知乎 首发于
数择云-数据中台从零构建

零成本构建私有化DataWorks平台 - Part 1

起点大数据中台
阿里数据中台深度参与者，拥有一整套的数据中台方法论。

4 人赞同了该文章

前言

数据中台的价值

知乎 首发于
数择云-数据中台从零构建

DataWorks在数据中台构建的核心价值

如下节选阿里云官网DataWorks产品定位介绍：

> DataWorks致力于为数据开发者、数据分析师、数据资产管理者打造一个具备开放自主开发能力与全栈数据研发能力的一站式、标准化、可视化、透明化的智能大数据全生命周期云研发平台。DataWorks赋予用户仅通过单一平台即可实现数据传输、数据计算、数据治理、数据分享的各类复杂组合场景的能力。
> 同时DataWorks持续打造符合企业级数仓、数据中台构建要求的功能模块，为企业业务数字化转型提供最大程度支持。

DataWorks从功能上划分，主要包括如下维度的功能：

- 数据同步
- DataStudio（数据开发）
- 数据运维
- 数据治理
  - 数据质量（DQC）
  - 数据地图（元数据管理）

使用社区开源方案构建DataWorks平台

DataWorks涉及的功能比较多，限于篇幅，基于社区开源方案构建DataWorks平台会分三篇文章：

- 基于社区解决方案搭建**数据同步**平台。这块核心价值点和功能：

知乎 首发于 数择云-数据中台从零构建

- 基于社区解决方案搭建**数据开发**平台。数据开发和数据运维这块是不分离的，所以这块附带的把数据运维的功能也集成进来。这块涉及核心功能点：
  - 任务IDE
  - 工作流调度引擎
  - 数据运维
- 基于社区解决方案搭建**数据治理**平台。这块会引入两个非常成熟的平台：
  - 数据质量（DQC）
  - 元数据管理平台

本文会详细介绍**基于社区解决方案搭建数据开发平台。**

数据开发平台构建介绍

这块会涉及三个比较大的核心功能：

- 任务IDE
- 工作流调度引擎
- 数据运维

这块需要说明些，为啥会把**任务IDE**跟**工作流调度引擎**分离开来，主要原因是：

- 任务IDE模块需要做的事情是蛮多的：
  - 智能化的检测各种类型脚本（Python、Shell等）语法、HQL（Hive SQL、Spark SQL等各种大数据计算引擎）语法。
  - 任务多版本管理：这块跟Devops相关联，维护开发的任务的各个版本，便于查看Change log、Diff，便于回滚和版本切换。

Oozie、Airflow、Easy Scheduler对IDE的支持非常弱。使得在数据开发环境，完全依赖于工作流调度引擎会比较痛苦。

综上原因，我们是需要分开IDE与工作流调度引擎环节的。

任务开发IDE - Scriptis

任务开发IDE这块目前开源的解决方案并不多，不过深入调研了一段时间发现还是有具体的解决方案。**Scriptis是微众银行**开源的一个非常nice的任务开发IDE：

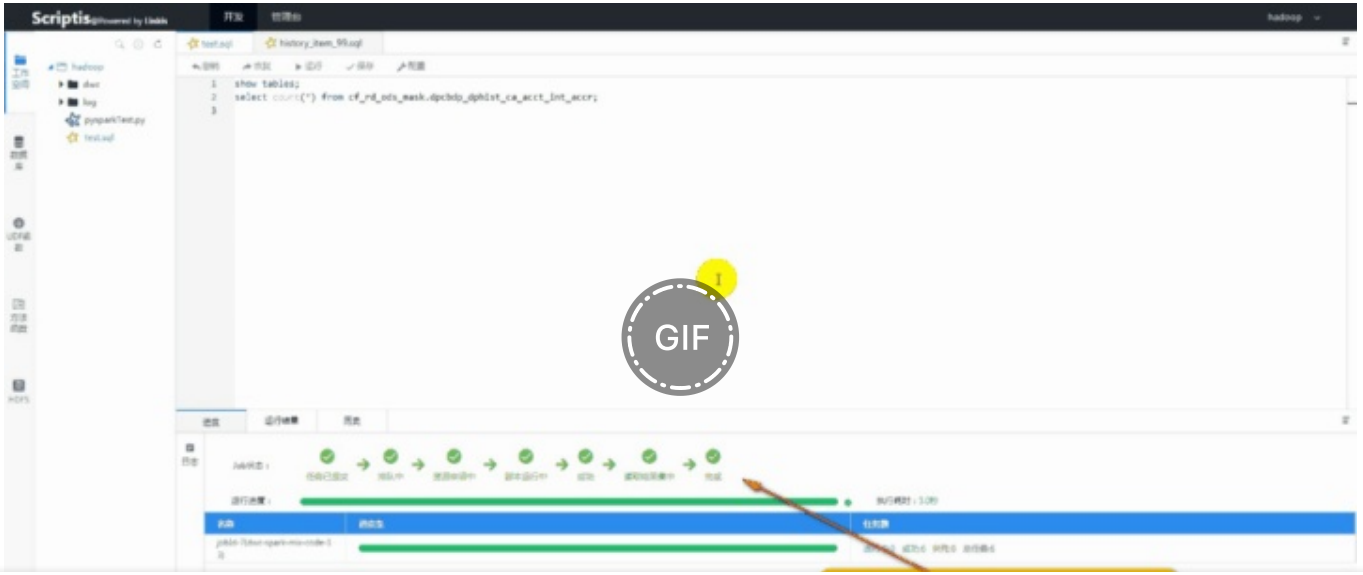
> Scriptis是一款支持在线写SQL、Pyspark、HiveQL等脚本，提交给Linkis（Linkis是什么？点我了解）执行的数据分析Web工具，且支持UDF、函数、资源管控和智能诊断等企业级特性。

Scriptis系统核心优点：

- 脚本编辑：多语言、自动补全、语法高亮、SQL语法纠错；

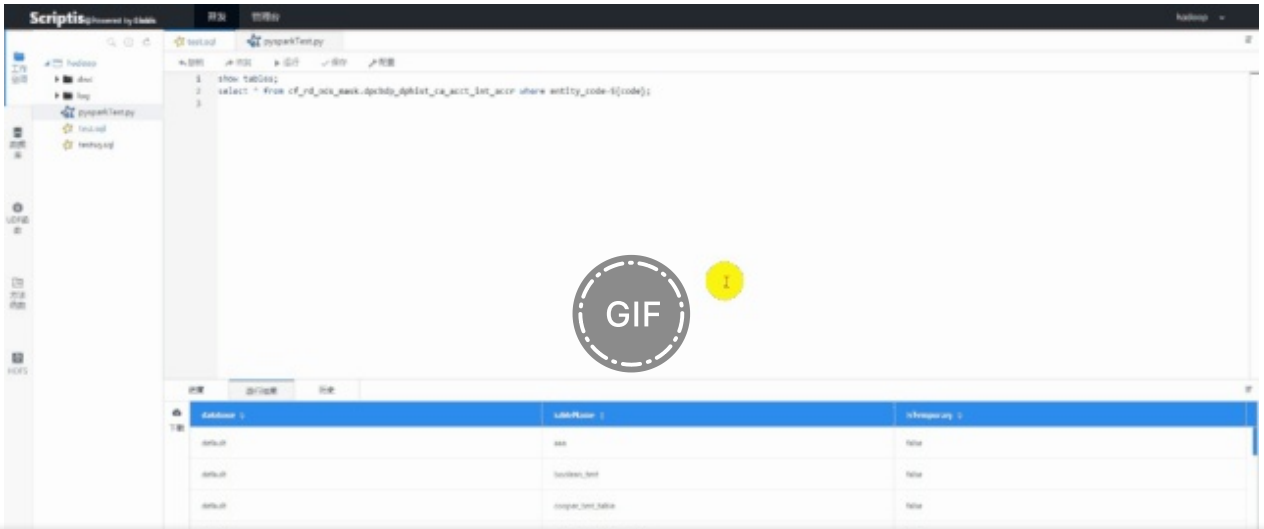
知乎 首发于 数择云-数据中台从零构建

- 算引擎：基于Linkis对接多计算引擎：Spark、Hive、TiSpark等；
- 运行时功能：Job全生命周期展示、错误码；



知乎 首发于 **数择云-数据中台从零构建**

- 结果集：多结果集支持、结果集别名自定义、结果集一键发布到报表系统；



知乎 首发于 数择云-数据中台从零构建

- 管理台：多租户资源管控、引擎参数个性化配置、任务和会话管理。

Scriptis整体的功能做的挺好的，并且完全能够满足我们对IDE的需求，如果能够在如下地方做些完善就真的能够跟DataWorks相关功能一决高下了。

- 目前**Scriptis** IDE开发界面完全都是白色的，对于大量数仓任务需要开发的**数据开发同学**而言容易视力疲劳，如果能够支持暗黑系皮肤供数据开发同学自定义切换体验就更好。
- 目前**Scriptis** 初步能够支持任务多版本支持，每次任务的运行系统会记录其对应的脚本具体内容。不过这块做的比较粗糙，以下情况很难满足：
  - 任务只有执行后，系统才会记录其具体内容。这样不够灵活，理论上有个发布的操作，系统能够根据发布的情况自动的为其分派版本，便于回滚以及Diff查看。
  - 无法查看不同版本之间（或者是不同执行任务之间）的Diff情况，不太利于一些场景下问题的排查。

引入的原因

虽然**Scriptis**有些小的瑕疵，但是整体的功能还是挺不错的，满足我们的需求，而且它是唯一的选择。另外由于如下原因，让我们有理由相信它的爆发力以及未来的潜质：

- 有微众银行大数据团队的全力持续投入，以及社区的力量不断贡献于其中。
- 微众银行自身的业务是该系统的重度用户，能够很好的支撑微众银行的业务，支持中小型互联网公司的业务需求完全不在话下。
- 任何开源的系统在最开始都会有不稳定、不完善的地方，只要走在对的路上，我觉得离成功就不远了。而且我相信做数据中台平台最好的IDE是一条对的路，我觉得**Scriptis**肯定能够做的非常好的。

工作流调度引擎 - Easy Scheduler（集成数据运维功能）

- Airflow

Easy Scheduler是由国内易观公司开源的工作流调度引擎，从个人角度而言具备以下优点：

- 中文支持比较好、比较符合国内开发者的使用习惯；
- 支持任务类型比较丰富，能够满足基本的离线数仓和实时数仓的需求；
- 系统部署简便，非常易用；
- 比较易于集成，系统提供大量对外Restful接口，便于集成到其他系统中去。
- 对HA的支持比较不错，Master和Worker完全无状态，便于增加工作节点提升系统吞吐量和服务能力。

如下是Easy Scheduler与其他工作流引擎的详细对比情况（选自官网，深度使用过，对比完全可信）：

知乎 首发于
数择云-数据中台从零构建

以下是Easy Scheduler的部分使用截图：

知乎 首发于
数择云-数据中台从零构建

知乎 首发于
数择云-数据中台从零构建

界面风格跟DataWorks比较接近，如果任务DAG依赖关系不需要支持小时任务依赖分钟级任务等细粒度任务实例之间的相互依赖，Easy Scheduler完全符合需求。

有几点感觉做的非常好的地方：

- Easy Scheduler对数据源做了一个专门的维护和管理，这块是做的非常好的地方。便于后续细粒度的平台权限控制以及对应功能的支持。
- 数据中台平台，从底层的数据同步、数据开发、数据治理、数据服务到上层的BI可视化，每层之间的相互关联是以**数据源**为枢纽。把数据源单独抽离出来，做一个统一的入口，便于不同层级系统间的依赖解耦，另外后续的整个大平台的权限易于设计和维护。
- Easy Scheduler监控中心Dashboard做的挺好的，对系统各个组件的监控非常直观以及便于运维：

知乎 首发于 **数择云-数据中台从零构建**

- Easy Scheduler集成了对**数据运维**的支持：
  - 首页的**任务状态统计**、**流程状态统计**、**队列统计**、**命令状态统计**、**流程定义统计**基本上满足我们对T + 1场景下离线数仓开发的需求和T + 0场景下实时数仓开发的需求。
  - 任务维度的补数据、重跑任务基本上满足我们补数以及任务失败后重跑的需求，整体上能够支持我们的需求。如果能够支持更细粒度的重跑机制会更好。
- 支持工作流之间的相互依赖，而且可以把其他工作流做为一个子流程配置到当前工作流中，这个是非常重要的一个功能。

总结

对Easy Scheduler功能以及在使用过程中的一些总结：

- Easy Scheduler满足中小型互联网公司对工作流调度引擎的一切想象，在中大型互联网公司也同样适用；

知乎 首发于
数择云-数据中台从零构建

集成的数据开发平台如何Work

上面介绍了**任务开发IDE系统Scriptis**和**工作流调度引擎Easy Scheduler**，那大家会比较关心：这两个系统如何协作？

如下是我们在实践过程中感觉不错的经验：

- 在**Scriptis**平台开发各个任务，使用**Scriptis** IDE功能。边开发边测试，测试OK了，再在**Easy Scheduler**做如下操作：
  - 创建工作流，如果之前没有创建的话；
  - 创建任务；
  - 设置当前任务跟工作流中其他任务的相互依赖关系；
  - 配置任务的SQL语句，如下图所示：

知乎 首发于
数择云-数据中台从零构建

让**Scriptis**聚焦于：

- 任务开发IDE
- 任务开发、调试
- 任务版本管理

让**Easy Scheduler**聚焦于：

- **周期性**工作流、任务DAG依赖关系的配置和调度
- 多资源组的支持；
- 多数据源的支持；
- 数据运维的支持；
- 任务监控报警的支持；

使得我们拥有一个完善、强大的数据开发平台，功能丝毫不属于阿里云DataWorks的Data Studio平台（数据开发平台）。

注意事项

在**Scriptis**与**Easy Scheduler**之间协调工作的过程中，以下点着重需要注意：

- **Scriptis**和**Easy Scheduler**必须要在同一个环境中，底层基于的Hadoop环境必须保持一致；
- **Scriptis**和**Easy Scheduler**都支持任务自定义环境变量，在**Scriptis**任务运行过程中设置环境变量之后，一定记得在**Easy Scheduler**相关任务中也配置对应的环境变量，否则会导致任务失败。
- 记得把**Scriptis**中创建的HDFS、资源文件此都在**Easy Scheduler**中做同样的操作，否则会导致

知乎 首发于
数择云-数据中台从零构建

总结

基于社区的**Scriptis + Easy Scheduler**，让打造一个强大的、完全私有化部署的数据开发平台成为可能。

后续会分两篇文章，深入介绍如何强大的**数据同步**平台如何私有化构建？强大的**数据治理**平台如何私有化搭建？让企业私有化搭建强大的**DataWorks**平台成为可能！

发布于 2019-10-12

数据中台 大数据 数字化

文章被以下专栏收录

数择云-数据中台从零构建
物竞天择，企竞数择。数字化转型，企业存亡生死攸关。数择云，国际领先的数据中...

进入专栏

推荐阅读

知乎 首发于 **数择云-数据中台从零构建**

起点大数据 发表于数择云-数...

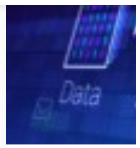
从零构建数据中台平台

生：你算什么？魔都垃圾分拣阿
姨：你是什么垃圾？ What？？？
自从上海开始推行垃圾分类后...

Informatica 数据管理

NBI 发表于大数据可视...

数据治理中的数据血缘关系是什



大数据在...

冷思考：

2 条评论 ⇌ 切换为时间排序

写下你的评论...

就是这么骚 1 个月前

你好，文章中体现了数据开发平台构建的思路，在你的文章中没找到搭建数据同步平台和搭建数据治理平台，能否共享一下这两方面的资料

赞

我来也 27 天前

期待续集part2

赞

▲ 赞同 4 ▼ 2 条评论 分享 收藏 ···