

MMAv3使用文档

[MMA介绍](#)

[功能](#)

[原理](#)

[Hive数据迁移原理](#)

[通过Hive UDTF迁移数据到MaxCompute](#)

[通过OSS迁移Hive数据到MaxCompute](#)

[MaxCompute数据迁移原理](#)

[同region project迁移](#)

[emr + dlf + oss迁maxcompute](#)

[跨region project迁移](#)

[MMA任务与子任务](#)

[安装与配置](#)

[环境准备](#)

[MaxCompute权限准备](#)

[程序](#)

[配置文件](#)

[启动](#)

[停止](#)

[初次访问与MMA配置](#)

[通过Hive UDTF迁移HIVE数据](#)

[准备事项](#)

[网络环境要求](#)

[创建用于迁移数据的HIVE UDTF](#)

[如果Hive配置了kerberos访问认证](#)

[添加数据源](#)

[创建数据迁移任务](#)

[迁移多个table](#)

[迁移多个分区](#)

迁移单个database

通过OSS迁移Hive数据

准备事项

网络环境要求

如果Hive配置了kerberos访问认证

添加数据源

MaxCompute数据迁移

准备事项

同region project迁移

如果能够通过同一个账号访问源和目标project

如果不能通过同一个账号访问源和目标project

emr + dlf + oss迁maxcompute

跨region project迁移

添加数据源

创建数据迁移任务

迁移多个table

迁移多个分区

迁移单个database(project)

分区过滤表达式说明

迁移任务查看与操作

增量迁移

自动增量迁移

手动增量迁移

MMA介绍

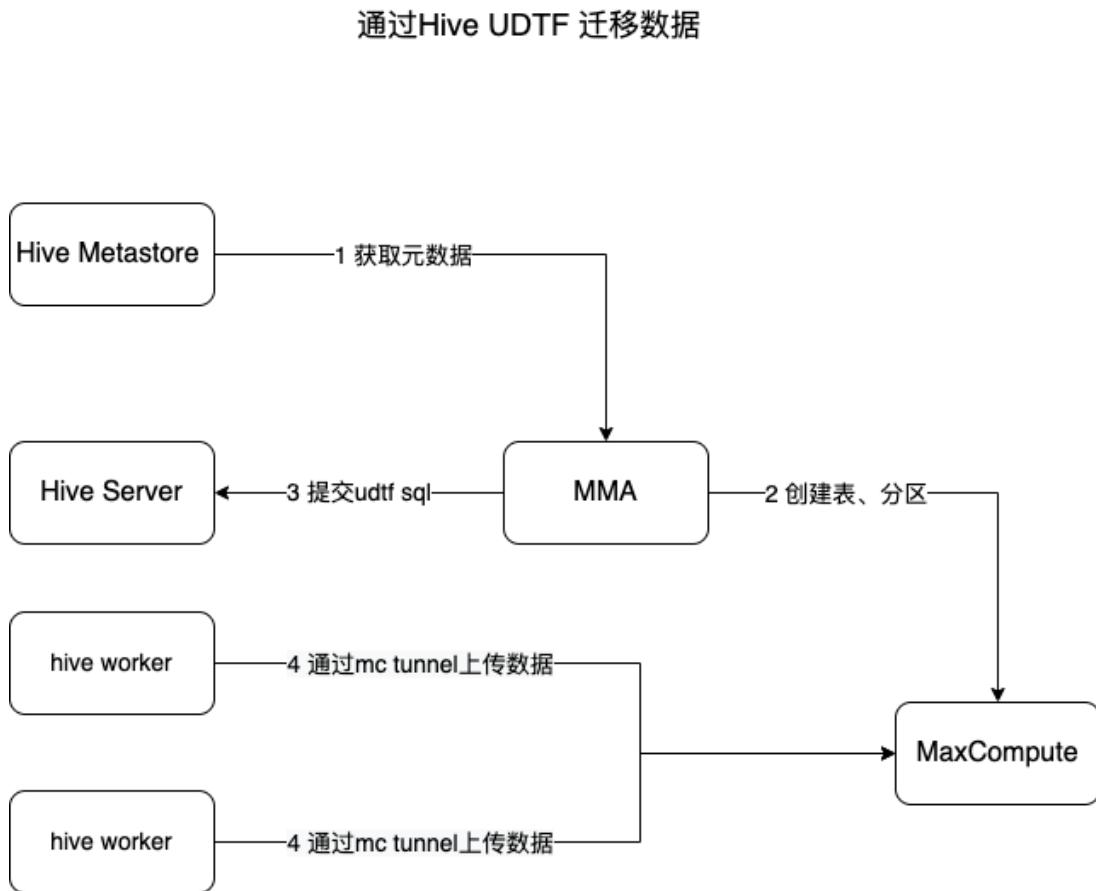
功能

1. Hive数据迁移到MaxCompute
2. MaxCompute跨project数据迁移
3. emr + dlf + oss迁移MaxCompute project

原理

Hive数据迁移原理

通过Hive UDTF迁移数据到MaxCompute



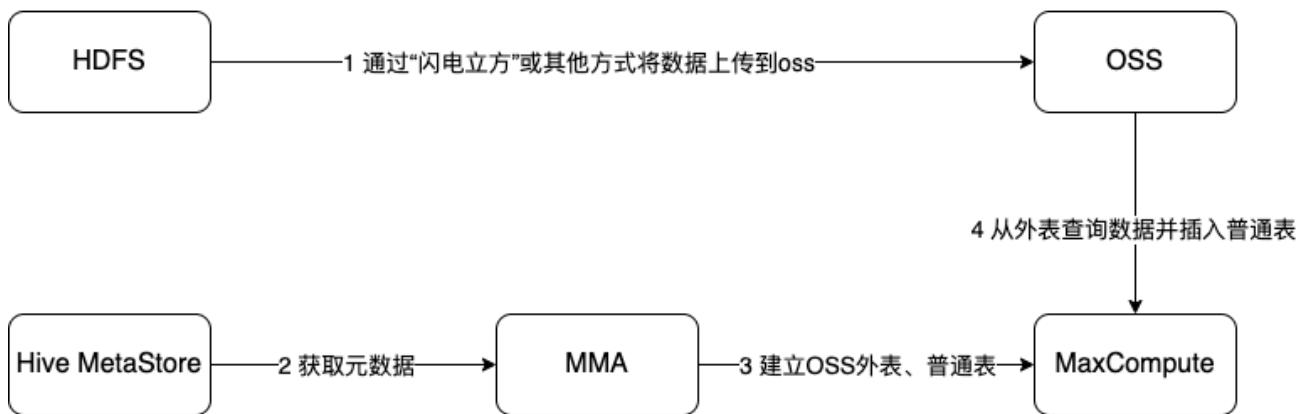
这种迁移方式通过Hive的分布式能力，实现Hive数据向MaxCompute的高并发传输。数据迁移的主要过程如下

1. MMA通过Hive MetaStore获取元数据：有哪些表、表的schema、分区信息
2. MMA在MaxCompute端根据获取到的schema建表、分区
3. MMA向Hive提交执行UDTF的SQL
4. UDTF会调用MaxCompute的tunnel sdk向MaxCompute写表数据
5. 数据校验。目前提供的校验方式是在Hive端和MaxCompute端对同一个表(或多个分区)执行select count(*), 然后对比两端的行数。

这种迁移方式的前置条件是：Hive集群各节点需要能够访问MaxCompute

通过OSS迁移Hive数据到MaxCompute

通过OSS迁移Hive数据



这种迁移方式会先将数据迁移到oss, 然后通过MaxCompute读取oss数据。主要过程如下

1. 通过阿里云"闪电立方"服务或`distcp`、`juicesync`将数据从HDFS迁移到OSS
2. MMA通过Hive MetaStore获取元数据：有哪些表、表的schema、分区信息等
3. MMA在MaxCompute端根据获取到的schema, oss路径信息建MaxCompute oss外表和外表对应的普通表
4. 通过：`insert 普通表 from select oss外表` 将数据从oss导入到MaxCompute

MaxCompute数据迁移原理

同region project迁移

1. 获取源project里的所有表、分区，在目的project里建表、分区
2. 使用sql：`insert overwrite 目的表 from 源表` 迁移数据

emr + dlf + oss迁maxcompute

这种场景需要通过maxcompute + dlf建立maxcompute的external project，然后迁移操作与“同region project迁移”一致。

跨region project迁移

跨region project迁移使用的是maxcompute的copytask任务，copytask可以将表数据从一个region的project，复制到另一个region的project。

这种迁移方式的前提条件是源project开启允许使用copytask的开关。

MMA任务与子任务

1. MMA可以以单个db、多个table、多个partition为单位提交迁移任务
2. 迁移任务会以“分区”和“非分区表”为单位进行子任务分割，子任务是实际执行迁移操作的单位。一个子任务迁移一个非分区表或一个/多个分区。

安装与配置

环境准备

1. 操作系统: linux
2. java版本 >= 8
3. mysql >= 5.7

其他环境根据要求根据不同的数据源会有所不同，具体详见不同的数据源迁移说明。

MaxCompute权限准备

程序

MMAv3.jar

配置文件

新建config.ini文件,示例文件如下

```
Properties | 复制代码

1 [mysql]
2 host = mysql-host
3 port = 3306
4 ; can be any database name
5 db = mmav3
6 username = user
7 password = pass
8
9 [mma]
10 listening_port = 6060
```

启动

建议使用nohup启动程序

```
Shell | 复制代码

1 nohup java -jar MMAv3.jar -c config.ini > nohup.log &
```

停止

可以直接找到mma程序的pid,然后kill掉

Shell | 复制代码

```
1 ps aux | grep MMAv3.jar | grep -v grep | awk '{print $2}' | xargs kill -9
```

初次访问与MMA配置

mma程序启动后，可以通过在浏览器端打开 "<http://ip:6060>" 进行访问(6060通过config.ini中的listening_port进行配置)。浏览器打开mma后，会跳入mma的配置页面。

配置页面内容如下：

MMA配置

保存 | 重置

配置项	配置值	描述
mc.endpoint	http://service.cn.maxcompute.aliyun.com/api	maxcompute endpoint (必填)
mc.data.endpoint	http://service.cn.maxcompute.aliyun-inc.com/api	用于数据传输的maxcompute endpoint (可选)
mc.tunnel.endpoint	http://dt.cn-zhangjiaokou.maxcompute.aliyun-inc.com	maxcompute tunnel endpoint (可选)
mc.auth.access.id	[REDACTED]	maxcompute access id (必填)
mc.auth.access.key	请输入	maxcompute access key (必填)
mc.default.project	mma_test	maxcompute default project (必填)
mc.projects	mma_test,mma_test2	要迁往的maxcompute项目列表 (可选)
task.max.num	3	数据搬迁任务最大并发量 (可选)

配置项的说明：

这里的MaxCompute配置项是目的MaxCompute的信息

配置项	说明
mc.endpoint	用于MMA访问MaxCompute的endpoint, 要求MMA所在机器能够连通mc.endpoint, 具体的endpoint信息可以参考 https://help.aliyun.com/document_detail/34951.html
mc.data.endpoint	可选 <ol style="list-style-type: none"> 通过Hive UDTF迁移数据时用于UDTF访问MaxCompute, 这时要求hive集群的节点能够连通这个地址 MaxCompute跨区域迁移时, 这个地址为目的端MaxCompute所在region的VPC或经典网endpoint MaxCompute同区域迁移时不填
mc.tunnel.endpoint	可选 <ol style="list-style-type: none"> 通过Hive UDTF迁移数据时用于UDTF访问MaxCompute, 这时要求hive集群的节点能够连通这个地址 MaxCompute跨区域迁移时, 这个地址为目的端MaxCompute所在region的VPC或经典网tunnel endpoint MaxCompute同区域迁移时不填
mc.auth.access.id	用于访问MaxCompute的Access ID
mc.auth.access.key	用于访问MaxCompute的Access Key
mc.default.project	MMA会使用这个项目的配额在MaxCompute上执行sql
mc.projects	要迁往的MaxCompute项目名列表。多个项目名之间以英文逗号(,)分隔
task.max.num	用于迁移数据的MMA任务最大并发数, 这个参数是调节迁移速度的重要参数之一 如 <ol style="list-style-type: none"> 在通过Hive UDTF迁移数据时, 这个参数是MMA向 Hive同时提交的sql任务最大数量。 在通过OSS迁移Hive数据时, 这个参数是MMA执行 <code>insert overwrite ... select</code> sql的最大数量 在同MC 同reqion项目迁移时, 这个参数是执行“<code>insert overwrite select</code>” sql的最大数量

通过Hive UDTF迁移HIVE数据

准备事项

网络环境要求

1. Hive集群各节点需要能够访问MaxCompute
2. MMA所在的机器能够访问hive metastore server、hive server

创建用于迁移数据的HIVE UDTF

1. 下载udtf的jar包，可以在“帮助”页面下载。
2. 上传mma-udtf.jar到HDFS

Shell | 复制代码

```
1 hdfs dfs -put -f mma-udtf.jar hdfs:///tmp/
```

3. 使用beeline或hive命令登录hive、创建Hive UDTF

SQL | 复制代码

```
1 DROP FUNCTION IF EXISTS default.odps_data_dump_multi;
2 CREATE FUNCTION default.odps_data_dump_multi AS 'com.aliyun.odps.mma.io.McD
ataTransmissionUDTF' USING JAR 'hdfs:///tmp/mma-udtf.jar';
```

如果Hive配置了kerberos访问认证

需要将几个文件拷贝到MMA所在的机器

1. hive.keytab文件
2. gss-jass.conf文件，注意：gss-jass.conf中有keytab文件的路径，要与MMA所在机器上hive.keytab文件所在路径一致
3. krb5.conf文件，注意：krb5.conf中有kdc地址，MMA所在的机器要能够访问这些地址

添加数据源

第一步：点击“添加数据源”按钮，进入添加数据源页面

数据源名	类型	db数	table数	partition数	最新更新	操作
mma_test	ODPS	1	202	55353	2022-09-15 09:56:09	更新

第二步：数据源类型选择“HIVE”

添加数据源

1 选择数据源类型 2 配置数据源

* 数据源类型:

HIVE

下一步

第三步：配置数据源

添加数据源

1 选择数据源类型 ————— 2 配置数据源

* 数据源名

* 数据源类型

* hive metastore url

hive metastore client socket timeout

* hive jdbc url

* hive jdbc user name

hive jdbc password

Hive数据源配置项说明如下

配置名	配置值
数据源名	数据源的名称，可随意自定义，注意不能包含字符、数字、汉字之外的特殊字符
hive metastore url	如: thrift://192.168.0.212:9083
hive jdbc url	如: jdbc:hive2://192.168.0.212:10000/default;principal=hive/emr-header-1.cluster-48497@EMR.48497.COM
hive jdbc user name	必填
hive metastore是否开启了kerberos认证	如果是，则下面kerberos相关的配置需要填写
kerberos principal	要与kr5.conf里的kdc_realm值保持一致
kerberos keytab	keytab文件在MMA机器上的路径

又1位直	
kerberos gss-jass.conf文件位置	gss-jass.conf文件在MMA机器上的路径
kerberos krb5.conf文件位置	krb5.conf文件在MMA机器上的路径
单个任务处理的最多分区数量	默认50。一次MMA任务迁移的分区数量，通过批量的分区迁移，可以减少提交hive sql的次数，节约hive sql提交时间
单个任务处理的最大数量	默认5。一次MMA任务迁移的所有分区的大小之和的上限
hive job配置	用于mr, spark, tez等引擎。 注意: 默认是mr任务的一些配置，如果hive用的引擎不是mr，则需要通过指定"hive.execution.engine"值来指定hive用的引擎，并且用于不同引擎的任务参数需要自己调整。这项配置用于解决yarn container内存不足、指定spark运行队列等问题
meta api访问并发量	默认值为3，访问Hive MetaStore的并发量，用于提高获取Hive元数据的速度
数据库白名单	需要迁移的Hive database, 多个值之间以英文逗号分隔
数据库黑名单	不需需要迁移的Hive database, 多个值之间以英文逗号分隔
表黑名单	不需要迁移的Hive table。单个表的格式为"db名字.table名字", 多个table之间以英文逗号分隔
表白名单	需要迁移的Hive table。单个表的格式为"db名字.table名字", 多个table之间以英文逗号分隔

第四步：点击"提交"按钮，如果所填的配置无误、MMA所在机器能够访问metastore url, jdbc url, 这时MMA会通过metastore url拉取Hive的元数据：库、表、分区信息。

否则，将会报错。这时需要检查个配置项，重新填写并提交配置

第五步：等拉取元数据的进度条为100%后，页面会跳到"数据源"页面

创建数据迁移任务

MMA可以创建三个级别的迁移任务

1. 单库，迁移单个database
2. 多表，迁移1个或多个table
3. 多分区，迁移1个或多个partition

迁移多个table

1. 进"数据源"页面，点击表格里想要迁移的数据源的名称，进入数据源的详情页面
2. 点击要迁移的"数据库名"，进入"数据库"的详情页
3. 选择要迁移的table，点击"新建迁移任务"
4. 这时会弹出"新建迁移任务"的对话框，如下图所示

新建迁移任务

X

* 名称: mma_test@2022-11-1/14:36:18

x

数据源: hive3_2

库名: mma_test

* 任务类型: hive UDTF

* mc项目: mma_test3

table列表: test_orc_partitioned_10x1k,test_rcfile_partitioned_10x1k

开启校验:

增量更新:

只迁schema:

分区过滤:

test_orc_partitioned_10x1k

p1 > 'uXaBc' and p2 > 10



+ 添加一行数据

表名映射:

test_orc_partitioned_10x1k

test_orc_10x1k



+ 添加一行数据

取消

确定

有关新建任务配置项的说明:

配置	说明
名称	最好填入有意义的任务名称，以便于整理迁移记录
任务类型	hive数据源目前只有hive UDTF类型的任务
mc项目	目的mc项目
table列表	要迁移的table列表，多个table之间以英文逗号分隔
开启校验	是否开启校验
增量更新	是否开启增量更新，开启后分区表已经迁移过的分区不会被重新迁移
只迁schema	只迁移表结构、分区值
分区过滤	详见“分区过滤表达式说明”
表名映射	一个table迁移到目的project后的名字

5. 点击"确定"后，如果迁移任务的配置没有错误，则新的迁移任务可以在"迁移任务/任务列表"中查看，相应的子任务可以在"迁移任务/子任务列表查看"

迁移多个分区

1. 进"数据源"页面，点击表格里想要迁移的数据源的名称，进入数据源的详情页面
2. 点击要迁移的"数据库名"，进入"数据库"的详情页
3. 点击"partition列表"，进入partition列表tab页面
4. 选择要迁移的partition，点击"新建迁移任务"，这是会弹出"新建迁移任务"的对话框，如下图所示

新建迁移任务

X

* 名称:

测试迁移多个分区

X

数据源:

hive3_2

库名:

mma_test

* 任务类型:

hive UDTF

V

* mc项目:

mma_test3

V

开启校验:



只迁schema:



partition列表:

test_partitioned_100x10k.p1=ARBqa/p2=4758
test_partitioned_100x10k.p1=AhXOR/p2=27
test_partitioned_100x10k.p1=AnBRB/p2=2001

表名映射:

+ 添加一行数据

取消

确定

5. 点击"确定"后，如果迁移任务的配置没有错误，则新的迁移任务可以在"迁移任务/任务列表"中查看，相应的子任务可以在"迁移任务/子任务列表查看"

迁移单个database

1. 进"数据源"页面，点击表格里想要迁移的数据源的名称，进入数据源的详情页面
2. 如下图，点击要迁移的database所在行的"迁移"操作

hive2

[数据信息](#) [配置信息](#)

数据库: 1

表: 21

分区: 159

最新更新时间: 2022-09-08 03:10:18

状态

库名	table总数	迁移完成table数	分区总数	迁移完成分区数	大小	行数	状态	操作
mma_test	21	1	159	11	2.45 GB	1171002	部分完成	迁移

第 1-1 条/总共 1 条 < 1 >

3. 这时会弹出"新建迁移任务"的对话框, 如下图所示

新建迁移任务

X

* 名称: 测试迁移整库 ×

数据源: hive3_2

库名: mma_test

* 任务类型: hive UDTF ▼

* mc项目: mma_test3 ▼

table白名单 table黑名单

table白名单:

开启校验:

增量更新:

只迁schema:

分区过滤: ▼ × □ □

+ 添加一行数据

表名映射: ▼ × □ □

+ 添加一行数据

取消

确定

有关新建任务配置项的说明:

配置	说明
名称	最好填入有意义的任务名称，以便于整理迁移记录
任务类型	hive数据源目前只有hive UDTF类型的任务
mc项目	目的mc项目
table白名单	要迁移的table列表，多个table之间以英文逗号分隔
table黑名单	不迁移的table列表，多个table之间以英文逗号分隔
开启校验	是否开启校验
增量更新	是否开启增量更新，开启后分区表已经迁移过的分区不会被重新迁移
只迁schema	只迁移表结构、分区值
分区过滤	详见“分区过滤表达式说明”
表名映射	一个table迁移到目的project后的名字

5. 点击"确定"后，如果迁移任务的配置没有错误，则新的迁移任务可以在"迁移任务/任务列表"中查看，相应的子任务可以在"迁移任务/子任务列表查看"

通过OSS迁移Hive数据

准备事项

通过阿里云"闪电立方"服务或[distcp](#)、[juicesync](#)将数据从HDFS迁移到OSS

网络环境要求

MMA所在的机器能够访问hive metastore server、hive server

如果Hive配置了kerberos访问认证

需要将几个文件拷贝到MMA所在的机器

1. hive.keytab文件
2. gss-jass.conf文件, 注意: gss-jass.conf中有keytab文件的路径, 要与MMA所在机器上hive.keytab文件所在路径一致
3. krb5.conf文件, 注意: krb5.conf中有kdc地址, MMA所在的机器要能够访问这些地址

Hive数据源配置项说明如下

添加数据源

第一步：点击“添加数据源”按钮，进入添加数据源页面

The screenshot shows a table with columns: Data Source Name, Type, db数, table数, partition数, Last Updated, and Operations. There is one row for 'mma_test' which is of type 'ODPS'. The 'Operations' column contains a 'Update' link. A blue button labeled 'Add Data Source' is located in the top right corner of the table area, with a red box highlighting it.

数据源名	类型	db数	table数	partition数	最新更新	操作
mma_test	ODPS	1	202	55353	2022-09-15 09:56:09	更新

第二步：数据源类型选择“HIVE_OSS”

The screenshot shows the first step of a two-step wizard. The title is 'Add Data Source'. It has two tabs: '1 选择数据源类型' (selected) and '2 配置数据源'. Below the tabs is a section titled 'Data Source Type:' with a dropdown menu containing 'HIVE_OSS'. A blue 'Next Step' button is at the bottom.

第三步：配置数据源

添加数据源

1 选择数据源类型 ————— 2 配置数据源

* 数据源名

* 数据源类型

* oss endpoint internal, use in mc sql

* oss endpoint external, use in mma if mma can not use internal endpoint

* oss access id

* oss access key

* oss bucket

oss路径, 去除bucket和db名字部分, a/b/

Hive OSS数据源配置项说明如下

配置名	配置值
数据源名	数据源的名称, 可随意自定义, 注意不能包含字符、数字、汉字之外的特殊字符
oss endpoint internal	oss 用于经典网络或VPC网络访问的endpoint
oss endpoint external	oss 用于外网访问的endpoint
oss access id	oss账号的access id
oss access key	oss账号的access key
oss bucket	表文件所在的bucket

oss路径, 去除bucket和db名字部分, a/b/	如oss://user:pass@oss-cn-zhangjiakou-internal.aliyuncs.com/mma-test/hive/test_hive.db/test_hive_table/ mma-test是bucket名, test_hive为hive db名, test_hive_table是table名, 则oss路径为 "hive"
oss path db template, like "\${db}.db"	如“oss”路径示例, hive_db名我test_hive.db, 则改项配置值为\${db}.db
maxcompute 迁移任务sql参数	用于maxcompute sql任务的flag, 默认为 <pre>{ "odps.sql.hive.compatible": "true", "odps.sql.split.hive.bridge": "true" }</pre> 一般情况下不需要修改
hive metastore url	如: thrift://192.168.0.212:9083
hive jdbc url	如: jdbc:hive2://192.168.0.212:10000/default;principal=hive/emr-header-1.cluster-48497@EMR.48497.COM
hive jdbc user name	必填
hive metastore是否开启了kerberos认证	如果是, 则下面kerberos相关的配置需要填写
kerberos principal	要与kr5.conf里的kdc_realm值保持一致
kerberos keytab文件位置	keytab文件在MMA机器上的路径
kerberos gss-jass.conf文件位置	gss-jass.conf文件在MMA机器上的路径
kerberos krb5.conf文件位置	krb5.conf文件在MMA机器上的路径
单个任务处理的最多分区数量	默认50。一次MMA任务迁移的分区数量, 通过批量的分区迁移, 可以减少提交hive sql的次数, 节约hive sql提交时间

单个任务处理的最大数量	默认15。一次MMA任务迁移的所有分区的大小之和的上限
hive job配置	用于mr, spark, tez等引擎。 注意: 默认是mr任务的一些配置, 如果hive用的引擎不是mr, 则需要通过指定"hive.execution.engine"值来指定hive用的引擎, 并且用于不同引擎的任务参数需要自己调整。这项配置用于解决yarn container内存不足、指定spark运行队列等问题
meta api访问并发量	默认值为3, 访问Hive MetaStore的并发量, 用于提高获取Hive元数据的速度
数据库白名单	需要迁移的Hive database, 多个值之间以英文逗号分隔
数据库黑名单	不需要迁移的Hive database, 多个值之间以英文逗号分隔
表黑名单	不需要迁移的Hive table。单个表的格式为"db名字.table名字", 多个table之间以英文逗号分隔
表白名单	需要迁移的Hive table。单个表的格式为"db名字.table名字", 多个table之间以英文逗号分隔

第四步：点击“提交”按钮，如果所填的配置无误、MMA所在机器能够访问metastore url, jdbc url，这时MMA会通过metastore url拉取Hive的元数据：库、表、分区信息。
否则，将会报错。这时需要检查个配置项，重新填写并提交配置

第五步：等拉取元数据的进度条为100%后，页面会跳到“数据源”页面

MaxCompute数据迁移

准备事项

同region project迁移

如果能够通过同一个账号访问源和目标project

源project权限要求：对project有List、CreateInstance权限，对table有Describe, Select权限

目的project权限要求：对project有Read, CreateInstance, CreateTable权限，或对所有table有Select, Alter, Update权限

如果不能通过同一个账号访问源和目标project

1. 需要通过“[跨项目访问资源权限控制](#)”允许目的project的账号访问源project。并依次将要迁移的table通过"add table <table_name> to package <package_name> with privileges select"命令添加select权限到package
2. 源project账号权限要求：对project有List、CreateInstance权限，对table有Describe, Select权限
3. 目的project账号权限要求：对project有Read, CreateInstance, CreateTable权限，或对所有table有Select, Alter, Update权限

emr + dlf + oss迁maxcompute

需要先构建maxcompute的湖仓一体，具体可参考

https://help.aliyun.com/document_detail/205439.html

之后的迁移操作与“同region project迁移”一致

跨region project迁移

要求源project开启copytask开关，具体的请联系阿里云同学

对账号权限要求

1. 源project账号权限要求：对project有List、CreateInstance权限，对要迁移的table有Describe, Select权限
2. 目的project账号权限要求：对project有Read, CreateInstance, CreateTable权限

添加数据源

第一步

The screenshot shows a table with a single row of data. The columns are: 数据源名 (Data Source Name), 类型 (Type), db数 (db count), table数 (table count), partition数 (partition count), 最新更新 (Last Update), and 操作 (Operations). The data in the table is: mma_test, ODPS, 1, 202, 55353, 2022-09-15 09:56:09, and a blue '更新' (Update) button. A red box highlights the blue '添加数据源' (Add Data Source) button in the top right corner.

数据源名	类型	db数	table数	partition数	最新更新	操作
mma_test	ODPS	1	202	55353	2022-09-15 09:56:09	更新

第二步

The screenshot shows the first step of the 'Add Data Source' wizard. It has two tabs: '1 选择数据源类型' (Step 1: Select Data Source Type) and '2 配置数据源' (Step 2: Configure Data Source). The first tab is active. It includes a dropdown menu labeled 'MAXCOMPUTE' and a blue '下一步' (Next) button.

第三步



选择数据源类型



配置数据源

* 数据源名

mma_test2



* 数据源类型

ODPS

* maxcompute endpoint

请输入

* maxcompute access id

[REDACTED]



* maxcompute access key

[REDACTED]



* maxcompute default project(用于执行sql的project)

请输入

* 要迁移的maxcompute projects

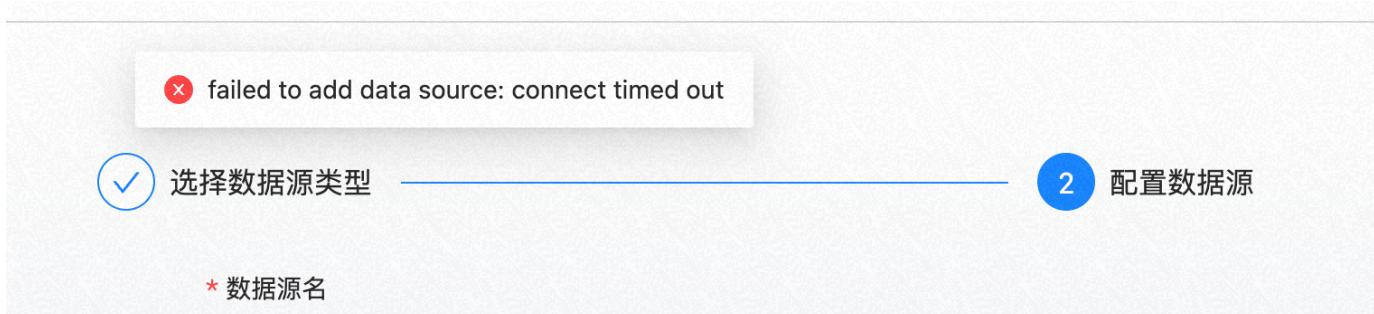
请输入

相关配置说明:

配置	说明
数据源名	要求: 英文字母、数字, 全局唯一
maxcompute endpoint	源project所在region的endpoint , 具体可参考 https://help.aliyun.com/document_detail/34951.html
maxcompute access id	用于访问MaxCompute的Access ID
maxcompute access key	用于访问MaxCompute的Access Key
maxcompute default project	用于执行有关源project sql的project, 如源project为A, default project为B, 则执行select * from A.table的sql语句会使用B的配额组
要迁移的maxcompute projects	要迁移的project列表, 多个project之间以英文逗号分隔
instance number of one copyTask	仅用于跨region project迁移。 跨region project迁移时每个copytask使用的并发数
maxcompute 迁移任务sql参数	仅用于同region project迁移。 执行迁移sql时使用的一些flag, 大多数时这个配置使用默认值即可。 如果任务执行期间有sql报错, 可以把logview的信息发给阿里云的同学。
单个任务处理的最多分区数量	仅用于同region project迁移。 一个mma子任务可以同时同一个表的多个分区, 这个配置指定可以同时迁移的分区的数量。
meta api访问并发量	获取源project时的访问并发量, 建议值20
表黑名单	格式为 db.table, 多个表名之间以英文逗号分隔
表白名单	格式为 db.table, 多个表名之间以英文逗号分隔

第四步：点击“确定”

如果有配置错误，提示错误原因。这时需要检查配置，修改配置后重新点击"确定"。



如果配置正确，MMA会拉取元数据，元数据拉取完毕后，页面会自动跳转到数据源页面



创建数据迁移任务

迁移多个table

1. 进"数据源"页面，点击表格里想要迁移的数据源的名称，进入数据源的详情页面
2. 点击要迁移的"数据库名"，进入"数据库"的详情页
3. 选择要迁移的table，点击"新建迁移任务"
4. 这时会弹出"新建迁移任务"的对话框，如下图所示

新建迁移任务

X

* 名称: 第一批次 X

数据源: mma_test2测试

库名: mma_test2

* 任务类型: mc同region ▼

* mc项目: mma_test ▼

table列表: copytask_speed_test,lemming_test_20220228

开启校验:

增量更新:

只迁schema:

分区过滤: lemming_test_20220228 ▼ dt > '2019-01-26' and dt < '201 X □ □

+ 添加一行数据

表名映射: + 添加一行数据

取消

确定

有关任务的配置说明如下

配置	说明
名称	最好填入有意义的任务名称，以便于整理迁移记录
任务类型	同region项目迁移 跨region项目迁移 mc校验：对比源和目的project所有相同表的数据
mc项目	目的mc项目
table列表	要迁移的table列表，多个table之间以英文逗号分隔
开启校验	是否开启校验
增量更新	是否开启增量更新，开启后分区表已经迁移过的分区不会被重新迁移
只迁schema	只迁移表结构、分区值
分区过滤	详见“分区过滤表达式说明”
表名映射	一个table迁移到目的project后的名字

5. 点击"确定"后，如果迁移任务的配置没有错误，则新的迁移任务可以在"迁移任务/任务列表"中查看，相应的子任务可以在"迁移任务/子任务列表查看"

迁移多个分区

1. 进"数据源"页面，点击表格里想要迁移的数据源的名称，进入数据源的详情页面
2. 点击要迁移的"数据库名"，进入"数据库"的详情页
3. 点击"partition列表"，进入partition列表tab页面
4. 选择要迁移的partition，点击"新建迁移任务"，这是会弹出"新建迁移任务"的对话框，如下图所示

新建迁移任务 X

* 名称: 第二批次 X

数据源: mma_test2测试

库名: mma_test2

* 任务类型: mc同region ▼

* mc项目: mma_test ▼

开启校验:

只迁schema:

partition列表: copytask_speed_test.p1=1GFXCJ7LCF/p2=9023049806080075876
copytask_speed_test.p1=3BDY3YA3BE/p2=8699746166586838680

表名映射: + 添加一行数据

取消 确定

5. 点击“确定”后，如果迁移任务的配置没有错误，则新的迁移任务可以在“迁移任务/任务列表”中查看，相应的子任务可以在“迁移任务/子任务列表查看”

迁移单个database(project)

如果一个project的数据量特别大，不建议直接迁移整个project，可以分批次建立“多个table”的任务。

1. 进“数据源”页面，点击表格里想要迁移的数据源的名称，进入数据源的详情页面
2. 如下图操作

数据源 / mma_test2测试

mma_test2测试

数据信息 配置信息

数据库: 1 表: 68 分区: 26267
最新更新时间: 2022-09-16 02:30:06

库名	table总数	迁移完成table数	分区总数	迁移完成分区数	大小	行数	状态	操作
mma_test2	68	0	26267	0	142.91 GB	--	未迁移	迁移

第 1-1 条/总共 1 条 < 1 >



3 填写新建迁移任务的配置

新建迁移任务

X

* 名称： ×

数据源：

库名：

* 任务类型： ▼

* mc项目： ▼

table白名单 table黑名单

table白名单：

开启校验：

增量更新：

只迁schema：

分区过滤： + 添加一行数据

表名映射： + 添加一行数据

取消 确定

配置说明如下

配置	说明
名称	最好填入有意义的任务名称，以便于整理迁移记录
任务类型	同region项目迁移 跨region项目迁移 mc校验：对比源和目的project所有相同表的数据
mc项目	目的mc项目
table白名单	要迁移的table列表，多个table之间以英文逗号分隔
table黑名单	不迁移的table列表，多个table之间以英文逗号分隔
开启校验	是否开启校验
增量更新	是否开启增量更新，开启后分区表已经迁移过的分区不会被重新迁移
只迁schema	只迁移表结构、分区值
分区过滤	详见“分区过滤表达式说明”
表名映射	一个table迁移到目的project后的名字

5. 点击"确定"后，如果迁移任务的配置没有错误，则新的迁移任务可以在"迁移任务/任务列表"中查看，相应的子任务可以在"迁移任务/子任务列表查看"

分区过滤表达式说明

例子： `p1 >= '2022-03-04' and (p2 = 10 or p3 > 20) and p4 in ('abc', 'cde')`

例子说明：

1. p1, p2, p3为分区名
2. 分区值只有字符串和数字两种，字符串被双引号或单引号包裹。除int/bigint类型的分区列值外，其他所有类型的分区值都只能取字符串值。
3. 比较操作符包括: >, >=, = , <, <=, <>
4. 支持"in"操作符
5. 逻辑操作符包括: and, or

6. 支持括号

迁移任务查看与操作

可以在"迁移任务/子迁移任务列表里"看到子迁移任务的信息，子迁移任务的详情信息里可以看到任务执行的log

增量迁移

自动增量迁移

MMA的现在只能对"新增分区"进行自动增量迁移，已经迁移过的分区和非分区表无法进行增量迁移。

注意: 在进行增量迁移前，需要更新数据源，这时候MMA会拉取新增的表、分区等元数据。

实现增量迁移的方法有

1. 创建迁移任务时打开"增量迁移"开关，这是如果一个表已经迁移了部分分区，则已经迁移过的分区不会被重新迁移
2. 创建任务时通过"分区过滤"表达式，指定只迁移新增的分区

如果已经迁移过的分区源端数据有变动时，需要重新创建迁移这些分区的迁移任务，并把"增量迁移"开关关闭。

手动增量迁移

如果已经迁移过的分区有数据变化，这时候需要

1. 更新数据源
2. 进页面：目标数据源->目标database->partition列表 tab页面
3. 筛选出"元数据有更新"的partition，选择这些partition、创建迁移任务

table列表 [partition列表](#)

表名	分区值	是否	元数据有更新	数据最后修改时间	状态
<input type="checkbox"/> test_partitioned_25000x1k_copy	p1=003LNPF74Z/p2=6751365152984271782	是	--	否	2022-05-18 21:03:32 未迁移
<input type="checkbox"/> test_partitioned_25000x1k_copy	p1=004M1C8APT/p2=1011913816458194045	--	--	否	2022-05-18 21:03:31 未迁移
...

这里由于hive的元数据可能和数据不同步，所以即使更新了数据源、分区可能也会没有被标记为“有更新”，这种情况下需要

1. 先修正hive数据, 再在mma上更新数据源
2. 如果已知哪些分区是有更新的, 可以直接过滤出相应的分区, 然后创建迁移任务