知乎 首发于
数择云-数据中台从零构建

从零构建数据中台平台

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。

最近在使用社区开源系统从零构建数据中台平台，有挺多的收获，后续会记录从零构建数据中台平台具体实操经验。希望能够帮助传统企业以及中小型互联网公司低成本、更好的数字化转型。

数据中台平台从技术角度，需要支持以下维度的功能：

  - 工作流调度引擎
- 数据运维
- 数据治理
  - 数据质量
  - 元数据管理
- 数据安全
- 平台权限
- DevOps
- 监控报警
- 数据服务
- BI可视化展示

IAAS

底层引擎支撑

- 离线开发
  - CDH（Hadoop发行版，包括Hive、Spark、HDFS、YARN、MR、ZK等基础组件）
- 实时开发
  - Flink
  - 数据源：Kafka、关系型数据库（主要是Mysql）
  - 目标源：Kafka、Hbase、ES

数据同步

- DataX（离线场景）
- Flume（实时场景）

如果能够做到把DataX和Flume做整合，给上层提供统一的数据同步接口会更好。

数据开发

这块涉及两个重要环节。数据开发IDE、工作流调度引擎。

数据开发IDE

- 支持各种脚本、SQL　IDE功能。

工作流调度引擎

可选的引擎主要有如下四种：

- **Easy Scheduler**
- Azkaban
- Oozie
- Airflow

数据运维

能够支持对数据开发运行的任务实例进行：

知乎 首发于 数择云-数据中台从零构建

- 基线报警&报警配置

数据治理

数据治理主要涉及两个大的功能：

- 数据质量（DQC）
- 元数据管理

数据质量（DQC）

DQC这块开源比较好的选择是eBay开源的Griffin。

数据质量在数据科学领域是至关重要的。在大数据时代，企业决策调整，商机发现等越来越依赖于大数据的数据分析和数据挖掘，而数据质量的保证是所有一切数据分析和数据挖掘的基础。

Apache Griffin是一个应用于分布式数据系统中的开源数据质量解决方案。在Hadoop, Spark, Storm等分布式系统中，提供了一整套统一的流程来定义和检测数据集的质量并及时报告问题。

元数据管理

元数据管理核心在更好的维护数据血缘关系，能够支持表级别、字段级别数据血缘关系。

为后续的数据发现、数据追朔、标签体系构建、数据资产运营等提供支撑。

- Atalas

初定Atalas，具体的对比可以参考后续markdown文档。

数据安全

数据安全涉及几大类：

- 底层Hadoop层面存储、计算能力的数据安全，这块可以用apache的**Kerberos**或者Ranger解决。
- 组件之间通信协议数据安全：
  - 工作流调度Master与Slave之间通信协议（主要是**HTTP**）数据的加密，以及安全认证。**这个主要是混合部署模式下非常有价值和必要。**
  - 数据服务对外暴露API需要有访问安全控制，以及根据用户需要对数据进行加密。

平台权限

多用户&多租户资源隔离
Kerberos或Ranger只解决Hadoop层面多用户的隔离。但是对于数据中台平台，目标的对象不仅仅是Hadoop的存储、计算能力，而是数据开发（包括实时和离线开发）这个重要环节。所有的DQC、元数据管理、数据安全都是在努力为它服务。

一个公司，拥抱数据中台，就自然而然意味着数据中台平台是整个公司数字化转型的基础设施。

就自然而然会使得如下职责人员在同一个平台协作：

知乎 首发于 数择云-数据中台从零构建

- 业务方（主要是使用数据服务提供的API接口）
- 运营（主要是使用HQL、Adhoc场景取数，以及使用BI报表）

所以，使得多租户权限控制和隔离非常重要。

多租户隔离需要做到几点：

- 数据源使用权限的严格限定；
- 开发任务访问、执行权限的严格限定；
- 数据服务对外暴露API权限的严格限定

DevOps

数据中台平台在DevOps场景，主要是**多环境支持的需求**。这个其实前期可以做的比较弱，后期可以慢慢加强。

核心解决问题是多套环境以及规范测试开发、上线流程。

严格意义上，一个完整的上线规范包括：测试 -> 预发 -> 生产，这套流程是需要在我们的数据中台平台上完成的。我们平台需要有套机制帮助用户更好的：

- 规范流程
- 提供预发、生产变更机制
- 提供预发、生产回滚机制
- 多线下版本维护
- 细粒度权限控制
- 监控&报警支持

知乎 首发于 **数择云-数据中台从零构建**

需要有套完整的监控报警服务，能够根据自定义上报上的数据或者是采集的数据，根据配置的阈值以及规则进行**有效的报警**，并**对报警的情况进行监控。**

数据服务

数据服务主要解决问题是：把底层关系型数据库的数据API化，使得上层的应用能够更好的使用数据中台平台开发好的数据。

这块涉及两个问题要解决：

- SQL Proxy
- API网关

BI可视化

BI可视化解决问题是把底层各种关系型数据库数据、多维聚合数据图表化展示。

主要支持数据源

- 传统关系型数据库
- Kylin等多维离线聚合OLAP引擎
- Hbase等NoSQL场景
- ES

丰富图表类型支持

要求尽量支持如下图表类型：

- KPI
- 漏斗图
- 桑基图
- 雷达图
- 气泡图
- 对比图
- 标签云
- 热点图
- 关系图

END

关注我们，如果对从零构建数据中台平台有兴趣，也可以一起加入我们。扫码添加个人微信号，备注“数据中台”，我会第一时间通过，一起会数据中台开源贡献力所能及的力量。

知乎 首发于 **数择云-数据中台从零构建**

发布于 2019-10-12

大数据 数据中台 大数据时代

文章被以下专栏收录

▲ 赞同 4 ▼ 添加评论 分享 收藏 ···

知乎 首发于 **数择云-数据中台从零构建**

推荐阅读

阿里腾讯极其看重的数据中台，我用大白话给你说清楚

帆软 发表于商业智能研...

五种大数据框架你必须要知道

云吞铺子

建数据中台，治理烟囱式应用

智领云科技

2020全 路径+教

天才少年

还没有评论

写下你的评论...

▲ 赞同 4 ▼ 添加评论 分享 收藏 ...