：

```
{
  "success": "true",
  "message": "ok"
}
```

删除任务

接口（GET）：http://{BDSMaster}:12311/hbase/incr/{jobId}/del

返回值：

```
{
  "success": "true",
  "message": "ok"
}
```

重试任务

接口（GET）：http://{BDSMaster}:12311/hbase/incr/{jobId}/retry

返回值：

```
{
  "success": "true",
  "message": "ok"
}
```

使用BDS通过公网进行HBase数据的迁移

BDS集群默认不能访问公网，这给部分想通过公网将本地HBase迁移到云HBase的客户带来了不小的麻烦。本文主要介绍如何开通BDS集群访问公网的能力。

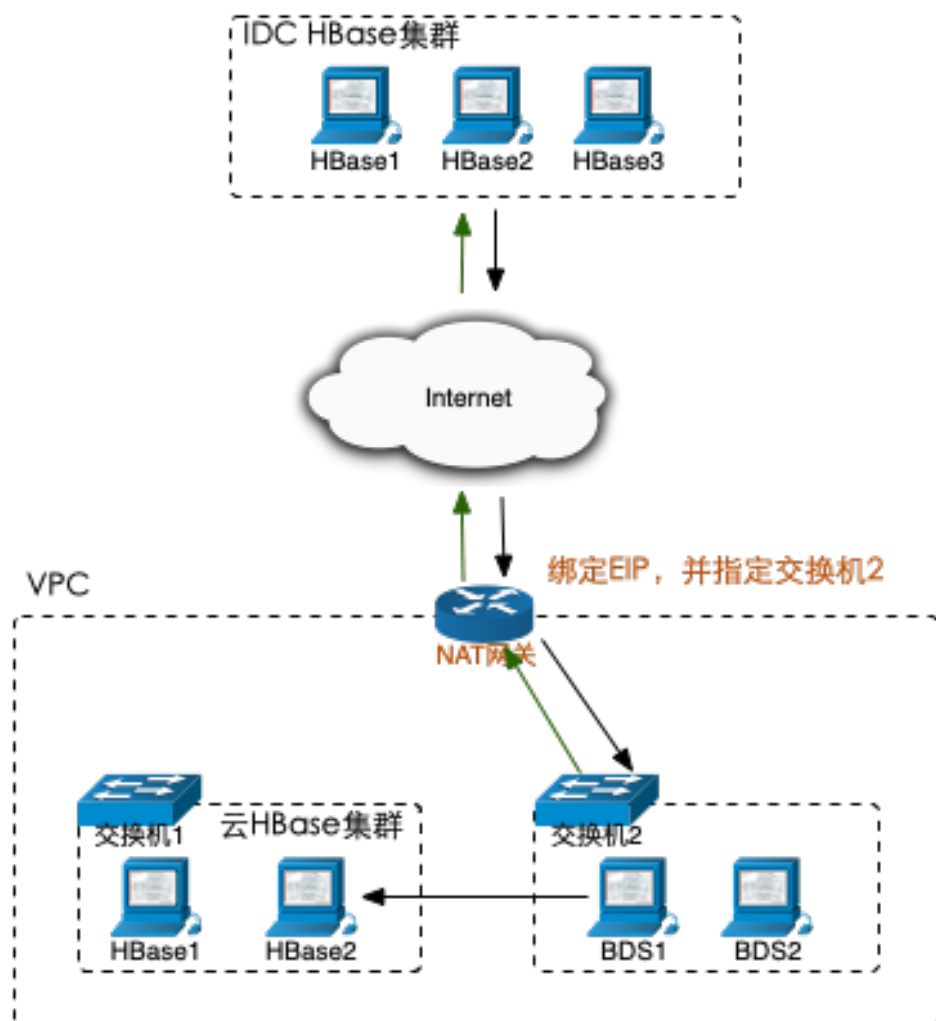
前提条件

1. 本地集群对公网开放
2. 对BDS开通本地集群访问的权限
3. 购买BDS

BDS打通公网流程

BDS默认情况下是不能访问公网的，为了打通BDS访问公网的能力主要分为以下几个步骤：

1. 对BDS所在VPC开通NAT网关
2. 根据迁移的规模开通对应规格的公网EIP
3. 绑定EIP到NAT网关，并指定BDS所在的交换机

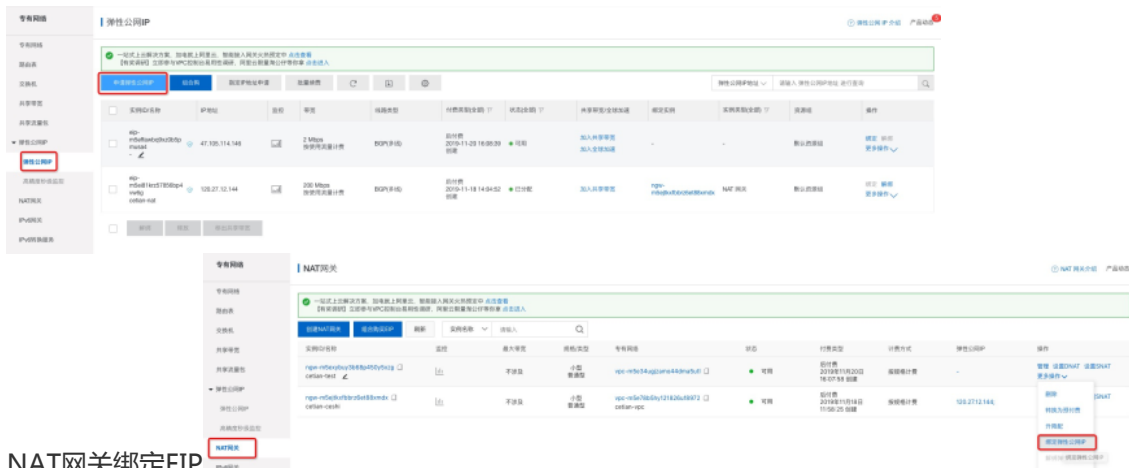


具体步骤

1. 针对BDS所在的VPC开通NAT网关，详见NAT网关产品文档

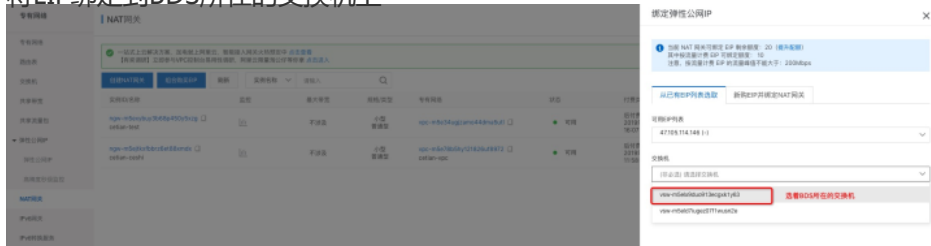


2. 创建对应规格的林IP



3. NAT网关绑定EIP

将EIP绑定到BDS所在的交换机上



4. 绑定成功后，对应VPC交换机下的BDS就有了访问公网的能力

5. 登陆BDS页面，添加HBase数据源，这里需要注意一下添加本地HBase数据源时需要填写

etc/hosts，并把内网的ip替换成公网对外的ip

添加HBase数据源

集群名

uf66hv2r32fy8132

HDFS地址

hdfs://{namenode1}:{port}/{namenode2}:{8020}}

HBase版本

HBase2x

HBase根目录

hbase-site.xml的hbase.rootdir配置，例如：/hbase

HBase连接信息

hb-proxy-pub-1061w2r32fy8132-master1-001.hbase.rds.aliyuncs.com

集群etc/hosts(自建、EMR集群需要)

47.101.253.187 hb-uf66hv2r32fy8132-master1-001.hbase.rds.aliyuncs.com
47.101.253.187 hb-uf66hv2r32fy8132-master2-001.hbase.rds.aliyuncs.com
100.100.100.100 hb-uf66hv2r32fy8132-master3-001.hbase.rds.aliyuncs.com
47.101.253.158 hb-uf66hv2r32fy8132-core-001.hbase.rds.aliyuncs.com

替换成本地集群外网ip

添加

6. 提交实时同步任务和全量数据迁移任务

DTS订阅通道

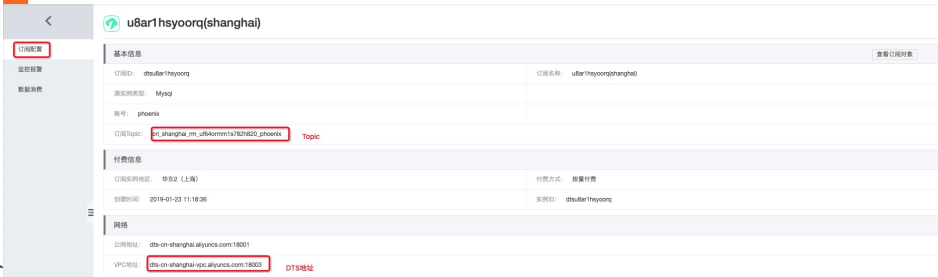
1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
2. 开通DTS订阅通道，订阅RDS的实时数据

1. 登陆BDS操作页面，数据源管理->DTS订阅通道

2. 录入DTS订阅通道的信息

数据源名称

23



获取DTS连接地址，Topic

消费通道ID，消费通道账户，消费通道密码



HBase数据源

前置条件

- 1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
- 2. 打通BDS和HBase迁移集群的网络

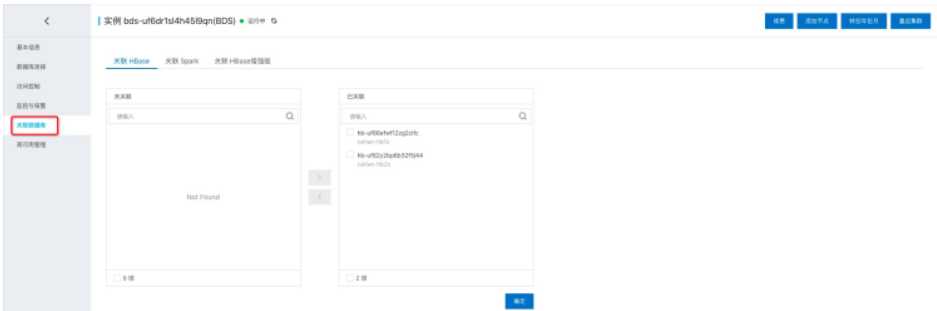
控制台绑定HBase数据源

使用场景：

- 1. 于BDS处于同一个VPC下的云HBase

添加步骤：

- 1. 进入BDS控制台页面 -> 关联数据库，选择需要添加的HBase集群



2. 关联完成之后，进入BDS的操作页面，点击**数据源管理 -> HBase数据源**



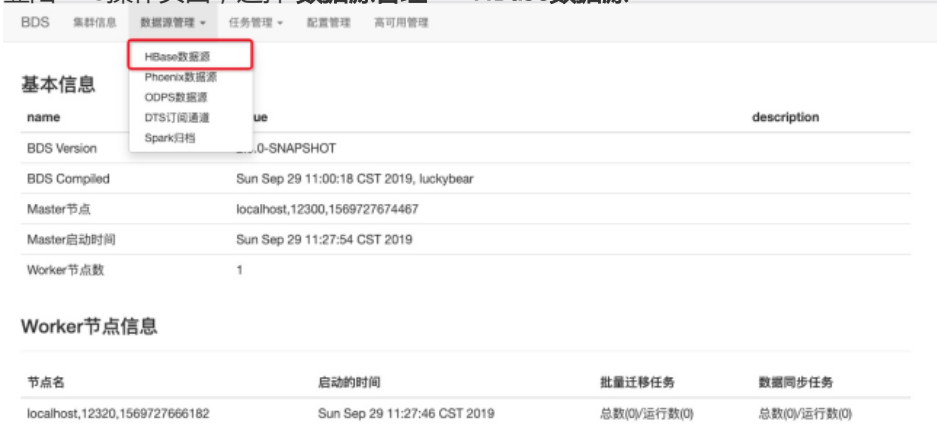
手动添加HBase数据源

使用场景（不能通过关联数据源进行关联的集群）：

1. 自建HBase集群
2. EMR HBase集群
3. 不同与BDS所在VPC的云HBase集群

添加步骤：

1. 登陆BDS操作页面，选择 **数据源管理 -> HBase数据源**



2. 进入HBase数据源的管理界面， 点击**添加数据源**



3. 录入集群的信息，点击添加

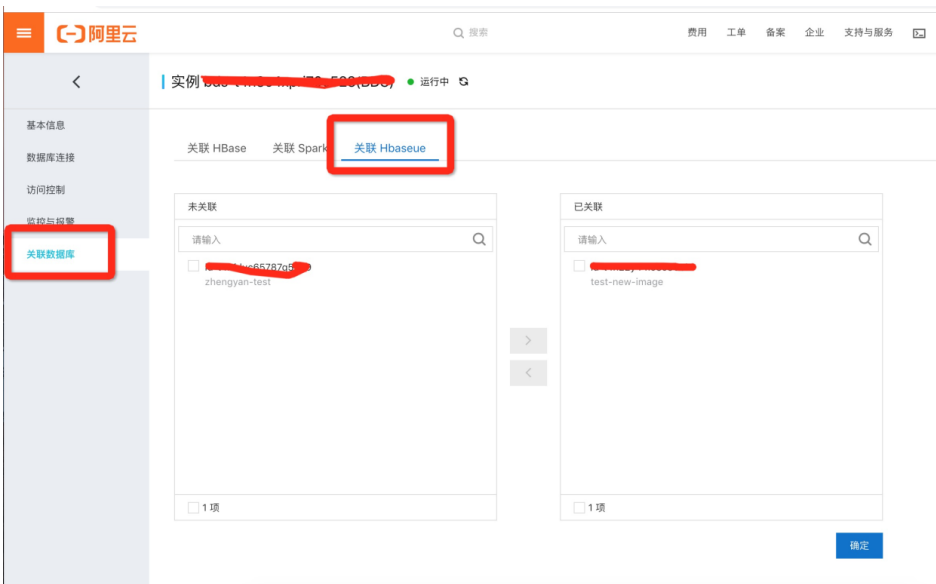


HBase增强版

BDS中添加HBase增强版数据源，请按以下步骤操作

1. 关联HBase增强版集群

进入阿里云BDS集群控制台 “关联数据库” -> “关联HBase增强版：



关联后在BDS WebUI可以查看

到新关联的集群（注：关联后可能需要稍等1分钟左右生效）：

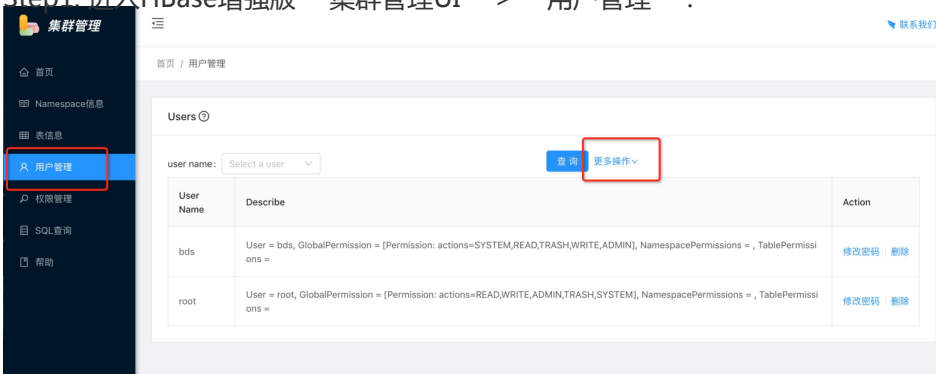


2. 创建/获取用户信息

BDS访问HBase增强版需要配置一个用户，该用户需要有SYSTEM/READ/WRITE/ADMIN权限，您可以选择：

1. 使用系统默认的root用户：参考 [用户和ACL管理](#)
2. 或手动创建一个新用户，创建方式：

Step1. 进入HBase增强版 “集群管理UI” > “用户管理”：



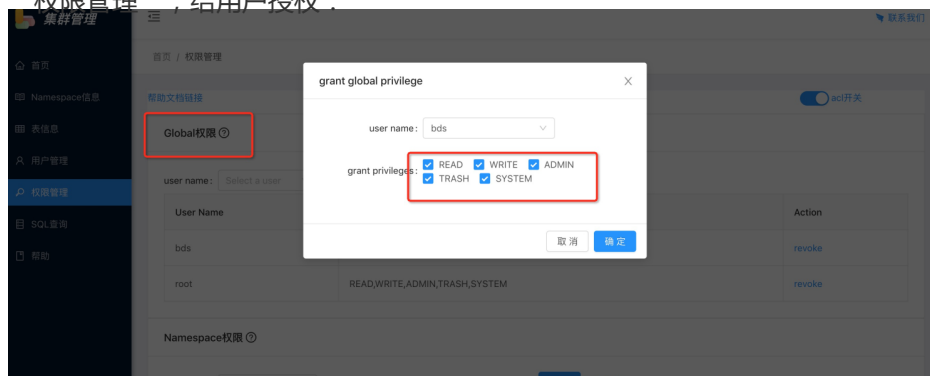
Step2.选择“更多操作” ->



“create user”：

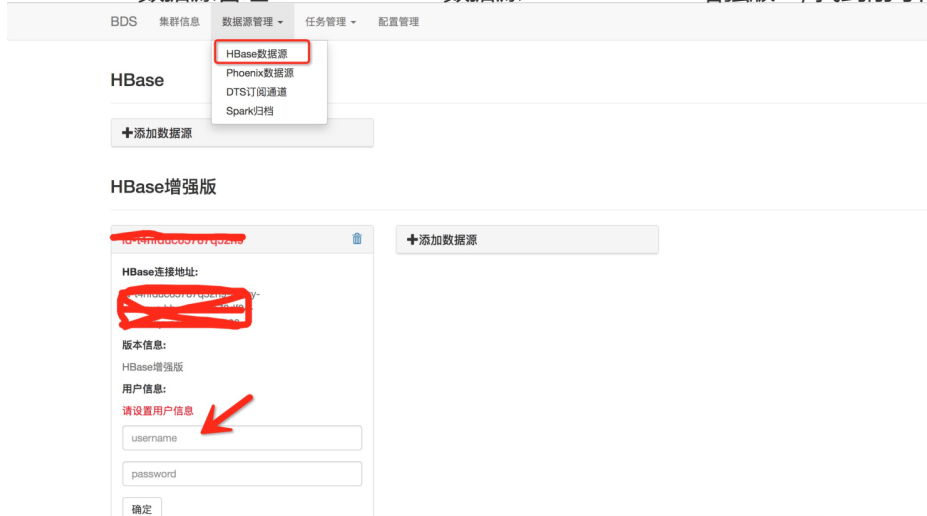
Step3.进入

“权限管理”，给用户授权：



3. 配置用户信息

进入BDS webUI -> “数据源管理” -> “HBase数据源” -> “HBase增强版”，找到刚才添加的集群，添加



用户信息：

ODPS数据源

前提条件

1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
2. 开通ODPS

数据源添加

步骤：

1. 进入BDS操作页面，**数据源管理 -> ODPS数据源**



2. 录入ODPS的相关信息，**点击创建**

BDS 集群信息 数据源管理 任务管理 配置管理 高可用管理

添加ODPS数据源

数据源名称 **数据源别名**

hbase2odps

AccessId **账户AK信息**

AccessKey

EndPoint **根据ODPS所在的区域填写对应区域服务地址**

http://service.cn.maxcompute.aliyun.com/api

TunnelEndPoint

http://dt.cn-shanghai.maxcompute.aliyun.com

ODPS Project **需要关联的ODPS的Project**

hbase2odps

创建

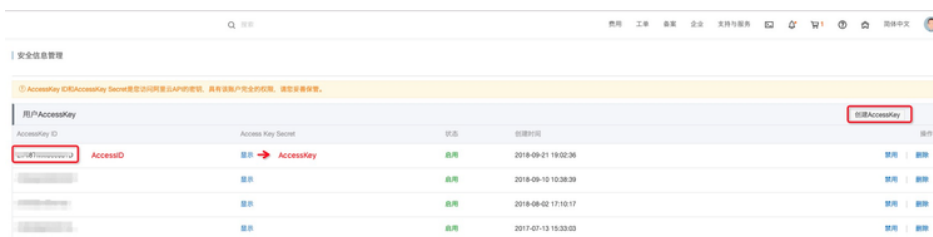
参数说明：

获取 AccessId、AccessKey

DataWorks / 工作空间列表

创建工作空间 搜索输入工作空间名称

工作空间名称/别名	模式	创建时间	管理员	状态	开通服务	操作
hbase2odps	简单模式	2019-09-29 19:03:22	cloud_hbase	正常	√	工作空间配置 进入数据服务
hbase2odps	简单模式	2019-09-29 19:03:22	cloud_hbase	正常	√	工作空间配置 进入数据服务
bds_odps	标准模式	2019-08-15 12:40:50	cloud_hbase	正常	√	工作空间配置 进入数据服务
bds	标准模式	2019-08-15 12:40:50	cloud_hbase	正常	√	工作空间配置 进入数据服务
catian	简单模式	2019-08-20 11:57:10	cloud_hbase	正常	√	工作空间配置 进入数据服务
catian	简单模式	2019-08-20 11:57:10	cloud_hbase	正常	√	工作空间配置 进入数据服务
dingqi_test	简单模式	2019-05-29 10:00:50	cloud_hbase	正常	√	工作空间配置 进入数据服务
dingqi_test	简单模式	2019-05-29 10:00:50	cloud_hbase	正常	√	工作空间配置 进入数据服务



获取EndPoint、TunnelEndPoint

根据ODPS所在地域，查阅 https://help.aliyun.com/document_detail/34951.html 获取VPC网络下的EndPoint地址。测试案例中，ODPS在华东2，因此选择华东2的VPC网络下EndPoint地址

- VPC网络下Region和服务连接对照表

在VPC网络下访问MaxCompute，只能使用如下Endpoint和Tunnel Endpoint。

Region名称	所在城市	开服状态	VPC网络Endpoint	VPC网络Tunnel Endpoint
华东1	杭州	已开服	http://service.cn-hangzhou.maxcompute.aliyun-inc.com/api	http://dt.cn-hangzhou.maxcompute.aliyun-inc.com
华东2	上海	已开服	http://service.cn-shanghai.maxcompute.aliyun-inc.com/api	http://dt.cn-shanghai.maxcompute.aliyun-inc.com

MaxCompute已开服的Region
和服务连接对照表

访问原则

获取ODPS Project



RocketMQ数据源

前置条件

1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
2. 确保BDS所在的VPC能够访问自建的或者云上MQ

数据源添加

1. 首先进入BDS操作页面，数据源管理 -> RocketMQ数据源



添加RocketMQ数据源

数据源名称

mq

NameSrvAddr
http://MQ_INST_170260740406241...internet-oss-cn-mq-internet.aliyuncs.com:90

HBase版本
RocketMQ4.5

是否是阿里云MQ
☒

AccessKey
LWU*****D

AccessSecretKey
5I4Q7Hn*****m

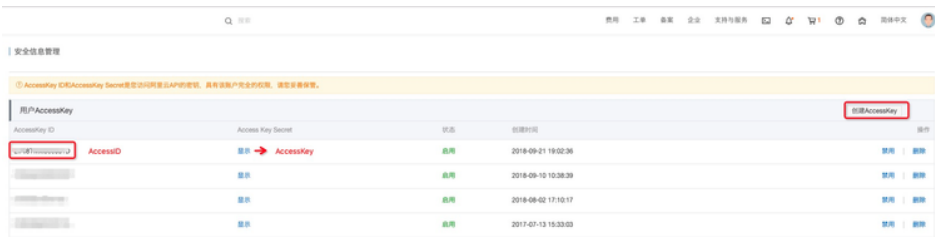
创建

获取NameSrvAddr

获取

获取 AccessId、AccessKey

The screenshot shows the 'DevWorks' workspace configuration page. The table lists configurations for 'Hbase2019gs' and 'kudu_jdk8'. The 'Hbase2019gs' configuration is highlighted, showing it is a 'Simple Mode' configuration with status 'Success' and a 'View Details' button. The 'kudu_jdk8' configuration is also shown with status 'Success' and a 'View Details' button. The 'View Details' button is highlighted with a red box.



BDS使用说明

HBase历史数据迁移

前提条件

- 1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
- 2. 打通BDS和HBase迁移集群的网络
- 3. 添加HBase集群数据源

版本

- 1. 自建HBase1.x、2.x
- 2. EMR HBase
- 3. 标准版云HBase、增强版云HBase (单机版本、集群版本)

创建任务

- 1. 进入BDS操作页面，点击**任务管理->HBase历史数据迁移->创建HBase迁移任务**



- 2. 选择源HBase集群和目标HBase集群，输入需要迁移的表

HBase历史数据批量迁移

源集群

hb-j5e517a16xb80rkb

目标集群

ld-j5e64ub2z4fx9x828

操作

☒ 是否在目标集群建表 ☒ 是否进行数据装载 ☐ 迁移前是否进行flush

迁移的表

namespace1:srcTable/namespace2:targetTable {"startTs": "20191001000000", "endTs": "20191010000000", "startKey": "aaa", "endKey": "zzz"}
namespace1:srcTable1
namespace1:srcTable2 {"cols": ["f1:a", "f1:b"]}
namespace1:table3 {"copyType": "api"}

创建

3. 查看迁移进度，当所有的任务都完成之后
批量数据迁移任务: workflow-1567670649703

建表详情

表名	任务状态	开始时间	结束时间
default:unid -> default:unid_bds_realtime	SKIPPED	2019-09-05 16:04:10	2019-09-05 16:04:10

数据迁移详情 (总共: 1 完成: 1 失败: 0)

迁移大小: 687.7 GB 迁移速度: 0.00MB/s

任务名	表名	任务状态	开始时间	结束时间	完成进度	操作
fulldata-1567670652710	default:unid -> default:unid_bds_realtime	SUCCEEDED	2019-09-05 16:04:28	2019-09-05 18:20:47	202/202	详情 终止

数据装载 (总共: 1 完成: 1 失败: 0)

任务名	表名	任务状态	开始时间	结束时间	操作
bulkload-1567670652717	default:unid -> default:unid_bds_realtime	SUCCEEDED	2019-09-05 18:20:47	2019-09-05 18:22:13	终止 重试

参数说明

```
hbaseTable/targetTable {"cols":["f1:a", "f1:b", "f1:c"]}
hbaseTable/targetTable {"copyType": "api"}
hbaseTable/targetTable {"startTs": "20190101000000", "endTs": "20200101000000", "startKey": "a", "endKey": "z"}
hbaseTable
```

指定导出的列

例如：hbaseTable/targetTable {"cols":["f1:a", "f1:b", "f1:c"]}

支持不同的迁移模式

历史数据迁移支持两种模式：文件拷贝模式和api模式默认情况是走文件拷贝的方式，迁移过程不会和源HBase集群交互，只和HDFS进行交互api模式走的是HBase的API，对于大表的迁移，会影响源集群的业务。适用于单机版HBase，或者不方便对BDS开放hdfs的集群迁移

指定迁移数据的范围

入参：table1 {"startKey": "xxx", "endKey": "zzz", "startTs": "20191001153000", "endTs": "20191001233000"}

迁移支持指定任意startKey、endKey、startTs、endTs参数，可以控制导出的时间范围和rowkey的区间

指定迁移到不同HBase表

支持HBase原表和目标表不同表名, 默认情况目标表和源表名相同例如: srcTable/targetTable

HBase实时数据同步

前提条件

- 1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
- 2. 打通BDS和HBase迁移集群的网络
- 3. 添加HBase集群数据源

版本

- 1. 自建HBase1.x、2.x
- 2. EMR HBase
- 3. 标准版云HBase、增强版云HBase (仅支持集群版本)

任务创建

- 1. 进入BDS操作页面，点击任务管理->HBase实时数据同步->创建HBase实时同步

BDS

集群信息

数据源管理

任务管理

配置管理

HBase实时数据同步

正在执行的通道

创建HBase同步通道

结束通道

删除所有已结束的通道

同构数据源

HBase历史数据迁移

HBase实时数据同步

HBase数据校验

异构数据源

DTS实时数据订阅

HBase实时数据归档

通道名	源集群	目标	同步起始点位	同步点位	任务队列	同步时延	操作
unid_realtime_prod	hb-2zei3t982b592iti5	bigdata-prod	KILLED	20190905144232	20191013092600	2019-09-05 14:47:32	2019-10-13 09:49:35
test	bigdata-test	bigdata-prod	KILLED	20190905141837	20190905141621	2019-09-05 14:23:37	2019-09-05 14:41:42

- 2. 指定源集群HBase和目标集群HBase，指定需要同步的表

通道名

源集群

hb-j5e517al6xlb80rkb

目标集群

hb-j5e517al6xlb80rkb

操作

☒ 是否在目标集群建表

同步的表

请勿填写phoenix表,填写格式如下:
namespace:tablename
namespace1:tablename1

创建同步通道

总结

同步集群	通道状态	延迟	处理耗时	读详情	写详情	通道
hb-2z-  no8r3ts6 -> hb-2-  0q8873	RUNNING	8 ms	min=4ms/max=9ms/avg=7ms	209.7 TB 0.30 MB/s	171.2 TB 0.00 MB/s	运行

表信息详情

表名	PUT	DELETE	起始时间	最近更新时间
default:vk_log	80617668	0	2019-08-07 14:20:30	2019-10-11 14:20:10
vk_class:vk_class_realtime_course	8595843615	0	2019-07-10 17:30:04	2019-10-14 11:24:20
bi_log:classroom_bicall	717333648	7	2019-07-10 15:17:07	2019-10-14 11:25:43
vk_class:vk_class_core_map_userlist	45671983	0	2019-07-10 17:30:55	2019-10-14 11:24:20
bi_log:classroom_event	10859324952	0	2019-07-10 15:17:05	2019-10-14 11:25:49
bi_log:classroom_mapping_new	53442605	0	2019-10-08 16:48:38	2019-10-14 08:02:52
server_log:classroom_server_list	9677964673	0	2019-07-10 17:30:04	2019-10-14 11:24:20
vk_class:vk_class_all_course	15325331	0	2019-07-11 08:51:31	2019-10-14 08:51:34
server_log:classroom_server_call	322042399	0	2019-07-10 17:30:04	2019-10-14 11:24:21
bi_log:classroom_mapping	95104418	0	2019-08-21 17:34:01	2019-10-08 10:41:38
bi_log:classroom_role_tag	5090097	0	2019-09-02 20:14:58	2019-10-14 09:37:58
default:vk_class_tag_new	82661376	0	2019-07-10 15:51:03	2019-10-14 09:37:41
default:vk_log_new	18481789	0	2019-09-16 10:17:55	2019-10-14 11:25:49
server_log:classroom_server_event	839216978	0	2019-07-10 17:30:21	2019-10-14 11:24:20

日志信息详情

日志文件	状态	读详情
hb-2ze276uwwnc-r3lts6-core-001.hbase.rds.aliyuncs.com%2C16020%2C1561712760204.default.1563837495362	SKIPPED	0 0.00 MB/s
hb-2ze276uwwnc-r3lts6-core-002.hbase.rds.aliyuncs.com%2C16020%2C1566585058925.default.1567521317524	SUCCEEDED	4.1 MB 0.00 MB/s
hb-2ze276uwwnc-r3lts6-core-001.hbase.rds.aliyuncs.com%2C16020%2C1567513743806.default.1567520567187	SUCCEEDED	143.9 MB 8.53 MB/s

3. 查看同步状态

更新表

1. 进入BDS操作页面，点击任务管理->HBase实时数据同步

	目标集群	通道状态	同步起始点位	同步点位	任务队列	同步时延	操作
mo8r3ts6	hb-2ze88nkt0bp9q8873	RUNNING	20190710161603	2019-10-14 11:35:33(1571024133560)	运行中(3) / 队列中(0)	7 ms	更新表 禁止

結束通道 刪除所有已結束的通道

通道名	源集群	目标集群	状态	同步开始点位	同步结束点位	任务开始时间	任务结束时间	操作
-----	-----	------	----	--------	--------	--------	--------	----

HBase实时数据同步表更新

通道名

hbase-hot-cold

源集群

hb-2ze3r3ts6

目标集群

hb-2ze9p9q8873

同步的表

bi_l

bicall

bi_l

event

ser

class

call

ser

classro

event

ser

classro

list

vk

cli

userlist

vk_class

real

course

2. 修改表列表

HBase历史数据导出ODPS

前提条件

1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
2. 打通BDS和HBase迁移集群的网络
3. 添加HBase集群数据源
4. 添加ODPS数据源

版本

1. 自建HBase1.x、2.x
2. EMR HBase
3. 标准版云HBase、增强版云HBase（集群版本）

创建任务

1. 进入BDS操作页面，点击**任务管理->HBase历史数据导出ODPS**

BDS 集群信息 数据源管理 任务管理 配置管理 高可用管理

HBase历史数据批量导出到ODPS

正在执行的任务 **创建迁移任务**

任务名	源集群名	目标集群名	状态	开始时间	结束时间	操作
-----	------	-------	----	------	------	----

已经完成的任务 [删除所有已完成的任务](#)

任务名	源集群名	目标集群名	状态	开始时间	结束时间	其他操作
workflow-1569498958918	11.164.87.151.bak2	odps	SUCCEEDED	2019-09-26 19:55:58	2019-09-26 19:56:21	重试 删除
workflow-1569498773070	11.164.87.151.bak2	odps	SUCCEEDED	2019-09-26 19:52:53	2019-09-26 19:53:11	重试 删除
workflow-1569498461338	11.164.87.151	odps	SUCCEEDED	2019-09-26 19:47:41	2019-09-26 19:47:50	重试 删除
workflow-1569497911798	11.164.87.151	odps	SUCCEEDED	2019-09-26 19:38:31	2019-09-26 19:39:00	重试 删除

2. 选择源HBase集群和目标ODPS集群，输入需要迁移的表
HBase历史数据批量导出ODPS

源集群

11.164.87.151

目标集群

hbase2odps

迁移的表

wal-test-1220974600/bdsTest1 {"cols":["cf1:string|string","cf1:int|int","cf1:double|double"],"tableMode":"wideTable"}
wal-test-1220974600/bdsTest2 {"tableMode":"KVTable"}

创建

3. 查看迁移进度，当所有的任务都完成之后

批量数据迁移任务: workflow-1569756429256

建表详情

表名	任务状态	开始时间	结束时间
bdsTest1_tmp	SUCCEEDED		2019-09-29 19:27:12
bdsTest2	SUCCEEDED		2019-09-29 19:27:12

数据迁移详情 (总共: 2 完成: 2 失败: 0)

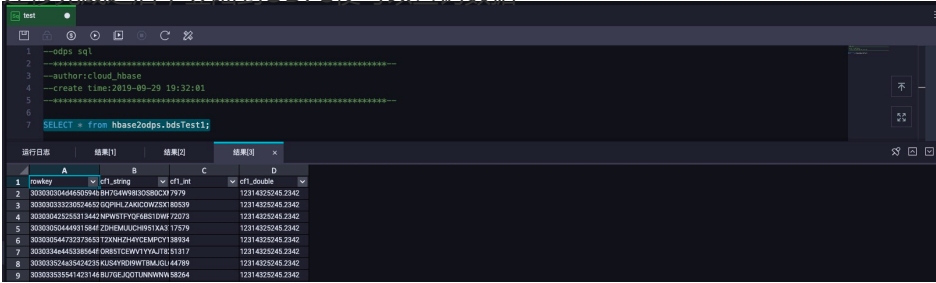
迁移大小: 75.9 MB 迁移速度: 0.00MB/s

任务名	表名	任务状态	开始时间	结束时间	完成进度	操作
fulldata-98331d9fb73a484bab31a9a428ecf7b	wal-test-1220974600 -> bdsTest2	SUCCEEDED	2019-09-29 19:27:37	2019-09-29 19:27:43	10/10	详情 终止
fulldata-2f6db1cc60db4799997e2433929891f7	wal-test-1220974600 -> bdsTest1	SUCCEEDED	2019-09-29 19:27:32	2019-09-29 19:27:37	10/10	详情 终止

Table Merge (总共: 1 完成: 0 失败: 0)

任务名	表名	运行状态	任务进度	耗时	起始时间	结束时间	操作
merge-4c9e196e296e40a2b20c3247a4bb2f74	wal-test-1220974600 -> bdsTest1	RUNNING	50.00 %	0.00 ms	20190929192737		中断 重试

4. 迁移完成之后，登录到ODPS便可以查询数据



参数说明

hbaseTable/odpsTable {"tableMode": "wideTable", "cols":["f1:a", "f1:b", "f1:c"]}
hbaseTable/odpsTable {"tableMode": "wideTable", "cols":["f1:a|string", "f1:b|boolean"]}
hbaseTable/odpsTable {"tableMode": "wideTable", "cols":["f1:a|string", "f1:b|boolean"], "startTs": "20190101000000", "endTs": "20200101000000", "startKey": "a", "endKey": "z"}
hbaseTable/odpsTable {"tableMode": "KVTable"}

HBase数据导出到ODPS，支持两种模式：KV模式、宽表模式

KV模式

入参：tablexxx {"tableMode": "KVTable"}

结果输出：

	A	B	C	D	E	F
1	rowkey	version	op	family	qualifier	val
2	3737375a4c424f5033!1569404372597	Put	cf1	bigdecimal		00000040ddb105bd3
3	3737375a4c424f5033!1569404372597	Put	cf1	boolean	ff	
4	3737375a4c424f5033!1569404372597	Put	cf1	double		4206efed87e9dfa4
5	3737375a4c424f5033!1569404372597	Put	cf1	float	4499e77f	
6	3737375a4c424f5033!1569404372597	Put	cf1	int	0000641a	
7	3737375a4c424f5033!1569404372597	Put	cf1	long		0000002dfdc1c3e
8	3737375a4c424f5033!1569404372597	Put	cf1	null		
9	3737375a4c424f5033!1569404372597	Put	cf1	short	0001	
10	3737375a4c424f5033!1569404372597	Put	cf1	string		4b43524b5039315241
11	37373836555a52474f!1569404344057	Put	cf1	bigdecimal		00000040ddb105bd3
12	37373836555a52474f!1569404344057	Put	cf1	boolean	ff	
13	37373836555a52474f!1569404344057	Put	cf1	double		4206efed87e9dfa4

每一行代表HBase中的一个KV，rowkey表示的是HBase表KV的rowkey，version代表是KV的版本，op代表是KV的类型（Put、DeleteFamily、DeleteColumn），family代表的是列族，qualifier代表的是列名，value代表是KV的具体value（转化成HexString的格式）

宽表模式

入参：hbaseTable/odpsTable {"tableMode": "wideTable", "cols":["cf1:string|string", "cf1:int|int", "cf1:double|double"]}

宽表模式需要指定导出的列，例如“cols”:[“f1:a”，“f1:b”，“f1:c”]，列值默认是转化成HexString的格式。可以指定列的类型，例如“cols”:[“f1:a|string”，“f1:b|boolean”，“f1:c|int”]。列支持的类型：string、int、long、short、decimal、double、float、boolean，默认类型 hex，将bytes数组转化为HexString

结果输出:

	A	B	C	D
1	rowkey	cf1_string	cf1_int	cf1_double
2	303030304d4650594b8H7G4W98I3OSB0CXI7979			12314325245.2342
3	303030333230524652GQPIHLZAKICOWZSX180539			12314325245.2342
4	303030425255313442NPW5TFYQF6BS1DWF72073			12314325245.2342
5	30303050444931584fZDHEMUUCHI951XA317579			12314325245.2342
6	303030544732373653T2XNHZH4YCEMPCY138934			12314325245.2342
7	3030334e445338564fOR85TCEWV1YYAJT8:51317			12314325245.2342
8	303033524a35424235KUS4YRDI9WTBMJGLI44789			12314325245.2342
9	303033535541423146BU7GEJQOTUNNWNW58264			12314325245.2342
10	30303446304f49513469PKNZ07IOY0KM5GI54219			12314325245.2342
11	303035414e5543365209U6DML9TUWXEKQ339945			12314325245.2342
12	303035493230335356HY93A9Y0XQUSIPT4:9360			12314325245.2342

指定导出数据的范围

入参：table1 {"startKey": "xxx", "endKey": "zzz", "startTs": "20191001153000", "endTs": "20191001233000"}

HBase导出ODPS不管是KV模式还是宽表模式，导出支持指定任意startKey、endKey、startTs、endTs参数，可以控制导出的时间范围和rowkey的区间

指定导出到不同ODPS表

HBase导出ODPS支持HBase表和ODPS表不同表名，默认情况和HBase表名相同（hbase表名中的‘-’、‘.’

会被替换成‘ _ ’) 例如: hbaseTable/odpsTable {"tableMode": "KVTable"}

HBase实时数据归档ODPS

前提条件

- 1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
- 2. 打通BDS和HBase迁移集群的网络
- 3. 添加HBase集群数据源
- 4. 添加ODPS数据源

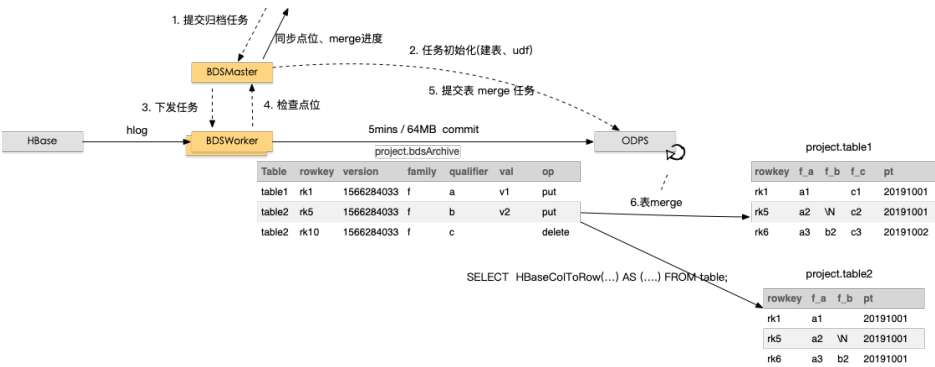
版本

- 1. 自建HBase1.x、2.x
- 2. EMR HBase
- 3. 标准版云HBase、增强版云HBase（集群版本）

限制

实时数据归档基于HBase的日志，因此不能导出bulkload的数据

归档原理



- 1. 提交归档任务，包括HBase集群信息、ODPS连接信息、需要归档的表、归档的时间粒度
- 2. BDS会在ODPS表对应的project上创建{{project}}.bdsArchive的临时表（临时表表名可以指定），并上传UDF到ODPS指定的Project下
- 3. BDSMaster构建任务并下发任务给BDSWorker
- 4. BDSWorker读取日志，每隔5分钟或者数据达到64MB将日志数据归档到ODPS的临时表{{project}}.bdsArchive中
- 5. BDSMaster检查归档任务的点位同步的点位，触发ODPS表merge的任务，将

{{project}}.bdsArchive中对应区间的数据进行行转列，最后将新数据写入到{{project}}.{{targetTable}}的对应分区中

其中临时表{{project}}.bdsArchive表结构如图所示，table字段表示HBase的表名，rowkey就是表的rowkey，version就是HBase中KV的版本（默认是KV入库的时间戳），family 表示的是HBase表的列族，qualifier 表示HBase表的列，val 表示KV的值，op表示KV的操作类型。每一行代表一个HBase中的一个KV。其中rowkey和val默认转化成HexString的格式

提交归档任务

1. 进入BDS操作页面，点击任务管理->HBase实时数据导出ODPS

BDS 集群信息 数据源管理 任务管理 配置管理 高可用管理

HBase实时数据归档OD

正在执行的任务 创建同步任务

任务名	源集群名	目标集群名	状态	开始时间	结束时间	操作
workflow-1570500348372	HBase历史数据导出ODPS	hbase2odps	RUNNING	2019-10-08 10:05:48		停止
	HBase实时数据导出ODPS					

已经完成的任务 删除所有已完成的任务

任务名	源集群名	目标集群名	状态	开始时间	结束时间	其他操作
workflow-1569975081005	hb-uf62y2bp6b32f5j44	hbase2odps	KILLED	2019-10-02 08:11:21		重试 删除
workflow-1569972805554	hb-uf62y2bp6b32f5j44	hbase2odps	KILLED	2019-10-02 07:33:25		重试 删除

2. 点击创建同步任务，进入任务创建页面，选择对应的HBase源集群和ODPS目标集群，指定需要导

HBase实时数据归档到ODPS

源集群

11.164.87.151

目标集群

hbase2odps

迁移的表

wal-test-301333235 ["cols":["cf1:a|long","cf1:b|short","cf1:c|decimal","cf1:d|double"],"mergeInterval":1800000,"mergeStartAt":"20190930100000"]

高级配置

odpsTmpTable=xxxx

出的HBase表 创建 如图
所示，对HBase表wal-test-301333235进行实时数据归档到ODPS，归档的列包括cf1:a、cf1:b、cf1:c、cf1:d四列。mergeInterval 表示的是归档的间隔（单位毫秒），默认指定 86400000（一天），mergeStartAt（格式为yyyyMMddHHmmss）表示从2019年9月30号0点整作为起始时间（可以指定一个过去的值）

归档任务: workflow-1570502364824

建表详情

表名	任务状态	开始时间	结束时间
bdsArchive	SUCCEEDED		2019-10-08 10:39:26

实时同步通道

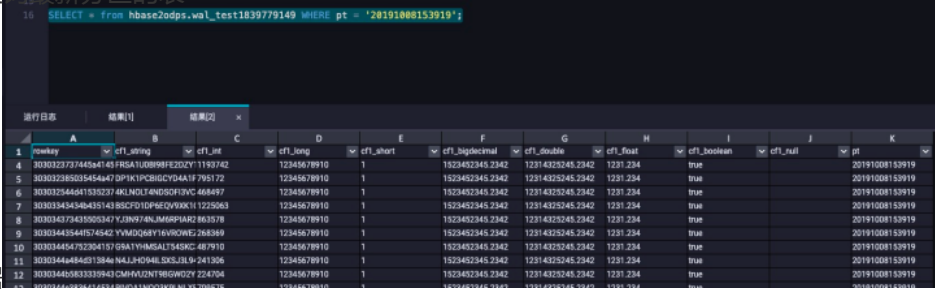
通道名	通道状态	同步起始点位	同步点位	任务队列	同步时延	操作
wal-1e5ddb45e0a4ee9a3e4e746fbcadf7	RUNNING	20191008103427	2019-10-08 11:14:23(1570504463886)	运行中(2) / 队列中(0)	4520 ms	

Table Merge

任务名	表名	运行状态	任务进度	耗时	起始时间	结束时间
merge-1d7e3406d206474bae624fd34adea625	wal-test1286719731 -> wal_test1286719731	RUNNING	47.50%	0.00 ms	20191008111411	

3. 查看表归档进度 实
时同步通道展示的是日志同步的延迟、点位，Table Merge 展现的是表Merge的任务，Merge完成

即可在ODPS上查询最新分区的表



4. ODPS查询表数据

参数说明

导出表格式如下：

```
hbaseTable/odpsTable {"cols": ["cf1:a|string", "cf1:b|int", "cf1:c|long", "cf1:d|short", "cf1:e|decimal",
"cf1:f|double", "cf1:g|float", "cf1:h|boolean", "cf1:i"], "mergeInterval": 86400000, "mergeStartAt": "20191008100547"}
hbaseTable/odpsTable {"cols": ["cf1:a", "cf1:b", "cf1:c"], "mergeStartAt": "201910080000000"}
hbaseTable {"mergeEnabled": false} // 不进行merge操作
```

导出表主要包括三个部分 {{hbaseTable}}/{{odpsTable}} {{tbConf}}

hbaseTable 指定的是HBase表

odpsTable可选，指定目标表名，默认情况和HBase相同（ODPS表名不支持包含“.-”等字符，统一转化成“_”）

tbConf 指定表的归档行为，支持的参数如下

参数	说明	例子
cols	指定需要导出表的哪些列，以及列的类型, 默认转化为HexString	"cols" : ["cf1:a" , "cf1:b" , "cf1:c"]
mergeEnabled	是否需要进行KV表转化成宽表，默认为true	"mergeEnabled" : false
mergeStartAt	merge的起始时间，可以指定一个过去的时间，格式是yyyyMMddHHmmss	"mergeStartAt" : "20191008000000"
mergeInterval	merge的时间间隔，单位毫秒，默认是一天，即按天归档	"mergeInterval" : 86400000

HBase实时数据导出RocketMQ

前置条件

- 1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面

2. 打通BDS和HBase迁移集群的网络
3. 添加HBase集群数据源
4. 添加RocketMQ数据源

版本支持

- 开源RocketMQ 4.5
- 阿里云消息队列 RocketMQ

限制

- HBase实时数据同步MQ目前还不支持保序
- BDS任务重试，MQ自身问题导致消息会被重复消费，需要业务方做到消费幂等
- BDS同步过程中不支持点位回退
- 只同步更新的列，相同Rowkey的其他未更新的列不会被同步
- BDS不支持同步Bulkload进来的数据
- BDS不支持Phoenix表数据的转码，需要客户自行处理Phoenix表的原生数据

任务创建

点击 任务管理 -> HBase实时数据同步RocketMQ -> 创建通道

BDS 集群信息 数据源管理 HBase集群迁移 HBase导入导出 高可用管理 备份恢复 配置管理

基本信息

name	value	description
BDS Version	2.4.0	
BDS Compiled	Thu Nov 14 15:47:38 CST 2019, luckybear	
Master节点	localhost,12300,1573718130838	
Master启动时间	Thu Nov 14 15:55:30 CST 2019	
Worker节点数	0	

HBa

历史数据导出ODPS

实时数据导出ODPS

历史数据导出HDFS

实时数据归档HDFS

RDS实时数据订阅

HBase实时数据同步RocketMQ

选择源HBase集群、目标MQ数据源，指定默认同步到MQ的Topic，以及同步通道所属的Group

通道名

test

源HBase数据源

11.164.87.151

目标MQ数据源

mq

RocketMQ GroupId

GI0-cetian

RocketMQ Topic

TopicTest

同步的表

t1 ("tag": "TagA")

t2

创建同步通道

总结

同步集群	通道状态	延迟	处理耗时	读详情	写详情	通道队列	同步起始点位
11.164.87.151 -> mq	RUNNING	272.00 ms	min=0ms/max=178ms/avg=59ms	122 B 0.00 MB/s	10 B 0.00 MB/s	运行(3) / 队列(0)	2019-10-31 14:

表名	PUT	DELETE	起始时间	最近更新时间
t1	1	0	2019-10-31 14:39:46	2019-10-31 14:39:46

日志文件	状态	读详情	写详情	日志延迟	处理耗时	同步 ¹
11.164.87.151%2C60023%2C1554712946764.default.1572503983124	RUNNING	122 B 0.00 MB/s	10 B 0.00 MB/s	291 ms	178 ms	local
11.160.211.12%2C60023%2C1554712953665.default.1572503951253	RUNNING	0 0.00 MB/s	0 0.00 MB/s	0 ms	0 ms	local
11.162.179.138%2C60023%2C1554712951716.default.1572503580159	RUNNING	0 0.00 MB/s	0 0.00 MB/s	0 ms	0 ms	local
11.164.87.151%2C60023%2C1554712946764.default.1572503383065	SUCCEEDED	0 0.00 MB/s	0 0.00 MB/s	0 ms	0 ms	local

订阅MQ指定Topic的消息，对消息中body进行转化，转化成字符串，就能得到KV对应的JSON格式

[illegible]

任务创建同步表支持如下的填写格式

```
t1 {"tag": "xxxx"} // t1表的message带上xxxx的标签，方便业务消息订阅时对tag标签进行消息过滤
t2 {"tag": "xxxx", "topic": "topic1"} // t2表的数据同步到topic1中，不特殊指定就走默认topic
t3 // t3表默认tag为空字符串
```

BDS解析HBase WAL日志中的数据，并对每个KV转化成对应的JSON格式写入到MQ，JSON格式如下所示

```
{
  "namespace": "default",
  "tableName": "t1",
  "rowKey": "727878",
  "opCode": 4,
  "family": "6631",
  "qualifier": "61",
  "value": "7631",
  "timestamp": 1572503986205
}
```

43

KV所包含的rowkey、family、qualifier、value在HBase都是byte数组形式进行存储，BDS同步KV默认会把byte数组转化成HexString，可以通过如下方式将HexString转化会原生的数组

```
import org.apache.commons.codec.binary.Hex;
...
Hex.decodeHex(rowkey.toCharArray())
Hex.decodeHex(family.toCharArray())
...
```

timestamp 表示的是KV的版本，业务没有特殊指定，默认是KV入库的时间戳

opCode 表示的是KV的类型，不同的值对应如下信息

```
enum Type {
    Put((byte) 4),

    Delete((byte) 8),

    DeleteFamilyVersion((byte) 10),

    DeleteColumn((byte) 12),

    DeleteFamily((byte) 14);
}
```

RDS实时数据同步HBase、Phoenix

前提条件

1. 购买BDS数据迁移同步服务，配置BDS操作页面账户密码，并登陆BDS操作页面
2. 打通BDS和HBase迁移集群的网络
3. 添加HBase集群数据源

版本

1. 自建HBase1.x、2.x
2. EMR HBase
3. 标准版云HBase、增强版云HBase (单机版本、集群版本)
4. 云HBase Phoenix

任务创建

1. 登陆BDS操作页面，任务管理->RDS实时数据订阅



2. 选择对应的DTS订阅通道和HBase(Phoenix)集群，指定需要同步表的mapping

DTS实时数据订阅

通道名

test

DTS数据源

rds_payment

目标数据源(HBase、Phoenix)

bigdata-test

表映射

```
{
  "config": {
    "startOffset": 1570032000
  },
  "mapping": [
    {
      "columns": [
        {
          "name": "f1:*"
        }
      ]
    }
  ],
  "config": {
    "skipDelete": false
  }
}
```

创建同步通道

参数说明

HBase表映射

```
{
  "mapping": [
    "srcTableName": "hhh_test.test",
    "targetTableName": "default:_test",
    "columns": [
      {
        "name": "cf1:hhh",
        "value": "{{ concat(title, id) }}"
      },
      {
        "name": "cf1:title",
        "value": "title"
      },
      {
        "name": "cf1:*"
      }
    ],
    "config": {
      "skipDelete": true
    }
  },
}
```

```
"rowkey": {
"value": "{{ concat('idg', id) }}"
}
}
]
}
```

参数	描述	必选
mapping[y].srcTableName	RDS源表表名	是
mapping[y].targetTableName	HBase目标表名	是
mapping[y].columns	RDS表和HBase表列的对应	是
mapping[y].columns[x].name	对应的HBase的列名	是
mapping[y].columns[x].value	对应的HBase列的计算表达式，该计算表达式为jtwig语法，当用户需要对源表的列进行简单计算得到rowkey时可以用	是
mapping[y].config	表的同步策略	否
mapping[y].rowkey	HBase表的rowkey的生成规则	是

- 支持简单的表达式，如下

```
{
"name": "cf1:hjh",
"value": "{{ concat(title, id) }}"
}
```

- 支持动态列，没有匹配到的列会走默认的匹配

```
{
"name": "cf1:*",
}
```

- 支持指定订阅的起始时间(DTS订阅通道中包含指定时间戳以后的数据)

```
{
"config": {
"startOffset":1569463200 // 秒
},
"mapping": [
"srcTableName": "hhh_test.test",
"targetTableName": "default:test",
"columns": [
{
"name": "cf1:*"
}
]
```

```
],
"config": {
"skipDelete": true
},
"rowkey": {
"value": "{{ concat('idg', id) }}"
}
}
]
}
```

- DML支持情况

操作	支持	备注
INSERT	是	对应HBase的PUT
UPDATE	是	对应HBase的PUT
DELETE	是	用户可以配置是否同步源端的DELETE，默认不同步

Phoenix表映射

```
{
"mapping": [
{
"srcTableName": "hhh_test.phoenix_test",
"targetTableName": "phoenix_test",
"config": {
"skipDelete": true
},
"columns": [
{
"name": "id",
"isPk": true
},
{
"name": "title",
"value": "title"
},
{
"name": "ts",
"value": "ts"
},
{
"name": "datetime",
"value": "datetime"
}
]
}
]
}
```


参数	描述	必选
mapping[y].srcTableName	RDS源表表名	是
mapping[y].targetTableName	Phoenix目标表名	是
mapping[y].columns	RDS表和Phoenix表列的对应	是
mapping[y].columns[x].name	Phoenix的列名	是
mapping[y].columns[x].value	对应RDS列名	是
mapping[y].columns[x].isPk	指定主键列	是
mapping[y].config	表的同步策略	否
mapping[y].rowkey	HBase表的rowkey的生成规则	是