ALIBABA CLOUD

阿里云

大数据计算服务

工具及下载

文档版本：20201023

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	**单击设置> 网络> 设置网络类型。**
粗体	表示按键、菜单、页面名称等UI元素。	**在结果确认页面，单击确定。**
`Courier字体`	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.查询编辑器

MaxCompute控制台提供查询编辑器，方便您快速执行SQL语句并进行数据分析操作。本文为您介绍如何通过查询编辑器使用MaxCompute服务。

概述

MaxCompute的查询编辑器集成在DataWorks的数据分析工具上，实现编辑MaxCompute SQL、查询数据、分析数据（电子表格）、在线分享数据及下载数据等功能。您可以通过查询编辑器快捷使用MaxCompute服务的相关功能：

- 支持编辑及运行SQL命令、授权命令（例如ACL授权）。
- 默认开放MaxCompute公开数据集，您可以通过查询编辑器基于公开数据集体验及测试MaxCompute。
- 支持在线使用电子表格Web Excel对数据查询的结果进行分析、下载或者分享给其他成员。

使用场景

查询编辑器的使用场景如下：

- 初次体验及测试MaxCompute的使用者：可以通过查询编辑器，使用公开数据集快速体验MaxCompute的核心功能。
- 数据分析师：您可以通过查询编辑器查询数据，并通过分析模式的Web Excel对查询结果进行分析。您也可以下载查询结果至本地，减少数据的流动，更好的保障数据安全。
- 安全管理员：MaxCompute项目右侧的**项目权限管理**提供了角色权限管理功能，但是正在试用过程中，很多场景需要通过命令行进行权限管理。安全管理员可通过查询编辑器快速执行大部分安全命令操作。

进入查询编辑器

1. 登录MaxCompute控制台，在左上角选择区域，单击**查询编辑**，即可进入查询编辑器界面。

   

2. 在**选择数据源**对话框，选择**数据源类型**为**MaxCompute**，**工作空间**为已创建好的项目空间。

   

   > **说明** 如果您选择的**工作空间**模式为标准模式，在查询编辑器中提交作业实际是在开发项目（带dev标识）中提交。

3. 单击**确认**，即可进入查询编辑器界面。

功能介介绍

查询编辑器界面分为查询模式和分析模式两种，单击左上角的**查询模式**或**分析模式**即可切换模式。该功能集成在DataWorks的数据分析工具上，详情请参见数据分析。基本界面功能如下：

- 查询模式

  

序号	描述
①	编辑器模式为查询模式，单击即可切换为分析模式。
②	工作空间名称。

序号	描述
③	工作空间下的所有表、使用过的表及公开数据集信息。
④	编辑器。支持执行SQL、ACL授权命令等。您进入查询编辑器界面时，可直接编辑脚本，无需先创建文件，同时也可以保存文件。
⑤	您可以查看已保存的查询文件、查询历史、查询日志及查询结果信息。

- 分析模式

序号	描述
①	编辑器模式为分析模式，单击即可切换为查询模式。
②	电子表格编辑及管理工具栏。
③	查询结果。

2.客户端

本文为您介绍如何通过客户端使用MaxCompute服务的功能。

客户端 odpscmd 命令行工具

下载安装

下载客户端，安装并配置客户端后即可正常使用。详情请参见安装并配置客户端。

> 说明
> - 请不要依赖客户端的输出格式执行任何解析工作。客户端的输出格式不承诺向前兼容，不同版本间的客户端命令格式及行为有差异。更多版本客户端下载请参见aliyun-odps-console。
> - 客户端从0.28.0版开始支持JDK 1.9，之前的版本只支持JDK 1.8。
> - 客户端从0.27.0版本开始支持MaxCompute 2.0新数据类型，详情请参见新数据类型。

安装并配置好客户端后，您可以借助命令行工具执行以下操作。更多客户端命令介绍，请参见常用命令列表。

运行客户端

- 运行安装路径下bin目录中的 `.bat` 文件即可运行客户端，返回信息如下所示。

```
[admin: ~]$odpscmd
Aliyun ODPS Command Line Tool
Version 1.0
@Copyright 2012 Alibaba Cloud Computing Co., Ltd. All rights reserved.
```

在光标位置输入命令（以分号作为语句的结束标志），回车即可运行。

```
odps@ odps> INSERT OVERWRITE TABLE DUAL SELECT * FROM DUAL;
```

- 在Windows中使用CMD命令行运行客户端。

在CMD命令行中，进入到bin目录下，执行 `odpscmd` ，返回如下表示运行成功。更多命令请参见启动参数。

```
D:\maxcompute\bin>odpscmd
Aliyun ODPS Command Line Tool
Version 0.30.2
@Copyright 2018 Alibaba Cloud Computing Co., Ltd. All rights reserved.
```

获取帮助

- 使用Windows的CMD命令行执行命令。

在CMD命令行中，进入到bin目录下，执行如下命令。

```
odpscmd -h
```

- 使用MaxCompute客户端执行命令。

- 在客户端执行如下命令查看全部帮助信息。

```
help;
--等价于如下命令。
h;
```

- 在客户端执行如下命令查看与关键字有关的命令提示。

```
help [keyword];
```

例如，获取与表操作相关的命令提示，如下所示。

```
odps@ odps> help table;
Usage: alter table  merge smallfiles
Usage: show tables [in ]
    list|ls tables [-p,-project ]
Usage: describe|desc [.] [partition()]
Usage: read [.] [([,..])] [PARTITION ()] [line_num]
```

启动参数

在Windows下以CMD命令行启动时，您可指定一系列参数，如下所示。

```
Usage: odpscmd [OPTION]...
where options include:
  --help (-h)for help
  --project= use project
  --endpoint= set endpoint
  -u  -p  user name and password
  -k  will skip begining queries and start from specified position
  -r  set retry times
  -f <"file_path;"> execute command in file
  -e <"command;[command;]..."> execute command, include sql command
  -C will display job counters
```

启动参数说明如下。

参数	说明
-help (-h)	获取客户端帮助信息。
--project= use project	指定进入的项目空间名称。
--endpoint= set endpoint	指定使用的Endpoint信息，详情请参见配置Endpoint。
-u	指定使用的用户名称。

参数	说明
-p	指定用户的密码。
-k	表示忽略前面的语句，从指定位置的语句开始运行。 说明 • 当 skip<=0 时，从第一条语句开始执行。 • 每个以分号分隔的语句被视为一条有效语句。 • 在运行时会打印出当前运行成功或者失败的是第几条语句。
-r	设置重试次数。
-f	指定读取文件。
-e	指定执行的命令。
-C	显示作业计数器。

示例

- 指定 -f 参数读取本地文件。
  i. 准备本地脚本文件*script.txt*，假设存放在D盘，文件内容如下所示。

```
DROP TABLE IF EXISTS test_table_mj;
CREATE TABLE test_table_mj (id string, name string);
DROP TABLE test_table_mj;
```

  ii. 打开您的CMD命令行工具，进入客户端所在路径，运行以下命令。

```
odpscmd\bin>odpscmd -f D:/script.txt;
```

- 指定 -k 参数读取指定位置的参数。
  i. 假设文件*/tmp/dual.sql*中有三条SQL语句，如下所示。

```
drop table dual;
create table dual (dummy string);
insert overwrite table dual select count(*) from dual;
```

  ii. 执行如下语句忽略前两条语句，直接从第三条语句开始执行。

```
odpscmd -k 3 -f dual.sql
```

获取当前登录用户

执行如下命令即可获取当前登录用户的云账号、使用的Endpoint配置和项目名。

```
whoami;
```

示例

```
odps@ hiveut>whoami;
Name: odpstest@aliyun.com
End_Point: http://service.odps.aliyun.com/api
Project: lijunsecuritytest
```

退出

执行如下命令退出客户端。

```
odps@ > quit;
--等价于下面的命令。
odps@ > q;
```

3.MMA2.0迁移工具

3.1. MMA2.0迁移概述

MaxCompute Migration Assist（MMA）是一款MaxCompute数据迁移工具。本文为您介绍MMA2.0的迁移方案、技术原理以及功能改进点。

MMA2.0

迁移解决方案

- 方法一：Hive直接迁移到MaxCompute。

- 方法二：Hive先迁移至OSS，再迁移至MaxCompute。

MMA2.0技术架构和原理

上图中流程说明如下：

1. 安装UDTF。
2. 启动MMA-Server。MMA-Server向Task Scheduler提交任务，Task Scheduler调用Task Runner执行任务。
3. 启动MMA-Client。MMA-Client向MMA-Server提交迁移作业。
4. 通过ODPS SDK在MaxCompute上建表以及表分区。
5. 数据校验后，通过Hive JDBC提交数据迁移作业。

MMA2.0重构改进

MMA2.0与MMA1.0相比，改进点如下：

- C/S架构设计。
- Python编程改Java。
- 完整的断点续传能力。
- 新增自动重试功能。
- 基于JDBC提交Hive作业，替代Hive客户端。
- 基于ODPS SDK提交MaxCompute作业，替代MaxCompute客户端。
- UDTF持久化上传到HDFS。

3.2. MMA2.0运行准备工作

本文为您介绍MMA2.0运行前的环境准备和迁移数据预处理。

准备运行环境

- 下载与Hive版本对应的MMA工具。下载方式请提工单获取。
- MMA所服务器上需要安装JDK1.8及以上版本的Java。

- 安装Beeline客户端。
- 确认MaxCompute所在地域并获取该地域的Endpoint，详情请参见配置Endpoint。
- 获取Hive Metastore URI。

  > 说明 在*hive-site.xml*中查找 "hive.metastore.uris" 即可获取Hive Metastore URI。

- 获取Hive JDBC连接信息。Hive JDBC的格式为 jdbc:hive2://localhost:10000/default 。
- 确保Hive集群和MMA所在机器与MaxCompute服务所在地域保持网络连通。

  > 说明 专线场景路由配置说明
  >
  > 例如，本地IDC通过专线访问MaxCompute的Endpoint，需要在边界路由器（VBR）中将100.64.0.0/10网段的路由条目指向VPC方向的路由器接口，并在本地数据中心的网关设备上将100.64.0.0/10网段的路由指向VBR的阿里云侧互联IP，详情请参见本地IDC通过专线访问云服务器ECS。

- 确认Hive Metastore是否有安全配置。

  请准备一个有权限访问Hive Metastore服务和执行Hive SQL的用户在*hive-site.xml*中查看 "hive.security.authorization.enabled" 的值。如果值为True，需要配置安全信息，详情请参见基于Kerberos身份认证的配置。

预处理待迁移数据

您可以通过如下方法对待迁移数据进行预处理，可以提升迁移效率、提升数据进入MaxCompute后的查询效率以及提前发现并解决MaxCompute与Hive的不兼容问题。

- 分区合并

  尽可能减少分区数，可以加速迁移。例如，7 TB非分区表迁移用时15分钟，而30 GB、3万分区表迁移用时为1小时。

- 类型转换
  - MaxCompute在数据类型上与Hive存在不完全兼容的情况，例如，STRING类型不支持超8 MB。
  - MMA会自动进行类型转换。例如，Hive上的DATE类型分区字段，在MaxCompute中会自动转换成STRING类型分区字段。

  > 说明 MMA默认会打开新类型，即 set odps.sql.type.system.odps2=true; ，以2.0新类型来创建表，详情请参见数据类型版本说明。

- 使用闪电立方从HDFS上传数据到OSS时，存储路径格式为*oss://bucket_name/database_name/table_name/partition_name/* 。

  > 说明 MMA2.0默认以2.0新数据类型创建表（即 set odps.sql.type.system.odps2=true; ）。详情请参见2.0数据类型版本。

基于Kerberos身份认证的配置

在*mma_server_config.json*和*mma_client_config.json*文件中添加krbPrincipal、keyTab、krbSystemProperties配置信息，如下所示。

```
{
 "dataSource": "Hive",
 "hiveConfig": {
  "jdbcConnectionUrl": "jdbc:hive2://127.0.0.1:10000/default",
  "user": "Hive",
  "password": "",
  "hmsThriftAddr": "thrift://127.0.0.1:9083",
  "krbPrincipal": "xxx",
  "keyTab": "xxx",
  "krbSystemProperties": "xxx=xxx,xxx=xxx",
  "hiveJdbcExtraSettings": [
   "hive.fetch.task.conversion=none",
   "hive.execution.engine=mr",
   "mapreduce.job.name=data-carrier",
   "mapreduce.max.split.size=512000000",
   "mapreduce.task.timeout=3600000",
   "mapreduce.map.maxattempts=0",
   "mapred.map.tasks.speculative.execution=false"
  ]
 },
 "odpsConfig": {
  "accessId": "your_access_id",
  "accessKey": "your_access_key",
  "endpoint": "your_endpoint",
  "projectName": "your_project_name"
 }
}
```

3.3. MMA2.0安装和配置

本文为您介绍如何安装MMA2.0、准备配置文件和添加函数。

解压工具包

执行如下命令解压工具包。工具包请提工单获取。

```
tar vxzf odps-data-carrier.tar.gz
```

解压后的MMA2.0目录结构如下。

准备配置文件

- *odp_config.ini*

```
project_name= /* MaxCompute的项目名称 */
access_id=/* 阿里云账号的AccessKey ID */
access_key=/* 阿里云账号的AccessKey Secret */
end_point=/* MaxCompute服务所在地域的Endpoint */
```

- *hive_config.ini*

```
jdbc_connection_url=/* Hive JDBC连接串 */
user=/* Hive JDBC用户名 */
password=/* Hive JDBC密码 */
hms_thrift_addr=/* Hive Metastore Service的Thrift地址 */
```

- *table_mapping.txt*

  该配置文件呈现待迁移Hive表与MaxCompute表的对应关系，文件中每一行对应一个Hive表到MaxCompute表的迁移任务。格式如下。

```
<hive db>.<hive table>:<maxcompute project>.<maxcompute table>
```

- MMA Server配置文件

  执行如下命令生成MMA Server配置文件。

```
cd odps-data-carrier/conf/
sh ../bin/generate-config ./
--to_server_config ./
--hive_config hive_config.ini
--odps_config odps_config.ini
```

- MMA Client配置文件

  执行如下命令生成MMA Client配置文件。

```
cd odps-data-carrier/conf/
sh ../bin/generate-config ./
--to_migration_config ./
--hive_config hive_config.ini
```

- MMA迁移任务配置文件

  MMA2.0支持如下四种迁移作业模式，您可以根据迁移场景选择合适的迁移模式。

- 全服务迁移（All Databases To One Project）

  配置文件模板如下。

```
{
 "user": "Jon",
 "globalAdditionalTableConfig": {
  "partitionGroupSize": 100,
  "retryTimesLimit": 3
 },
 "serviceMigrationConfig":
  {
   "destProjectName": "test_project"
  }
}
```

  destProjectName表示目标MaxCompute项目的名称。

- 整库迁移（One Database To One Project）

  配置文件模板如下。

```
{
 "user": "Jerry",
 "globalAdditionalTableConfig": {
  "partitionGroupSize": 100,
  "retryTimesLimit": 3
 },
 "databaseMigrationConfigs": [
  {
   "sourceDatabaseName": "test_db",
   "destProjectName": "test_project"
  }
 ]
}
```

- 表级迁移（DB.table To Project.table）

  您可以基于*table_mapping.txt*文件，使用 `generate-config` 命令生成配置文件。命令格式如下。

```
cd odps-data-carrier/conf/
sh ../bin/generate-config ./
--to_migration_config ./
--table_mapping table_mapping.txt
```

- 分区级迁移（DB.table.partition To Project.table.partition）

  配置文件模板如下。

```
{
 "user": "Jerry",
 "globalAdditionalTableConfig": {
  "partitionGroupSize": 100,
  "retryTimesLimit": 3
 },
 "tableMigrationConfigs": [
  {
   "sourceDataBaseName": "test_db",
   "sourceTableName": "test_partitioned_table",
   "destProjectName": "test_project",
   "destTableName": "test_partitioned_table",
   "partitionValuesList": [
    ["foo", "123456"],
    ["bar", "234567"],
   ]
   "additionalTableConfig": {
    "partitionGroupSize": 100,
    "retryTimesLimit": 3
   }
  }
 ]
}
```

  partitionValuesList表示待迁移分区的分区值列表。假设表的分区列为 `c1 STRING` 和 `c2 BIGINT` ，示例模板中partitionValuesList表示的分区即为 `c1="foo",c2=123456` `c1="bar",c2=234567` 。

添加函数

上传配置文件和资源包，并通过Hive客户端Beeline添加函数。

1. 执行如下命令上传*odps_config.ini*配置文件。

```
cd odps-data-carrier/conf/
hdfs dfs -put -f odps_config.ini hdfs:///tmp/odps_config.ini
```

2. 执行如下命令上传UDTF的JAR包。JAR包为工具包*lib*目录下的*data-transfer-hive-udtf-1.0-SNAPSHOT-jar-with-dependencies.jar*。

```
cd odps-data-carrier/lib/
hdfs dfs -put -f data-transfer-hive-udtf-1.0-SNAPSHOT-jar-with-dependencies.jar
hdfs:///tmp/data-transfer-hive-udtf-1.0-SNAPSHOT-jar-with-dependencies.jar
```

3. 启动Beeline，并添加函数。

3.4. MMA2.0数据迁移命令说明

本文为您提供MMA2.0的迁移命令及处理迁移失败作业的方法。

启动MMA Server

- 命令格式

```
cd odps-data-carrier/bin
nohup sh ./mma-server --config ../conf/mma_server_config.json > mma-server.out 2>&1 &
```

- 命令说明

  MMA Server进程会一直保持运行状态。如果由于各种原因导致MMA Server运行中断，您可以直接执行以上命令重启MMA Server。一台设备只能存在一个MMA Server进程。

- 示例

  

  查看MMA Server控制台输出的 `mma-server.out` ，MMA Server已经开始运行，状态为等待迁移作业。

  

通过MMA Client提交迁移作业

- 命令格式

```
cd odps-data-carrier/bin
sh ./mma-client --config ../conf/mma_client_config.json --start ../conf/mma_migration_config.json
```

- 命令说明

  启动MMA Client，成功提交迁移作业后，MMA Client会自动退出。查看MMA Server控制台的输出结果，MMA Server已经接收并开始执行迁移作业。

- 示例

  

查看迁移作业进展

- 命令格式

```
cd odps-data-carrier/bin/
sh ./mma-client --config ../conf/mma_client_config.json --wait_all
```

- 示例

  

查看不同状态的迁移作业

- 命令格式

```
cd odps-data-carier/bin/
sh ./mma-client --config ../conf/mma_client_config.json --list succeeded/failed/running
```

- 命令说明

  迁移作业状态包含成功、失败和运行中。

- 示例

删除迁移作业

- 命令格式

```
cd odps-data-carier/bin/
sh ./odps-data-carrier/bin/mma-client --config configs/mma_client_config.json --remove db_name.table_name
```

- 示例

处理迁移失败的作业

迁移作业失败的主要原因如下：

- 集群负载过高。
- HDFS DataNode异常、Meta异常或配置有误。

MMA 2.0版本默认开启表粒度的自动重试功能。您可以通过设置重试次数，提高迁移成功率。建议您及时查看迁移作业的状态，对于自动重试仍然失败的作业，请根据日志报错信息进行处理，处理完成后重新提交迁移作业。常见处理方法如下：

- 手动重跑失败的迁移作业

  由于异常问题导致作业重跑失败时，解决异常问题后，您需要手动重跑失败作业。在重新提交迁移作业时，同时执行如下命令。

```
cd odps-data-carier/bin/
sh ./mma-client --config ../conf/mma_client_config.json --start ../conf/mma_migration_config.json
```

- 自定义重试次数

  MMA重试次数默认为3次。如果3次重试仍然失败，您可以修改参数retryTimesLimit的值提高重试次数。MMA支持自定义全局和表级别的重试次数。

- 调节分区表的分组大小

  默认100个分区为一组，您可以根据集群资源、表的分区数和表的数据量调节分组数量，以保证迁移成功率。

3.5. 其他类型作业迁移说明

本文为您介绍Hive之外其他类型作业迁移时的注意事项。

UDF、MapReduce迁移 外表迁移 Spark作业迁移

UDF和MapReduce迁移

- 支持相同逻辑的UDF和MapReduce输入、输出参数的映射转换，但UDF和MapReduce内部逻辑需要您自行维护。
- 不支持在UDF、MapReduce中直接访问文件系统、网络访问、外部数据源连接。
- Hive UDF兼容示例，请参见Hive UDF兼容示例。

外表迁移

- 原则上数据会全部迁到MaxCompute内部表。
- 如果必须通过外表访问外部文件，建议先将文件迁移到OSS，然后在MaxCompute中创建外部表，实现对文件的访问。
- MaxCompute外部表支持的格式包括ORC、PARQUET、SEQUENCEFILE、RCFILE、AVRO和TEXTFILE。

Spark作业迁移

- 如果作业无需访问MaxCompute表和OSS，可直接运行Jar包，请参见《MaxCompute Spark开发指南》准备开发环境和修改配置。

  > **说明** 对于Spark或Hadoop的依赖必须设成provided。

- 如果作业需要访问MaxCompute表，请参见《MaxCompute Spark开发指南》中访问MaxCompute表所需依赖编译Datasource并安装到本地Maven仓库，在*pom.xml*中添加依赖后重新打包即可。
- 如果作业需要访问OSS，请参见《MaxCompute Spark开发指南》中OSS依赖，在*pom.xml*中添加依赖后重新打包即可。

4.MaxCompute Studio

4.1. 认识MaxCompute Studio

MaxCompute IDEA

MaxCompute Studio是阿里云MaxCompute平台提供的安装在开发者客户端的大数据集成开发环境工具，是一套基于流行的集成开发平台IntelliJ IDEA的开发插件，帮助您便捷、快速地进行数据开发。本文将为您介绍MaxCompute Studio的功能界面和常用的应用场景。

基本用户界面

MaxCompute Studio是IntelliJ IDEA平台上的一套插件，共享了IntelliJ IDEA的基本开发界面。

MaxCompute Studio在IntelliJ的基础上提供以下功能：

- **SQL编辑器（SQL Editor）**：提供SQL语法高亮、代码补全、实时错误提示、本地编译、作业提交等功能。

  编译器视图（Compiler View）：显示本地编译的提示信息和错误信息，在编辑器中定位代码。
- **项目空间浏览器（Project Explorer)**：连接MaxCompute项目空间，浏览项目空间表结构、自定义函数、资源文件。

  表详情视图（Table Details View）：提供表、视图等资源的详情显示和示例数据（Sample Data）。
- **作业浏览器（Job Explorer）**：浏览、搜索MaxCompute的历史作业信息。
  - 作业详情视图（Job Details View）：显示作业的运行详细信息，包括执行计划和每个执行任务的详细信息，Logview工具能够显示的全部信息。
  - 作业输出视图（Job Output View）：显示正在运行的作业的输出信息。
  - 作业结果视图（Job Result View）：显示SELECT作业的输出结果。
- **MaxCompute控制台（MaxCompute Console）**：集成了MaxCompute客户端，可以输入和执行MaxCompute客户端命令。

连接MaxCompute项目空间

使用Studio的大部分功能前需要您先创建项目空间连接。建立项目空间连接后，即可在**项目空间浏览器**中查看相关的数据结构和资源信息。Studio会自动为每一个项目空间连接建立一个本地的元数据备份，以提高对MaxCompute元数据的访问频率和降低延时。

创建MaxCompute项目空间连接后，才可以通过Studio进行编辑SQL脚本、提交作业、查看Job信息、打开MaxCompute控制台等操作。MaxCompute项目空间的更多详情请参见项目空间。在Studio中管理项目空间的更多详情请参见项目空间连接。

管理数据

您可以通过Studio的**项目空间浏览器**快速浏览项目空间的表结构、自定义函数、资源文件。通过树形控件，可以列出所有项目空间连接下的数据表、列、分区列、虚拟视图、自定义函数名称、函数签名、资源文件及类型等，并支持快速定位。

双击某个数据表，即可打开**表详情视图**，查看数据表的元信息、表结构和示例数据。如果您没有项目空间的相应权限，Studio会提示对应的错误信息。

Studio集成了MaxCompute Tunnel工具，可以支持本地数据的上传和下载，更多详情请参见导入并导出数据。

编写SQL脚本

您可以在Studio中编写MaxCompute SQL脚本，非常方便。

1. 打开Studio，导航至File > New > Project或者File > New > Module...。
2. 创建一个MaxCompute Studio类型的项目或者模块。
3. 导航至File > New > MaxCompute Script 或者右击菜单New > MaxCompute Script，即可创建一个MaxCompute SQL脚本文件。

> **说明** 创建MaxCompute SQL脚本时，Studio会提示您选择一个关联的MaxCompute项目空间，您也可以通过SQL编辑器上的工具条最右侧的**项目空间选取器**进行更改，编辑器会根据SQL脚本关联的项目空间对SQL语句自动进行元数据（比如表结构等）的检查并汇报错误，提交运行时也会发送到关联的项目空间执行。更多详情请参见编写SQL脚本。

SQL代码智能提示

Studio提供的SQL编辑器可以根据您写入的代码，智能提示SQL语句的语法错误、类型匹配错误或者警告等，实时地标注在代码上。

通过代码补全功能，Studio可以根据代码上下文，提示您项目空间名称、表、字段、函数、类型、代码关键词等，并根据您的选择，自动补全代码。

编译和提交作业

- 编译作业

  单击SQL编辑器工具条上的

  图标，可以对SQL脚本执行本地编译，如果有语法或者语义错误，编译器窗口会报告错误。

- 提交作业

  单击SQL编辑器工具条上的

  图标，会在本地编译之后，把SQL脚本提交到MaxCompute指定的项目空间排队执行。

查看历史作业

打开**作业浏览器**，您即可查看指定项目空间上近期执行的作业。

> **说明** 这个列表只能显示以当前连接使用的用户ID提交的作业。

双击其中一个作业，便可查看作业的详情信息。

如果知道一个任务的Logview URL，可以导航至MaxCompute > Open Logview，打开该任务的详情页面。

开发MapReduce和UDF

Studio还支持MapReduce和Java UDF的开发。

连接MaxCompute客户端

Studio集成了最新版本的MaxCompute客户端，您也可以在Studio的配置页面中指定本地已经安装好的MaxCompute客户端路径。

您在项目空间浏览器中选定一个项目空间，右键单击菜单选择Open in Console即可打开MaxCompute控制台窗口。

后续步骤

现在，您已经学习了MaxCompute Studio的功能界面和常用的应用场景，您可以继续学习下一个教程。在该教程中您将学习如何安装MaxCompute Studio。详情请参见工具安装与版本信息。

4.2. 安装MaxCompute Studio

4.2.1. 安装IntelliJ IDEA

MaxCompute Studio是运行在IntelliJ IDEA上的插件，本文将为您介绍如何安装MaxCompute Studio的基础平台IntelliJ IDEA。

安装IntelliJ IDEA

前提条件

IntelliJ IDEA支持在Windows、macOS或者Linux操作系统上安装，硬件及系统环境要求请参见 Requirements for IntelliJ IDEA 。

操作步骤

1. 下载 IntelliJ IDEA。请根据操作系统（Windows、macOS、Linux）下载对应IntelliJ IDEA版本。本文以Windows操作系统为例，支持IntelliJ IDEA 14.1.4以上版本 。

   > 说明 IntelliJ IDEA支持PyCharm、Ultimate版本或免费的Community版本。

2. 下载完成后，双击安装程序，进入安装界面，单击Next。
3. 指定安装目录后，单击Next。
4. 选择相应的系统类型，单击Next 。如下图所示。
5. 单击Install，开始安装。
6. 安装完成后，单击Finish 。

后续步骤

安装MaxCompute Studio插件，详情请参见安装步骤。

4.2.2. 安装步骤

环境要求

IntelliJ IDEA 支持在 *Windows*， *Mac*， *Linux* 操作系统上安装，硬件及系统环境要求请参见 https://www.jetbrains.com/help/idea/2016.3/requirements-for-intellij-idea.html 。基于 IntelliJ IDEA 平台的 MaxCompute Studio 也可以安装在这些操作系统的客户端上。

MaxCompute Studio 对用户环境有以下要求：

- Windows，Mac OS，或者 Linux 系统客户端。
- 安装 IntelliJ IDEA 14.1.4 以上版本 （支持 Ultimate 版本、PyCharm 版本和免费的 Community 版本）。
- 已安装 JRE 1.8 （最新的 IntelliJ IDEA 版本捆绑了 JRE 1.8）。
- 已安装 JDK 1.8 （*可选*：如果需要开发和调试 Java UDF，则需安装 JDK）。

安装方式

MaxCompute Studio 是 IntelliJ IDEA 的插件，有以下两种安装方式：

- 通过插件库在线安装（推荐）。
- 通过本地文件安装。

在线安装（推荐）

MaxCompute Studio 插件已对全部公网用户开放，您可以通过 IntelliJ 官方插件库安装。

操作步骤

1. 在 IntelliJ IDEA 中打开插件配置页面 （Windows/Linux 用户导航至 **File > Settings > Plugins**， Mac 用户导航至**IntelliJ IDEA > Preferences > Plugins** ）。
2. 单击**Browse repositories...** 按钮，然后搜索 `MaxCompute Studio` 。
3. 找到 MaxCompute Studio 插件页面，单击绿色**Install** 按钮进行安装。
4. 确认安装后，重新启动 IntelliJ IDEA，完成安装。

本地安装

MaxCompute Studio 也可以在本地环境中进行安装。

操作步骤

1. 进入 MaxCompute Studio 插件页面 下载插件包。
2. 运行 IntelliJ IDEA。
   - 如果是第一次，会出现欢迎界面，单击欢迎界面中的**configure** （配置），选中弹出菜单中的 **Plugins**（插件），如下图所示：
   - 如果不是第一次运行，可以依次单击菜单 **File > Settings > Plugins** 进入相同的界面，如下图所示：
3. 在插件页面，单击 **Install plugin from disk...**（从本地磁盘安装插件），如下图所示：
4. 在弹出窗口中，通过单击目录名称前的灰色图标进行导航，找到插件文件并选中，单击 **OK**。
5. 回到插件首页后，单击**OK** ，开始安装本地插件。
6. 安装完成后，弹出重新启动的提示窗口，单击**Restart** ，重新启动 IntelliJ IDEA。

7. 重新启动后，界面如下所示：

后续步骤

现在，您已经学习了如何安装 MaxCompute Studio 插件，您可以继续学习下一个教程。在该教程中您将学习如何配置 MaxCompute Project 连接管理数据和资源。详情请参见 新建 MaxCompute 项目空间连接。

4.2.3. 查看和更新MaxCompute Studio版本

本文为您介绍如何查看和更新MaxCompute Studio版本。

查看MaxCompute Studio版本信息

查看MaxCompute Studio版本信息

1. 在顶部菜单栏，单击**File > Settings...** 。
2. 在**Settings**页面的左侧导航栏，单击**MaxCompute Studio**。
3. 在**MaxCompute Studio**页面的**Updates Checking**区域可以看到当前MaxCompute Studio的版本号，以及最近发布的版本信息。

检查新版本

- 默认情况下，MaxCompute Studio会自动检测新版本，当有新的可用版本时，会自动通知您。

  收到更新提示后，您可以选择：
  - 安装： 在更新提示中，单击**安装**，将会自动下载并安装此新版本，安装完成后重启IntelliJ IDEA即可。
  - 配置： 在更新提示中，单击**配置**，您可以配置是否自动检查新版本。
- 如果关闭了自动更新功能，您可以通过以下步骤检查MaxCompute Studio的版本并选择安装：
  i. 在顶部菜单栏，单击**File > Settings...** 。
  ii. 在**Settings**页面的左侧导航栏，单击**MaxCompute Studio**。
  iii. 在**MaxCompute Studio**页面的**Updates Checking**区域可以看到当前MaxCompute Studio的版本号，以及最近发布的版本信息。
  iv. 单击**Check new versions**检测最新可用版本并安装。

  > **说明** 您可以选中**Automatically checks for new versions**打开自动检测新版本功能。

4.3. 配置MaxCompute Studio

本文为您介绍如何配置MaxCompute Studio以及各项配置项信息。

配置 MaxCompute Studio
安装MaxCompute Studio插件后，在顶部菜单栏，单击**File > Settings...**，即可进入**MaxCompute Studio**配置页面。

MaxCompute Studio配置选项页

在**Settings**页面左侧导航栏上，单击**MaxCompute Studio**。**MaxCompute Studio**配置选项页提供以下配置项：

- 基本参数

- **Local meta store location**：本地元数据仓库存储路径。指定本地存储MaxCompute项目空间元数据的路径。MaxCompute Studio的默认设置是本地用户目录下的*.odps.studio\meta*目录。
- **Table preview rows**：表的最大预览行数。
- **本地作业保存目录**：MaxCompute Studio作业的本地保存路径。
- **Python path to resolve UDF**：Pyhton的本地安装路径。

- 版本更新选项
  - **Automatically checks for new version**：控制MaxCompute Studio是否自动检查更新版本。默认情况下是选中状态，支持自动更新。
  - **Check new versions**：用于手动检查新版本。 单击此按钮后，如果有新版本可以更新，将显示**Install new version** 按钮。单击**Install new version**安装最新版本软件，安装完成后需要重启IntelliJ IDEA。

SDK & Console配置选项页

在**Settings**页面左侧导航栏上，单击**MaxCompute Studio > SDK & Console**，进入SDK & Console配置选项页。

SDK & Console配置选项页面提供了**Installed Location**配置项，用以指定本地安装MaxCompute客户端的安装路径。MaxCompute Studio会自动检测路径中安装的MaxCompute客户端的版本，如果检测失败，会返回错误信息。

> **说明** MaxCompute Studio 2.6.1之后的版本自带了最新的MaxCompute客户端，不需要您特别指定。 如果您希望使用自己特定版本的MaxCompute客户端，可以在此处指定路径。

MaxCompute SQL配置选项页

在**Settings**页面左侧导航栏上，单击**MaxCompute Studio > MaxCompute SQL** ，进入MaxCompute SQL配置选项页。

MaxCompute SQL配置选项页面提供以下配置项：

- 语法高亮

  选中**Enable syntax coloring** ，启动语法高亮功能。

- 代码自动补全
  - 选中**Enable code completion** ，启动代码自动补全功能。
  - 选中**Invoke code completion when you enter a space**，启动输入空格时自动结束代码补全功能。

- 代码格式化

  选中**Enable code formatting** ，启动代码格式化功能。

- 脚本提交选项
  - 选中**Show job detail when script submitted**，脚本提交时会显示作业详情。
  - 选中**Record sql history when script submitted**，脚本提交时记录SQL历史记录。
  - 选中**Show sql cost confirm dialog when script submitted**，提交脚本时显示SQL成本确认对话框。
  - 选中**Pin sql execution result tab by default**，默认情况下锁定SQL执行结果选项卡。

- 作业名称Job name：
  - use script file name as default：默认使用脚本的名称作为工作名称。
  - input job name when first submit：第一次提交时输入作业名称。
  - input job name for every submit：每一次提交时输入作业名称。

- 编译器选项

  此处设置的选项为全局默认的编译器选项。 以下选项还可以在SQL编辑器的工具栏中为每个文件单独设置。

  - 编译器模式（Compiler Mode）
    - 单句模式（Statement Mode）：在该模式下，编译器将SQL文件中的单条语句作为单元进行编译、提交。
    - 脚本模式（Script Mode）：在该模式下，编译器将整个SQL文件作为单元进行编译、提交。 脚本模式有利于编译器和优化器最大程度地优化执行计划，提高整体执行效率。
  - 类型系统
    - 旧有类型系统（Legacy TypeSystem）：原有MaxCompute的类型系统。
    - MaxCompute 类型系统（MaxCompute TypeSystem）：MaxCompute 2.0引入的新类型系统。
    - Hive 类型系统（Hive Compatible TypeSystem）：MaxCompute 2.0引入的Hive兼容模式下的类型系统。
  - 编译器版本
    - 默认编译器（Default Version）：默认版本的编译器。
    - 实验性编译器（Flighting Version）：实验性的版本的编译器，包含正在测试中的编译器的新特性。

Accounts配置选项页

在Settings页面左侧导航栏上，单击MaxCompute Studio > Accounts，进入Accounts配置选项页。

Accounts配置选项页面用于管理访问MaxCompute的所用账户，关于账户更多信息请参见 用户认证。

> 说明 MaxCompute Studio需要通过用户指定的账号访问MaxCompute的项目空间和执行提交作业等操作，目前MaxCompute Studio支持的账号类型为阿里云账号（AccessKey）。

- 添加账户
  i. 在右侧导航栏上，单击+ > Aliyun Account By Accesskey。
  ii. 在Add MaxCompute Account页面配置参数。
    - Account Name： 该账户在MaxCompute Studio中的标识名称。
    - Using properties file：从配置文件中读取AccessKey ID和AccessKey Secret。如果您选择了此种方式，需要上传在用户认证中的配置文件*conf/odps_config.ini*。
    - Using properties：手动填入AccessKey ID和AccessKey Secret。此选项与Using properties file选项二选一即可。
      - Access Id：填入阿里云账号的AccessKey ID。
      - Access Key：填入阿里云账号的AccessKey Secret。

  iii. 单击OK。添加完成后账号会出现在Accounts配置选项页面的列表中。
- 删除账户

该操作仅在MaxCompute Studio配置中删除账户配置，对您账户本身不产生影响：

i. 在Account列表中选择要删除的账户名称。

ii. 在右侧导航栏上，单击-。

iii. 在弹出的确认对话框中，选择OK完成删除 。

- 修改账户信息

i. 在Account列表中选择要修改的账户名称。

ii. 在右侧导航栏上，单击 。

iii. 在弹出的Edit MaxCompute Account窗口中编辑Account配置信息如下：

- Access Id：填入阿里云账号的AccessKey ID。
- Access Key：填入阿里云账号的AccessKey Secret。

iv. 单击OK完成修改。

4.4. 管理项目连接

您必须通过MaxCompute Studio连接MaxCompute项目后，才可以在MaxCompute Studio上查看MaxCompute项目的信息，例如表、视图、自定义函数（UDF）或资源。本文为您介绍如何创建或修改MaxCompute项目连接。

连接MaxCompute项目

步骤一：创建MaxCompute Studio项目

1. 启动IntelliJ IDEA，在顶部菜单栏，单击File > New > Project。
2. 在New Project页面的左侧导航栏，选择MaxCompute Studio，单击Next。

3. 填写Project name，单击Finish，完成项目创建。

> **说明**
>
> 如果有已经打开的Project，将会提示您是否在当前窗口中打开，即关掉之前的Project，选择This Window。

步骤二：创建MaxCompute项目连接

1. 在顶部菜单栏，单击View > Tool Windows > Project Explorer。

2. 单击左上角的+，选择Add project from accessId/Key。

3. 在Add MaxCompute project对话框，配置Connection页签信息。

> 说明
> - 单击对话框左下角的 即可查看在线文档。
> - 如果出现超时错误，单击对话框中的Setting页签，修改数据同步相关参数：
>   - sync one table timeout(s)：同步表超时参数。默认为5s。
>   - sync one function timeout(s)：同步函数超时参数。默认为30s。

- 通过配置文件自动配置参数。

参数	说明
Properties File	上传MaxCompute项目客户端（odpscmd）的配置文件*odps_config.ini*，详情请参见安装并配置客户端。用于初始化Access Id、Access Key、Project Name和End Point配置项。
AK Account	单击右侧+，在Accounts页面，选择已存在的账号。 说明 如果没有账号信息，您需要在Accounts页面，单击+ > Aliyun Account By AccessKey，通过配置文件自动识别或手动方式添加账号信息。

- 手动配置参数。

参数	说明
Access Id	连接MaxCompute项目时的AccessKey ID。
Access Key	连接MaxCompute项目时的AccessKey Secret。
Project Name	MaxCompute项目的名称。
End Point	MaxCompute项目的Endpoint。

4. 配置完成后，单击OK。在左侧Project Explorer区域中会显示MaxCompute项目的信息，包括该项目中的表、视图、函数以及资源。

步骤三：修改MaxCompute项目连接

1. 在Project Explorer区域中，右键单击需要修改的MaxCompupte项目，选择Modify project properties。

   

2. 在Modify MaxCompute project对话框，修改MaxCompute项目的配置。

后续步骤

连接MaxCompute项目后，您可以管理和查看项目内的数据和资源，详情请参见管理数据和资源。

4.5. 管理数据和资源

4.5.1. 管理项目数据

本文为您介绍如何在MaxCompute Studio上查看项目空间中的表、视图、函数和资源。

前提条件

已连接MaxCompute项目，详情请参见管理项目连接。

背景信息

您可以在Project Explorer区域查看已添加连接的MaxCompute项目中的表、视图、函数和资源。

浏览和更新项目数据

1. 进入MaxCompute Studio页面，在左侧导航栏，单击Project Explorer，即可浏览已连接的MaxCompute项目。

2. 单击Tables & Views、Functions或Resources节点前的下拉箭头，可以查看该项目下的所有表、视图、函数和资源。

> 说明 此操作需要您拥有该MaxCompute项目的 `DESC TABLE` 权限。

3. 在工具栏中，单击 更新本地元数据。

> 说明 MaxCompute Studio会将MaxCompute项目的元数据下载到本地，当MaxCompute项目的元数据有更新时，需要手动触发一次刷新，将变化的元数据重新加载到本地。

查看表或视图详细信息

1. 在Tables & Views节点树中，单击表名展开节点，可以快速查看表的字段和字段类型。

2. 在表名上单击右键，选择Show table detail，可以查看表的详细信息。例如，表的所属者、表的大小、表的列信息和Schema。

3. 在表名上单击右键，选择Find usages，可以查询表被脚本引用的情况。

查看函数详细信息

1. 在Functions节点树中，单击UserDefined节点，可以查看您创建的函数。

> 说明 Functions节点树的BuiltIn节点下分类显示系统的内建函数，双击即可显示该函数的使用说明。内建函数详情请参见内建函数。

2. 双击创建的函数，即可打开函数对应的代码。

查看资源详细信息

您可以在Resources节点树中，查看创建的资源。

4.5.2. 导入导出表数据

MaxCompute Studio可以将CSV、TSV等格式的本地数据文件导入至MaxCompute表中，也可将MaxCompute表中的数据导出到本地文件。MaxCompute Studio通过Tunnel导入导出数据。

前提条件

- 导入导出数据使用MaxCompute Tunnel，因此要求MaxCompute Studio中添加的MaxCompute项目必须配置了Tunnel。详情请参见安装并配置客户端。
- 导入导出使用的账号必须具备MaxCompute项目中表的操作权限。

导入数据

1. 在Project Explorer区域，单击MaxCompute项目的Tables & Views节点前的下拉箭头，右键单击需要导入数据的表，选择Import data into table。

   

2. 在Importing data to table_name对话框中，配置导入文件参数。

   

   - Input File：导入数据文件的本地路径。
   - File charset：导入数据文件的编码格式。编码格式包含UTF-8、UTF-16、UTF-16BE、UTF-16LE、ISO-8859-1、US-ASCII和GBK。默认为UTF-8。
   - Column Separator：列分隔符。包含Comma(',')、Space(' ')和Tab('\t')。默认为Comma(',')。
   - Record Limit：导入数据的最大行数。
   - Size(MB) Limit：导入数据量最大值，单位为MB。
   - Error Record Limit：容错行数。
   - Include Column Header：是否导入列标题。
3. 单击OK，完成数据导入。
4. 提示Success，表示数据导入成功，您可以在表中查看导入的数据。

导出数据

1. 在Project Explorer区域，单击MaxCompute项目的Tables & Views节点前的下拉箭头，右键单击需要导出数据的表，选择Export data from table。

   

2. 在Exporting data from table_name对话框中，配置导出数据文件参数。

   

   - Output File：导出数据文件的本地路径。
   - File charset：导出数据文件的编码格式。编码格式包含UTF-8、UTF-16、UTF-16BE、UTF-16LE、ISO-8859-1、US-ASCII和GBK。默认为UTF-8。
   - Column Separator：列分隔符。包含Comma(',')、Space(' ')和Tab('\t')。默认为Comma(',')。
   - Record Limit：导出数据的最大行数。

- **Size(MB) Limit**：导出数据量最大值，单位为MB。
- **Error Record Limit**：容错行数。
- **Include Column Header**：是否导出列标题。

3. 提示Success，表示数据导出成功，您可以在导出文件中查看导出的数据。

4.5.3. 可视化管理表

MaxCompute Studio的Project Explorer提供了可视化表结构编辑器。本文为您介绍如何通过Project Explorer可视化创建、修改和删除表。

创建表

1. 在Project Explorer区域，右键单击MaxCompute项目下的Tables & Views，选择Create new table。
2. 在创建表/视图对话框，配置参数信息。

   

   界面参数配置原则请遵循MaxCompute相关要求，详情请参见表操作。

   > **说明**
   >
   > 可视化建表无法设置Flag，默认使用以下两个Flag：
   >
   > - `odps.sql.submit.mode=script`
   > - `odps.sql.type.system.odps2=true`

3. 单击执行，提示SUCCESS，完成创建。

修改表

1. 在Project Explorer区域，单击MaxCompute项目的Tables & Views节点前的下拉箭头，右键单击需要修改的表，选择Open table editor。
2. 在修改表对话框，对表进行编辑。您可以新增列，修改表名称、表注释、生命周期、列名称和列注释。详情请参见表操作。

   

3. 单击执行，完成表修改。

删除表

1. 在Project Explorer区域，单击MaxCompute项目的Tables & Views节点前的下拉箭头，右键单击需要删除的表，选择Drop table from server。

   

2. 在Confirmation Required对话框，单击OK，即可将表从MaxCompute项目中删除。

4.6. 开发SQL程序

4.6.1. 概述

本文为您介绍在MaxCompute Studio上开发SQL脚本的流程、SQL编辑器和编译相关参数的设置。

开发SQL脚本流程 SQL编辑器设置 编译设置
在MaxCompute Studio上开发SQL脚本流程如下：

1. 创建MaxCompute Script Module。
2. 编写SQL脚本。详情请参见开发及提交SQL脚本。
3. 将SQL脚本提交至MaxCompute服务端，运行SQL脚本。详情请参见提交SQL脚本。

编辑器设置

MaxCompute Studio不仅提供语法高亮、智能提醒、错误提示，还支持以下功能：

- **schema annotator**： 当鼠标悬停在表上时，显示其Schema。
  - 悬停在列上时，显示其类型。
  - 悬停在函数上时，显示其签名。
- **code folding**： 将子查询折叠起来，方便SQL的阅读。
- **brace matching**：鼠标单击高亮左括号，其匹配的右括号也会高亮，反之亦然。
- **go to declaration**：按住Ctrl键，单击表 ，即可查看表详情。单击函数，即可显示其源码。
- **code formatting**：支持对当前脚本格式化，可以通过快捷键（Ctrl+Alt+L）打开配置页面。可在如下页面自定义格式化规则，例如关键字大小写、是否换行等。
- **code inspect**：支持对当前脚本进行代码检查，某些检查还支持快速修复，可通过快捷键（Alt+Enter）打开。
- **find usages**：右键单击选中的某张表（或函数），选择**Find Usages** ，则会在当前IDEA项目下寻找所有使用该表（函数）的脚本。
- **live template**：MaxCompute Studio内置了部分SQL模板，可以在编辑器中使用快捷键（Ctrl+J）打开模板。
- **builtin documentation**：支持在系统内置函数处通过快捷键（Ctrl+Q）打开帮助文档。
- **Sql History**：通过MaxCompute Studio提交的运行记录都保存在本地。在工具栏上单击 图标，即可在**Sql History**窗口，查询曾经执行过的SQL。

编译设置

在SQL脚本提交前，您可以根据自己的需要设置相关编译参数。MaxCompute Studio提供了丰富的功能，可以在编辑器上方的工具栏中快速设置。

设置参数主要分为以下3种：

- 编辑器模式：
  - **单步模式**：将提交的脚本按英文分号（;）分隔，逐条提交到MaxCompute服务端执行。
  - **脚本模式**：将整个脚本一次性提交到MaxCompute服务端，由服务端提供整体优化，效率更高。推荐您使用此模式。
- 类型系统：类型系统主要解决SQL语句的数据类型兼容性问题。分为以下3种类型：
  - **旧有类型系统**：原有MaxCompute的类型系统。即MaxCompute 1.0数据类型版本。
  - **MaxCompute 类型系统**：MaxCompute 2.0引入的新的类型系统。即MaxCompute 2.0数据类型版本。

- **Hive 类型系统**：MaxCompute 2.0引入的Hive兼容模式下的类型系统。即Hive兼容数据类型版本。

- 编译器版本：
  - **默认编译器**：稳定版本。
  - **实验性编译器**：包含编译器最新特性。

4.6.2. 创建MaxCompute Script Module

使用MaxCompute Studio开发SQL程序前，需要先创建MaxCompute Script Module。本文为您介绍如何创建MaxCompute Script Module。

背景信息

创建MaxCompute Script Module时存在以下两种情况：

- 本地没有Script文件：需要通过IntelliJ IDEA创建一个全新的Module。
- 本地已有Script文件：假如本地某个文件夹下已经存在脚本，此时需要用MaxCompute Studio来编辑脚本，您可直接打开一个Module，无需全新创建。

本地没有Script文件时创建Module

1. 启动IntelliJ IDEA，在顶部菜单栏，单击**File > New > Project**。
2. 在**New Project**页面的左侧导航栏，选择**MaxCompute Studio**，单击**Next**。

   
3. 填写**Project name**，单击**Finish**，完成项目创建。

   

   > **说明**
   >
   > 如果有已经打开的Project，将会提示您是否在当前窗口中打开，即关掉之前的Project，选择**This Window**。

本地已有Script文件时创建Module

本地已有Script文件时无需新建Module，只需要在已有的Module目录下添加MaxCompute连接配置文件即可。

1. 在MaxCompute Studio的本地*.\IdeaProjects\MaxCompute_Studio_Project_Name\scripts*文件夹下新建一个MaxCompute的连接配置文件*odps_config.ini*，文件中包含MaxCompute连接的鉴权信息，示例如下。

```
# 连接的MaxCompute项目名称。
project_name=xxxxxxxx
# 云账号的AccessKey ID。
access_id=xxxxxxxxxx
# 云账号的AccessKey Secret。
access_key=xxxxxxxxx
# 连接的MaxCompute服务所在区域的Endpoint信息。
end_point=xxxxxxxxx
```

2. 启动IntelliJ IDEA，在顶部菜单栏，单击**File > Open**，选择本地*.\IdeaProjects\MaxCompute_Studio_Project_Name\scripts*文件夹下的*odps_config.ini*文件。

> **说明** MaxCompute Studio会自动查找该文件夹下的*odps_config.ini*文件，根据这个文件中的配置信息抓取MaxCompute服务端的元数据，然后编译文件夹下的所有脚本。

4.6.3. 开发及提交SQL脚本

本文为您介绍如何在MaxCompute Studio上开发SQL脚本。包括编写和运行SQL脚本。

前提条件

- 已连接MaxCompute项目，详情请参见管理项目连接。
- 已创建MaxCompute Script Module，详情请参见创建MaxCompute Script Module。

编写SQL脚本

1. 在**Project**区域下，右键单击**scripts**，选择**New > MaxCompute SQL脚本**。

2. 在**New MaxCompute SQL Script**对话框，配置参数信息，单击**OK**。

   - **Script Name**：脚本名称。
   - **MaxCompute Project**：目标MaxCompute项目。单击+即可新建一个MaxCompute项目连接，配置详情请参见管理项目连接。
3. 在脚本编辑界面中编写SQL。SQL语法详情请参见SQL概述。

> **说明**
>
> - 支持跨项目空间资源依赖。例如，脚本绑定了项目A的同时，允许访问项目B下的table1（ProjectB.table1）。
> - MaxCompute Studio支持设置SQL脚本编辑器，详情请参见概述。

提交SQL脚本

在提交SQL脚本前您需要根据自身需求进行相关设置。MaxCompute Studio提供了丰富的设置功能，您可以在编辑器页面上方的工具栏中快速设置。设置主要分为以下三种：

- 编辑器模式：
  - **单句模式**：会将提交的脚本文件按 ; 分隔，逐条提交到MaxCompute服务端执行。
  - **脚本模式**：为最新开发模式，可将整条脚本一次提交到MaxCompute服务端，由MaxCompute服务端提供整体优化，效率更高，推荐使用此模式。
- 类型系统：类型系统主要解决SQL语句的兼容性问题。分为以下三种类型：
  - **旧有类型系统**：MaxCompute旧类型的系统。
  - **MaxCompute类型系统**：MaxCompute 2.0引入的新类型系统。
  - **Hive类型系统**：MaxCompute 2.0引入的Hive兼容模式下的类型系统。
- 执行模式：
  - **默认**：稳定版本。
  - **查询加速**：包含查询加速（MCQA）新特性。
  - **加速失败重跑**：支持作业在查询加速失败时，重新执行作业。

1. 完成SQL脚本编写后，单击工具栏或侧边栏上的图标，即可将SQL脚本提交到MaxCompute服务端运行。

   

   > **说明** 当SQL中存在变量时（如上图中的${bizdate}），会弹出对话框，提示您输入变量值。

2. 在SQL任务运行前，IntelliJ IDEA会向您提示预估的SQL费用。确认费用后，在Confirmation对话框，单击OK。

   

   > **说明**
   > - 在工具栏上，单击图标，可以更新SQL脚本中使用的元数据，例如表、UDF。如果MaxCompute服务端存在表或函数，但MaxCompute Studio提示表和函数不存在时，请尝试使用该功能更新元数据。
   > - SQL依赖于您在Project Explore窗口中添加的项目元数据，系统先在本地进行编译，无编译错误后会提交到服务端执行。
   > - SQL执行过程中会显示运行日志。当SQL开始在MaxCompute服务端运行时，会自动打开任务详情页签，显示运行作业的基本信息。

3. 在控制台结果页签查看SQL运行结果。

   单句模式下存在多条语句时，系统会显示每条语句的运行结果。

   

4.7. 开发Java程序

4.7.1. 概述

本文为您介绍使用MaxCompute Studio开发Java程序的流程以及相关目录。

开发Java程序流程 Module目录 warehouse目录

开发流程

通过MaxCompute Studio开发Java程序的流程如下：

1. 创建MaxCompute Java Module。
2. 开发Java程序。您可以参考如下示例开发不同的Java程序：
   - 开发UDF
   - 开发MapReduce
   - 开发Graph
   - 查询非结构化数据
3. 打包、上传和注册。

Module目录

创建MaxCompute Java Module后，MaxCompute Studio会自动创建一个Module。Module目录内容如下：

- *examples*：示例代码，包括单元测试示例。您可以参考示例开发单元测试脚本。
- *src/main/java*：开发Java程序的源码。
- *warehouse*：存储MaxCompute项目的表（包括Schema和数据）和资源。

warehouse目录

*warehouse*目录存储MaxCompute项目的表（包括Schema和数据）和资源，用于执行UDF或MapReduce。

- *warehouse*目录包含项目名、资源（*_resources_*）、表（*_tables_*）、表名、表结构（*_schema_*）和表数据（*data*）。
- 表结构（*_schema_*）文件中配置项目名、表名、列名和类型，并通过冒号（:）分隔。分区表需要配置分区列。图中*wc_in1*为非分区表，*wc_in2*为分区表。
- *data*文件采用标准CSV格式存储表的数据：
  - 特殊字符为逗号（,）、单个双引号（"）和换行符（ `\n` 或 `\r\n` ）。
  - 列分隔符为逗号（,），行分隔符为换行符（ `\n` 或 `\r\n` ）。
  - 如果列内容包含特殊字符，需要在该列内容前后加上双引号（" "）。例如 `3,No` 写为 `"3,No"` 。
  - 如果列内容包含单个双引号（"），则所有的单个双引号（"）需要转义成双引号（" "）。例如 `a"b"c` 写为 `"a""b""c"` 。
  - `\N` 表示该列为NULL，如果该列内容为 `\N` （STRING类型），需要转义为 `"""\N"""` 。
  - 文件字符编码为UTF-8。

4.7.2. 创建MaxCompute Java Module

MaxCompute Studio支持开发Java UDF、MapReduce和Graph等程序，首先您需要新建一个MaxCompute Java Module。本文为您介绍如何新建MaxCompute Java Module。

前提条件

已连接MaxCompute项目，详情请参见管理项目连接。

操作步骤

1. 启动IntelliJ IDEA，在顶部菜单栏，单击File > New > Module...。
2. 在New Module对话框的左侧导航栏，单击MaxCompute Java。
3. 配置Module SDK文件位置，单击Next。
4. 填写Module name，单击Finish。

执行结果

完成上述步骤，MaxCompute Studio会自动创建一个Maven Module，同时完成下列内容：

- 引入MaxCompute相关依赖，详情请查看*pom.xml*文件。
- 创建*examples*示例代码目录。详情请参见Module目录。
- 创建*warehouse*目录用于存储本地调试所需的数据。详情请参见warehouse目录。

后续步骤

完成MaxCompute Java Module创建后，即可开发Java程序。详情请参见：

- 开发UDF
- 开发MapReduce
- 查询非结构化数据
- 开发Graph

4.7.3. 开发UDF

本文为您介绍如何在MaxCompute Studio上开发UDF，包括编写UDF和调试UDF。

开发调试UDF

前提条件

您需要完成以下操作：

- 管理项目连接
- 创建MaxCompute Java Module

背景信息

您可以按照本文介绍自行开发UDF，也可以单击MaxCompute > 创建UDF直接创建函数。如下图所示。

编写UDF

1. 在Project区域，右键单击Module的源码目录（即src > main > java），选择new > MaxCompute Java。

2. 填写Name和Kind，单击OK 。
   - Name：创建的MaxCompute Java Class名称。如果还没有创建Package，在此处填写packagename.classname，会自动生成Package。
   - Kind：选择类型为UDF。目前支持的类型包含自定义函数（UDF/UDAF/UDTF）、MapReduce（Driver/Mapper/Reducer）和非结构化开发（StorageHandler/Extractor/Outputer）等。
3. 创建成功后，在编辑界面开发Java程序。

   Java UDF示例请参见JSON字符串获取示例。

通过本地运行调试UDF

通过本地运行方式测试，查看UDF的运行结果是否符合预期。

1. 右键单击编写完成的Java脚本，选择Run。
2. 在Run/Debug Configurations页面上配置运行参数。

   
   - MaxCompute project：UDF运行使用的MaxCompute空间。本地运行时选择local。
   - MaxCompute table：UDF运行时需要使用的MaxCompute表的名称。
   - Table columns：UDF运行时需要使用的MaxCompute表的列信息。
3. 单击OK，开始运行。

> **说明**
> - 本地运行会读取*warehouse*中指定的表数据作为输入，您可以在控制台查看日志输出。
> - 如果指定的MaxCompute项目的表数据未被下载至*warehouse*目录中，会先下载数据；如果数据已经下载，则跳过此步骤。
> - 关于*warehouse*的说明，请参见warehouse目录。

通过单元测试调试UDF

参考*examples*目录下的单元测试实例，编写自己的测试用例。

后续步骤

完成开发和调试UDF之后，需要对UDF代码打包、上传和注册。详情请参见打包、上传和注册。

4.7.4. 开发MapReduce

本文为您介绍如何在MaxCompute Studio上开发MapReduce，包括编写Mapreduce、调试Mapreduce、打包、上传和运行Mapreduce。

开发MapReduce

前提条件

您需要完成以下操作：

- 管理项目连接
- 创建MaxCompute Java Module

编写MapReduce

1. 在Project区域下，右键单击Module的源码目录（即src > main > java），选择new > MaxCompute Java。
2. 创建Driver。填写Name和Kind，单击OK 。
   - Name：创建的MaxCompute Java Class名称。如果还没有创建Package，在此处填写packagename.classname，会自动生成Package。
   - Kind：选择类型为Driver。目前支持的类型包含自定义函数（UDF/UDAF/UDTF）、MapReduce（Driver/Mapper/Reducer）和非结构化开发（StorageHandler/Extractor/Outputer）等。

     > 说明 创建Mapper和Reducer时，请选择Kind分别为Mapper和Reducer。
3. Driver创建成功后，在编辑界面开发Java程序。

   Java模板已自动填充框架代码，您只需设置输入表、输出表、Mapper和Reducer类等信息。

   > 说明 MapReduce开发详情请参见编写MapReduce（可选）。
4. 以同样的方式创建Mapper和Reducer。

通过本地运行调试MapReduce

通过本地运行方式测试，查看Mapreduce的运行结果是否符合预期。

1. 右键单击编写完成的Java脚本，选择Run。
2. 在Run/Debug Configurations页面上选择此次运行的MaxCompute项目空间名称。
3. 单击OK，开始运行。

   > 说明
   > - 本地运行会读取*warehouse*中指定的表数据作为输入，您可以在控制台查看日志输出。
   > - 如果指定的MaxCompute项目的表数据未被下载至*warehouse*目录中，会先下载数据；如果数据已经下载，则跳过此步骤。
   > - 关于*warehouse*的说明，请参见warehouse目录。

通过单元测试调试MapReduce

您可以参考*examples*目录下的*WordCount*单元测试示例，编写测试用例。

打包上传

调试成功之后，将Java程序打成Jar包，并作为资源上传至MaxCompute服务端，详情请参见打包、上传和注册。

运行MapReduce

通过MaxCompute客户端运行MapReduce。

1. 在左侧导航栏，单击Project Explorer。
2. 右键单击项目名称，选择Open in Console。
3. 在Console区域，执行如下命令运行MapReduce。更多命令请参见Jar命令。

```
jar -libjars wordcount.jar -classpath D:\odps\clt\wordcount.jar com.aliyun.odps.examples.mr.WordCount wc_in wc_out;
```

4.7.5. 开发Graph

本文为您介绍如何使用MaxCompute Studio开发Graph，包括编写Graph、调试Graph、打包上传和运行Graph。

开发Graph

前提条件

您需要完成以下操作：

- 管理项目连接
- 创建MaxCompute Java Module

编写Graph

1. 在Project区域，右键单击Module的源码目录（即src > main > java），选择New > MaxCompute Java。
2. 填写Name和Kind，单击OK 。
   - Name：填写创建的MaxCompute Java Class名称，如果还没创建package，可以在此处填写packagename.classname，会自动生成package。
   - Kind：选择创建的类型。Graph支持GraphLoader或者Vertex。
3. 创建成功后，在编辑界面开发Java程序。更多Graph开发，请参见编写Graph（可选）。

调试Graph

通过本地运行方式测试，查看Graph的运行结果是否符合预期。

1. 右键单击编写完成的Java脚本，选择Run。
2. 在Run/Debug Configurations页面上配置运行参数。
   - MaxCompute project：选择运行Graph的Maxcompute项目。
   - Download Record limit：下载数据记录限制。默认为100条。
3. 单击OK，开始运行。

> 说明
> - 本地运行会读取*warehouse*中指定的表数据作为输入，运行过程中您可以在控制台查看日志输出。
> - 如果指定的MaxCompute项目的表数据未被下载至*warehouse*目录中，会先下载数据；如果数据已经下载，则跳过此步骤。
> - 每运行一次本地调试，都会在已有工程目录下新建一个临时目录。
> - 关于*warehouse*的说明，请参见warehouse目录。

打包上传

调试成功之后，将Java程序打成JAR包，并作为资源上传至MaxCompute服务端。详情请参见打包、上传和注册。

运行Graph

通过MaxCompute客户端运行Graph。

1. 在左侧导航栏，单击Project Explorer。
2. 右键单击项目名称，选择Open in Console。
3. 在Console区域，执行如下命令运行Graph。更多命令请参见JAR命令。

```
jar -libjars xxx.jar -classpath /Users/home/xxx.jar com.aliyun.odps.graph.examples.PageRank pagerank_in pagerank_out;
```

4.7.6. 查询非结构化数据

MaxCompute2.0支持通过外部表的方式直接访问OSS、OTS等。MaxCompute Studio对此提供了一些代码模板方便您快速进行非结构化数据查询开发。本文为您介绍如何使用MaxCompute Studio查询非结构化数据。

查询非结构化数据

前提条件

您需要完成以下操作：

- 管理项目连接
- 创建MaxCompute Java Module

编写StorageHandler、Extractor和Outputter

1. 在Project区域，单击Module的源码目录（即src > main），选择new > java，选择MaxCompute Java。
2. 创建Driver。填写Name和Kind，单击OK 。
   - Name：创建的MaxCompute Java Class名称。如果还没有创建Package，在此处填写packagename.classname，会自动生成Package。
   - Kind：选择类型为Extractor。目前支持的类型包含自定义函数（UDF/UDAF/UDTF）、MapReduce（Driver/Mapper/Reducer）和非结构化开发

（StorageHandler/Extractor/Outputer）等。

> 说明 创建StorageHandler和Outputter时，请选择Kind分别为StorageHandler和Outputer。

3. 创建Extractor成功后，在编辑界面开发Java程序。代码框中模板已自动填充框架代码，只需要自行编写需要的逻辑代码即可。
4. 以同样的方式创建StorageHandler和Outputter。

调试Extractor和Outputter

您可以参考*examples*目录下的*unit test*单元测试示例，编写测试用例调试Extractor和Outputter。

打包上传

调试成功之后，将Java程序打成Jar包，并作为资源上传至MaxCompute服务端，详情请参见打包、上传和注册。

查询非结构化数据

1. 在Project区域，右键单击scripts，选择new > MaxCompute SQL 脚本。
2. 在Script Name后输入SQL脚本名称，MaxCompute project中选择执行脚本的MaxCompute项目，单击OK。
3. 在编辑器中输入创建外表的SQL语句。
4. 输入查询语句，单击运行查询数据。

4.7.7. 打包、上传和注册

Java程序开发完成后，需要打包发布至MaxCompute上才可以使用。本文为您介绍如何打包、上传和注册资源。

打包 上传 注册

背景信息

UDF、MapReduce和Graph等Java程序发布到服务端供生产使用前，要经历打包、上传和注册三个步骤。MaxCompute Studio提供了一键发布功能（即在MaxCompute Studio上依次执行mvn clean package、上传JAR和注册三个步骤）。

打包

1. 右键单击已经编译成功的Java代码，选择Deploy to server...。
2. 在Package a jar and submit resource对话框中，配置相关参数。
   - MaxCompute project：指定目标MaxCompute项目的名称。
   - Resource name：指定打包的资源名。
   - Function name：指定打包的函数名称。
   - Force update if already exists：选择当资源或函数已存在时是否强制更新。

3. 单击OK，完成打包。

> 说明 如果您有特殊的打包需求，可以自行修改pom.xml打包相关配置。

上传JAR包

打包成功后，需要将该JAR包上传到MaxComptute服务端。

1. 在顶部菜单栏，单击MaxCompute > 添加资源。
2. 在Add Resource对话框中配置相关信息，单击OK。
   - MaxCompute project：指定目标MaxCompute项目的名称。
   - Resource file：指定JAR包路径。
   - Resource name：输入上传的资源名。
   - Force update if already exists：选择当资源或函数已存在时是否强制更新。
3. 在左侧导航栏，单击Project Explorer。
4. 在Project Explorer区域的Resources节点下可以看到该资源。

注册UDF

JAR包上传完成后，需要注册UDF函数后您才可以调用该函数。

1. 单击顶部菜单栏上的MaxCompute，选择创建UDF。
2. 在Create Function页面配置如下参数，然后单击OK。
   - MaxCompute project：选择要上传的Project名称。
   - Function name：函数名称。
   - Using resources：函数依赖的JAR包名称。
   - Main class：JAR的主类。
   - Force update if already exists：当资源或函数已存在时是否强制更新。
3. 在左侧导航栏，单击Project Explorer。
4. 在Project Explorer区域的Functions节点下看到该函数。

4.8. 开发Python程序

4.8.1. 配置Python开发环境

MaxCompute Studio支持您在Intellij IDEA中完成Python开发，例如UDF和PyODPS脚本。本文为您介绍如何配置Python开发环境。

配置Python开发环境

安装PyODPS

1. PyODPS是MaxCompute的PyODPS SDK，详情请参见安装指南及使用限制。
2. 运行Intellij IDEA，在顶部菜单栏上，单击File > Settings。

3. 在Settings页面左侧导航栏，单击MaxCompute Studio。
4. 设置Python path to resolve UDF为本地Python安装目录。

安装Python插件

在Intellij IDEA的插件仓库中搜索Python或者Python Community Edition插件并安装。

1. 在Intellij IDEA顶部菜单栏上，单击File，选择Settings。
2. 在Settings页面左侧导航栏，单击Plugins。
3. 在搜索框搜索Python Community Edition，并安装。

配置Python依赖

配置Studio Module对Python的依赖，即可进行MaxCompute Python的开发。

1. 在顶部菜单栏，单击File > Project Structure。
2. 增加SDK。
   i. 在左侧导航栏上，单击SDKs。
   ii. 单击上方的加号添加Python SDK。
   iii. 单击Apply完成添加。
3. 配置Modules。
   i. 在左侧导航栏上，单击Modules。
   ii. 单击+添加Module依赖Python Facets。
   iii. 单击Apply保存配置。

4.8.2. 开发Python UDF

MaxCompute Studio支持Python UDF开发，本文为您介绍如何开发、测试和注册发布Python UDF。

开发Python UDF 测试Python UDF 发布Python UDF

前提条件

您必须完成以下操作：

- 管理项目连接
- 配置Python开发环境

开发Python UDF

1. 在Project区域MaxCompute Studio目录下，右键单击scripts，选择New > MaxCompute Python。
2. 在Create new MaxCompute python class对话框中输入类名Name，选择类型为Python UDF，单击OK完成。
3. 在编辑框中编写UDF代码。

测试UDF

UDF开发完成后，需要测试代码是否符合预期。MaxCompute Studio支持本地测试，即下载表的部分示例数据在本地运行并进行调试。

1. 右键单击已经编辑完成的Python UDF脚本，选择**RUN**。
2. 在**Edit configuration**页面，配置相关参数，单击**OK**。
   - **MaxCompute project**：UDF运行使用的MaxCompute空间。本地运行时选择**local**。
   - **MaxCompute table**：UDF运行时需要使用的MaxCompute表的名称。
   - **Table columns**：UDF运行时需要使用的MaxCompute表的列信息。
   - **Download Record limit**：下载数据记录限制。默认为100条。

   > **说明**
   > - 如果已经下载数据，则不会再次重复下载。如果需要再次下载，请在MaxCompute客户端使用Tunnel命令下载数据。
   > - 默认下载100条数据，如果需要更多数据测试，请在MaxCompute客户端使用Tunnel命令或者MaxCompute Studio的表下载功能下载数据。
   > - 下载完成后，您可以在*warehouse*目录下该表的*data*文件中看到下载的示例数据。

3. 本地运行框架会根据您指定的列，获取*data*文件中指定列的数据，调用UDF本地运行。

   > **说明** 本地运行是通过PyODPS的pyou脚本实现的，命令为 `pyou hello.Plus<data` 。安装完PyODPS后可以使用相应的命令检查该脚本是否存在。
   > - 如果您是Windows系统，请运行 `${python}/../Scripts/pyou` 命令。
   > - 如果您是mac OS系统，请运行 `${python}/../pyou` 命令。

4. 您可以在控制台查看打印结果。

发布Python UDF

Python UDF测试通过后，即可发布到生产环境上使用。详情请参见打包、上传和注册。

4.8.3. 开发PyODPS脚本

本文为您介绍如何开发PyODPS脚本。

前提条件

已配置Python开发环境，详情请参见配置Python开发环境。

背景信息

PyODPS是MaxCompute Python版本的SDK，提供对MaxCompute对象的基本操作和DataFrame框架。您可以通过PyODPS在MaxCompute上分析数据。

操作步骤

1. 在Project区域右键单击**scripts**，选择**New > MaxCompute Python**。

MaxCompute Python

2. 在Create new MaxCompute python class对话框，填写Name，从Kind列表中选择PyODPS Script。

   Create new MaxCompute python class

3. 新建MaxCompute PyODPS脚本后，PyODPS脚本模板会通过PyODPS Room自动初始化 `odps` 和 `o` 两个对象。通过DataWorks开发PyODPS脚本时，系统会自动创建Room。通过Intellij IDEA开发PyODPS脚本时，需要创建Room，详情请参见PyODPS文档。

   模板

4.9. 管理MaxCompute作业

4.9.1. 作业浏览

通过MaxCompute Studio的Job Explorer可以方便查看当前用户提交的MaxCompute实例情况，包括运行状态、作业类型、起止时间等。

Job Explorer

打开Job Explorer

在顶部菜单栏，单击View > Tool Windows > Job Explorer打开Job Explorer。

查看项目的作业实例

Job Explorer支持按照状态、提交者、日期和用时等信息查询提交的作业列表。例如：

- 按照状态查询提交的作业列表。选择状态为失败，查看过去24小时内执行失败的作业。
- 按照日期查询提交的作业列表。从日期列表框，滑动指针选择时间，获取指定时间段的作业列表。

> 说明
> - 默认只显示符合条件的前1000条作业。如果需要超过1000条的作业信息，请考虑更新过滤条件。
> - 您可以单击Id与名称、提交时间对作业进行简单的排序。

查看作业排队队列

活动状态的作业如果正在排队队列中等待调度，队列位置会展示当前排队的位置，优先级会展现作业的全局优先级。

> 说明 在Running Instances下的作业状态、队列位置等信息会自动更新，作业结束后会从列表中移除。

保存作业日志

目前作业的Logview日志默认保存7天。如果需要查看更长时间的Logview信息，建议您将Logview保存在本地。

1. 双击列表中的作业，在右侧打开作业的详细信息。
2. 在工具栏上，单击**保存**，将日志保存到本地。

   如果需要自定义保存文件的目录，可以在MaxCompute Studio的**Settings**页面中进行配置。

4.9.2. 作业实例详情

本文为您介绍如何查看作业实例详情。

作业实例详情

查看作业实例详情

MaxCompute Studio支持2种方式查看MaxCompute作业实例详情。

- 通过Logview URL或本地的logview离线文件以只读方式打开作业详情。

  使用Logview查看作业的详细信息是比较常用的方式。使用Logview还可以查看其他用户在其他项目空间中提交的任务状态。MaxCompute Studio支持通过输入一个有效的Logview URL打开任意一个作业详情。

  在IDEA顶部菜单栏中，单击**MaxComput > 打开Logview**，选择导出本地的Logview离线文件或者将有效Logview URL复制到**Open job detail by logview**对话框中。

- 双击作业列表下的作业，可以查看该实例的详细信息。

作业详情视图

作业详情页面包括顶部的工具栏，左半部分的基本属性栏以及右半部分详细视图页，其中详细视图页主要包含四个视图：

- **执行图**：以DAG图的方式显示作业整体信息，可查看子任务间的依赖关系以及各个子任务的详细执行计划。
- **详情**：以表视图的方式展示作业详细信息，包括子任务列表、各子任务的Worker列表、Worker处理数据量、执行时间及状态信息等。
- **脚本**：显示该作业提交时所对应的SQL语句以及提交作业的参数配置信息。
- **结果视图**：显示该作业运行结果。

还包含如下信息：

- **时序图**：显示作业执行时间线，可以从不同粒度查看执行的时序，并提供了多种过滤器。
- **概要（JSON）**：以JSON格式显示作业运行详细信息。
- **分析**：提供作业执行散点图、长尾柱状图及数据倾斜图。

工具栏

- 、：页面左右折叠。用来收起或完全展开左右侧视图，允许您最大化某一个视图进行查看。
- ：停止作业。用来中断正在执行的作业，需要具有响应权限才能停止作业（项目所有者或管理员）。
- ：刷新详情。对于运行中作业的基本信息，例如状态、Quota等会自动刷新。但右侧各个详情视图不会自动刷新，需要您手动刷新。
- ：拷贝Logview。

- ▫：在浏览器打开作业详情。即生成Logview URL，并通过浏览器打开。
- ▫：将作业详情信息保存为本地文件。
- ▫：是否开启自动刷新。对于运行中作业，开启自动刷新后，MaxCompute Studio会对作业执行全量定时刷新。

基本信息页面

基本信息页面展示了作业的基本信息，包括ID、创建人、状态、起止时间、计算资源用量、输入项（作业的输入表）、输出项等（作业的输出表）。运行中作业的基本信息会自动定时刷新。

双击表名即可查看对应表的基本信息页面。

执行图

执行图作为日常主要使用工具，以可视化的方式展示Fuxi Job、 FuxiTask以及Operation的依赖关系，同时提供一系列辅助工具，例如作业回放、进度图和热度图等，是排查问题的好帮手。

▫

上图中各序号对应说明如下：

1. 可单击跳转其他层次。
2. 缩放辅助工具。
3. 依赖表。
4. Fuxi Task节点。
5. 鹰眼。
6. 展示。
7. 默认打开FuxiTask层依赖。

执行图可展示三个维度的作业依赖关系：Fuxi Job层、Fuxi Task层、Operation层。可单击向上箭头进行维度切换，默认会展示Fuxi Task层依赖关系。

- Fuxi Job层

  单击**MaxCompute Job**打开Fuxi Job层。Fuxi Job层节点内包含Fuxi Task名称、起止时间等。

  ▫

- Fuxi Task层

  双击任一Fuxi Job节点即可进入Fuxi Task层。

  当有多个Fuxi Job时，默认打开最后一个Fuxi Job的Fuxi Task层。该层可展示Fuxi Task的依赖关系，输入输出表及分区等信息。当作业结束后单击工具栏中的**进度图**，可以选择**热度图**、**输出热度图**、**Task时间热度图**和**Instance热度图**等。进度图表示节点的完成进度，热度图通过颜色区分节点热度。

  ▫

  Fuxi Task节点内容如下：

  - Instance Count：表示为 `a/b/c` ，指某一时刻正在运行的子任务实例个数为a，已结束任务实例个数b，总任务实例个数c。
  - I/O Records：同理为某一时刻的输入记录量和输出记录量。
  - 百分比与橙色进度条：表示该任务运行情况，该比例根据子任务运行实例分析得出。
  - 子任务间连线：显示输出的记录数量。箭头表示数据流动方向。

- Operation层

  双击任一Fuxi Task即可打开Operation层。

Operation层揭示了Fuxi Task内在的运行方式，单击任一节点即可显示Operaiton完整信息。

> 说明 非SQL类型作业，仅能展示Fuxi Job和Fuxi Task层作业，不展示Operation层。

详情页

主要针对SQL DML类作业，展示作业在计算集群上的Fuxi Task列表、计算节点列表等。通常一个作业对应一个或多个Fuxi Job，每个Fuxi Job拆分成多个Fuxi Task（阶段），每个Fuxi Task包含多个Fuxi Instance（Worker）。右键单击Fuxi Instance可以查看作业运行的标准输出、标准错误和Debug Info。

对应的序号说明如下：

1. Fuxi Job Tab。
2. Fuxi Task列表。
3. 每个Fuxi Task详细信息及计算节点列表。

作业回放

MaxCompute Studio支持作业回放功能，作业回放就像播放媒体文件一样，可在12s内回顾该作业执行的历史轨迹。该功能主要用于帮助用户了解MaxCompute实例在不同时刻运行状态，快速判断子任务级运行顺序及消耗时间，掌握作业执行关键路径，从而针对运行较慢的子任务进行优化。

单击 > 按钮即可开始播放，再次单击则暂停。您也可以手动拖动进度条。

> 说明 回放功能仅通过时间估算某一个时刻IO数据量，从而确定完成进度，并不能代表该时刻真实IO数据量。Running状态作业不支持回放功能。

时序图

以甘特图的方式展示作业分布式执行的详细数据，可以调整展示粒度，将每一个计算节点都在甘特图中展示。可以通过甘特图直观的看出作业运行的时间瓶颈和长尾节点等。同时提供多种过滤器，能够直接筛选出作业执行的关键路径、最大数据节点和最长时间节点等。

分析页

分析页展示作业的长尾节点（Worker）、数据倾斜节点（Worker）。展示节点散点图、柱状图和辅助作业执行瓶颈诊断。散点图和柱状图支持从图中节点的准详情页查看Fuxi Instance详情。

结果页

结果页会根据作业类型及提交作业时的参数设置展示不同页面。

- SELECT语句并且设置 `odps.sql.select.output.format = HumanReadable` ，结果以文本方式展示。
- SELECT语句并且未设置output format参数，结果以TABLE方式展示。
- 对于数据输出到表的脚本，展示输出表名及表详情页的链接。
- 对于异常作业，结果页显示异常详情。

4.10. 工具集成

4.10.1. 集成MaxCompute客户端

MaxCompute Studio集成了MaxCompute客户端，您可以在MaxCompute Studio中直接运行MaxCompute客户端。

MaxCompute Studio MaxCompute客户端

背景信息

MaxCompute Studio中已包含最新版MaxCompute客户端程序，并指定为默认客户端。您也可自行指定其他版本客户端程序。

操作步骤

1. 配置客户端安装路径。
   i. 在顶部菜单栏，单击File > Settings...。
   ii. 在Settings页面左侧导航栏，单击MaxCompute Studio > SDK & Console。
   iii. 在Installed Location后选择MaxCompute客户端的本地安装路径。
   iv. 单击OK，完成客户端安装路径配置。
2. 在MaxCompute Studio中打开MaxCompute客户端。
   i. 在IDEA左侧导航栏，单击Project Explorer。
   ii. 在选中的项目节点上，右键单击选择Open in Console，打开MaxCompute客户端。
   iii. 在Console区域，运行客户端命令。详情请参见常用命令列表

4.11. Studio视频介绍

- Studio 安装介绍
- 通过Studio管理数据
- 通过Studio编辑SQL
- Studio SQL Scripting
- 通过Studio开发UDF
- 通过Studio remote debug
- 通过Studio查看所有job

5.相关下载

本文将为您提供在使用MaxCompute过程中，可能用到的相关工具及插件的下载地址。

MaxCompute及插件

- SDK下载信息：如果您使用Maven，可以从Maven库中搜索odps-sdk，获取不同版本的Java SDK。
- 客户端：进入客户端下载页面，下载所需版本客户端。
- IntelliJ IDEA开发插件：下载IDEA工具 ，Studio插件即可下载所需的IntelliJ IDEA开发插件。
- JDBC：MaxCompute提供开源JDBC，您可以在GitHub下载JDBC。

  > 说明 您可以在云栖社区查看发布信息或提问。

- PHP SDK：您可以在GitHub下载PHP SDK。