

数据中台技术架构方法论与实践



TABLE OF CONTENTS

目录

- 1、建设背景与目标**
- 2、技术架构与思路
- 3、构建过程：
 - 1. PaaS
 - 2. DaaS
 - 3. DA
- 4、未来发展方向
- 5、建设经验总结



1、转转数据中台的背景与目标



1、转转数据中台的背景与目标



流程规范



烟囱模式



孤岛重复



指标重复



时间保障



数据安全



数据共享



形式单一



临时取数



响应及时

- 外部业务：数据脏、乱、差，业务不满意
- 内部研发：疲于奔命、四处救火，普遍苦恼SQL-Boy，人肉提数机
- 方案：数据中台建设
- 目标：复用、赋能、降本提效

TABLE OF CONTENTS

数据中台

- 1、建设背景与目标
- 2、技术架构与思路**
- 3、构建过程：
 1. PaaS
 2. DaaS
 3. DA
- 4、未来发展方向
- 5、建设经验总结



2、转转数据中台技术架构与思路

- 到底什么是数据中台？有什么特点？



2、数据中台技术架构与思路

数据应用 /业务反馈	DA (数据应用层)						服务业务化
数据统计/ 分析/挖掘	BI报表		数据产品		业务系统		应用治理
	渠道分析	商品分析	交易分析	智能挖掘	自助报表	精细化推送	指标字典
DaaS (Data-as-a-Service)							
数据建模 /存储	留存模型主题表	事件模型主题表	画像提取平台	实时自助框架	生命周期管理	质量安全管理	资产服务化
	用户主题	商品主题	交易主题	收入主题	广告主题	行为主题	
PaaS (Platform-as-a-Service)							
数据传输 实时/批量	数据计算层						
	MapReduce	Spark	Storm	Flink	Kylin	Druid	
数据存储层							
数据采集	HDFS	Hive	HBase	MySQL	TiDB	ZZRedis	数据资产化
	Flume	Sqoop	Kafka	Lego	WS	Server	业务数据化

TABLE OF CONTENTS

转转数据中台

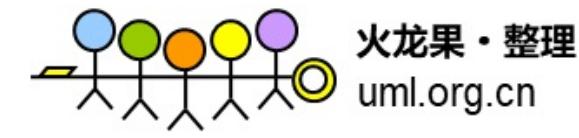
- 1、建设背景与目标
- 2、技术架构与思路
- 3、构建过程：**
 1. PaaS
 2. DaaS
 3. DA
- 4、未来发展方向
- 5、建设经验总结



3、转转数据中台构建过程：PaaS

- zzdp大数据平台
- 目标
 - 高可用、高性能、可扩展的大数据全链路一站式解决方案
- 核心组件/功能
 - Flink、Flume、Kafka、Hadoop、Spark、HBase 等存储计算框架
 - Docker 云平台日志采集系统
 - 苍鹰大数据管理平台
 - Skynet 调度平台

3、转转数据中台构建过程：PaaS



The screenshot shows a PaaS (Platform-as-a-Service) platform interface. At the top, there's a navigation bar with links: 首页 (Home), 数据开发 (Data Development), 元数据 (Metadata), 监控 (Monitoring), 工单 (Work Orders), 数据查询 (Data Query), 系统设置 (System Settings), and 集群管理 (Cluster Management). The '集群管理' dropdown is open, showing sub-options: 概览 (Overview), HDFS, YARN, and 配置管理 (Configuration Management). Below the navigation, there are three large cards: '用户数' (User Count) with a blue icon, '调度任务数' (Scheduling Task Count) with a red icon, and another card partially visible. Underneath these are sections for '任务运行状态' (Task Running Status) and '任务数量统计' (Task Quantity Statistics). A central feature is a circular diagram divided into segments, with the text 'PaaS (Platform-as-a-Service)' overlaid. To the left, there are three vertical tabs: 'Skynet调度平台' (Skynet Scheduling Platform), '苍鹰数据治理平台' (Cangying Data Governance Platform), and 'Lego日志采集平台' (Lego Log Collection Platform). To the right, there are three main layers of services:

- 数据计算层 (Data Computing Layer):** Flink, Spark, Storm, MapReduce, Kylin, Druid.
- 数据存储层 (Data Storage Layer):** HDFS, Kafka, TiDB, HBase, MySQL, ZZRedis.
- 异构数据源 (Heterogeneous Data Sources):** SDK, Docker, Server, DB, Spider, AD.

3、转转数据中台构建过程：PaaS

- 苍鹰大数据管理平台：
 - 为集群提供立体监控、自助化、可视化运维服务，保障高可用
- 核心功能
 - 集群日常使用情况报表统计与跟踪
 - 冷数据压缩、删除、小文件定期自动合并
 - 日常各类自动化运维操作、监控告警
 - 权限管理
 - 资产管理与优化治理：用户/任务/日志/表 总量、增量、异常数TOP
- 效果



10^1

小文件



1%

超长任务数



30%

高峰负载

3、转转数据中台构建过程：PaaS

- Skynet 调度平台
 - 轻量级、可维护、可扩展
 - 与 Hadoop 生态融合
- 核心功能
 - 任务精准时刻调度
 - 依赖方式灵活多样
 - 根据任务自建血缘关系
- 效果



20,000+

任务数



99.99%

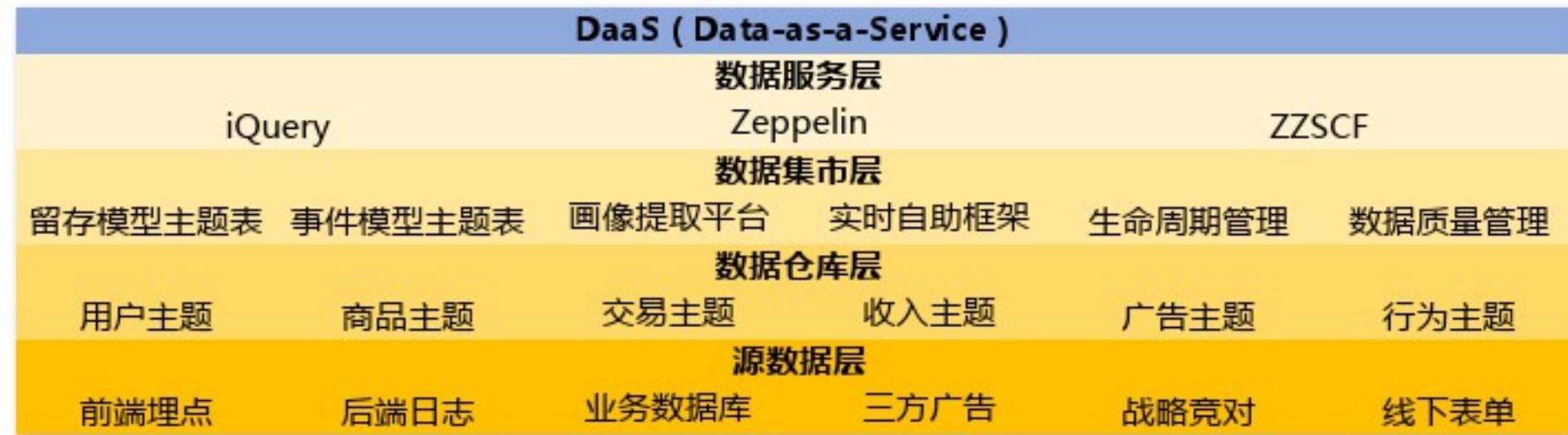
可用性

The screenshot shows the '新建任务' (New Task) interface of the Skynet scheduling system. It includes sections for '调度配置' (Scheduling Configuration) and '依赖配置' (Dependency Configuration). Key settings visible include:

- 调度配置 (Scheduling Configuration):**
 - 调度模式: 二级 (Secondary)
 - 调度次数: 0
 - 周期: 每天 (Daily)
 - 周期执行数: 2
 - 调度策略: 大 (Large)
 - 执行队列: 0
 - 并行数: 0
 - 并发数: 0
 - 调度时间间隔: 0
 - 调度时长: 0
 - 调度时长: 0
- 依赖配置 (Dependency Configuration):**
 - 每页: 10
 - 显示: 10
 - 操作: 插入任务依赖

任务ID	任务名称	任务状态	调度周期	运行时间	负责人	任务状态	操作

3、转转数据中台构建过程：DaaS



- 传统的数仓为何在数据中台地位如此重要？



- 目标：
 - 汇聚全域数据打破数据孤岛，沉淀企业完整 稳定 准确的数据资产
- 核心组件/功能
 - Galaxy 全域数据仓库
 - iQuery 自助式、可视化查询分析平台

3、转转数据中台构建过程：DaaS

➤ Galaxy 全域数据仓库目标：

- 统一的数据建模标准、规范
- 开放的数据存储、建模、计算能力
- 可落地、可扩展，满足转转未来2年，千万日活的业务体量

➤ 数据量



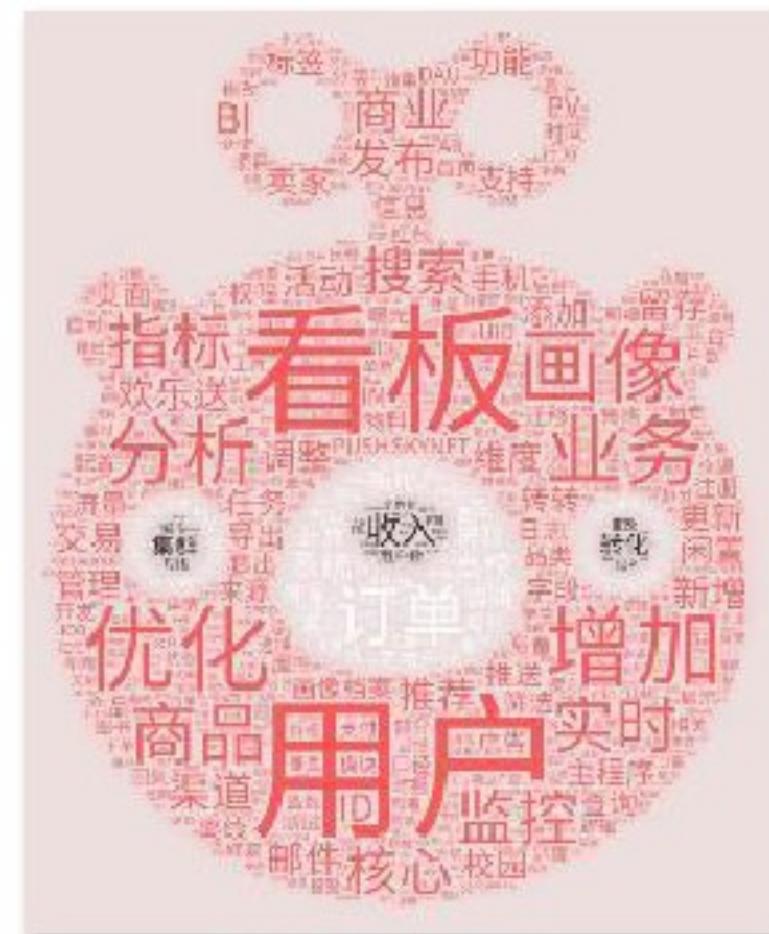
总数据 30PB+



日增量 50TB+

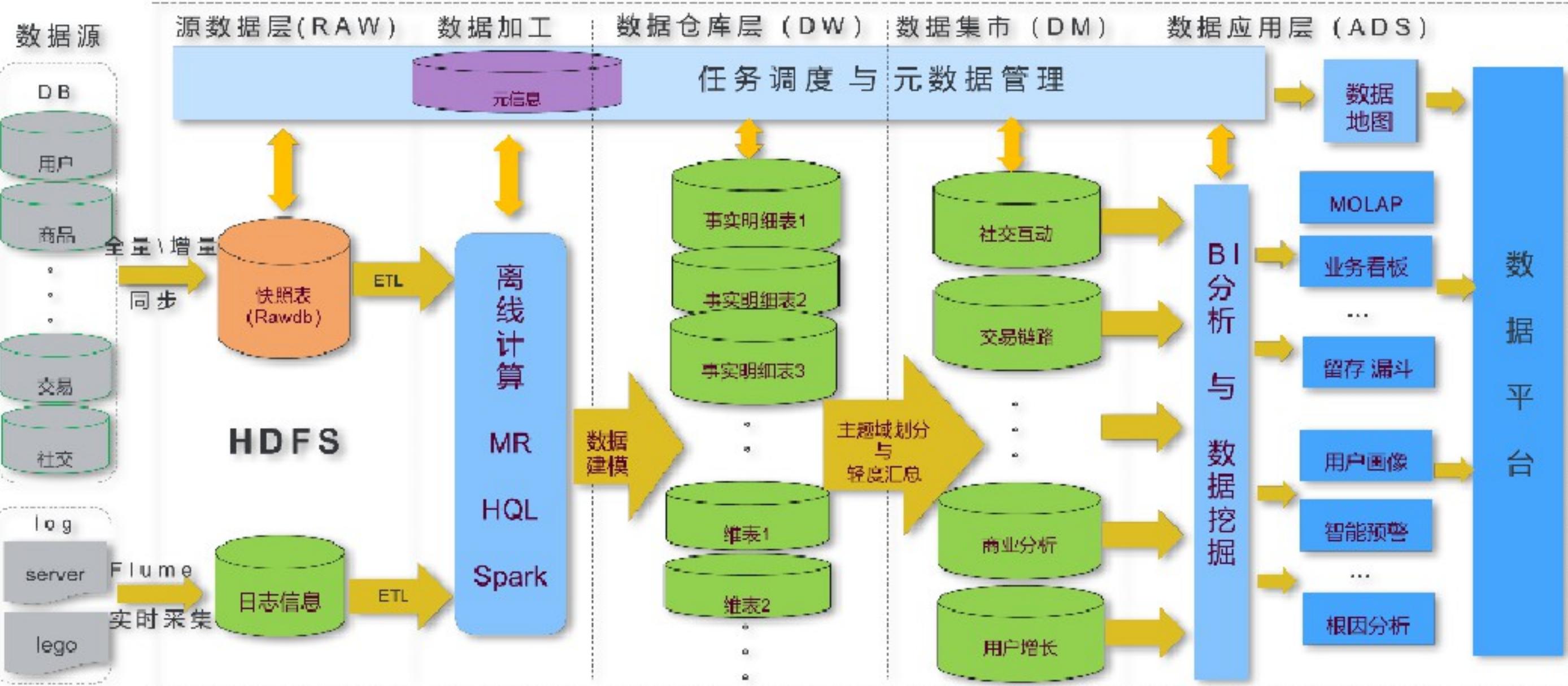


元数据 20,000+



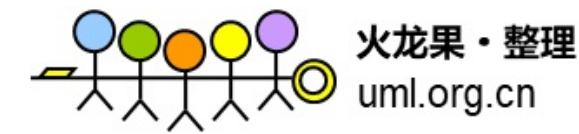
3、转转数据中台构建过程：DaaS

➤ 数据仓库构建之路：Galaxy 全域数据仓库离线整体流程



3、转转数据中台构建过程：DaaS

➤ Galaxy 全域数据仓库效果



时间段： 2015.11-2016.6

2016.6-2017.6

2017.7-至今

模式： 业务支撑

平台研发

业务共建、自治

业务需求： 500+

1400+

600+

业务场景： 分析

+监控+业务输出

+运营+线上服务

覆盖人群： 20%

40%

70%

3、转转数据中台构建过程：DaaS

➤ DaaS 目标

- 支撑数据服务化建设 → 数据价值输出

- 平台、工具、API → 服务化建设

- 面向 PM、运营、RD、分析师等多种角色 → 数据平民化，触达更多人&场景

运行时会使用以下配置，只有手动点击“保存”，才会将这些配置与文档一起储存，下次打开文档，配置仍然存在。

执行引擎 HiveSQL SparkSQL SparkSQL 优化参考 设置新规则默认执行引擎

变量替换 不使用 使用

模板参数 不使用 使用

符号替换 不使用 使用

结果输出到 文件 由插件 MySQL Http-Server Hive

Http Server

Http Server 起始的URL

结果通知 不使用 使用

通知条件(至少勾选一个)

行数成功时 行数失败时 查询结果为空时

接收人配置(至少配置一种类型)

微信消息通知(仅限于一号线上的连接，需要最基础的权限即可通过)。因为公司机房在隧道，公司得支付最高的带宽费用。

邮箱 英文账号分隔多个邮箱地址，每个邮箱地址每天限发送1000封邮件通知

短信 英文账号分隔多个口令名(即以用户名，而不是密码的形式)，每个短信账户每天限发送100条短信通知

推送 英文账号分隔多个手机号码，每个手机号每天限发送5条短信通知

文档附件 不使用 使用

执行方式 手动立即执行 定时执行

文档可见性 私有 公开 设置为“文档默认可见性”

运行脚本

运行全部

暂存脚本

脚本配置

列表

- 获取我有权限的数据库列表
- 获取我有权限的表列表
- 获取某人是审批人的表列表
- 获取某人是设计人的表列表
- 获取表元数据信息
- 批量认领 Hive 表
- 获取我的文档列表
- 获取文档详细信息
- 取消任务
- 删除任务
- 运行 SQL 语句
- 获取任务运行进度
- 获取任务失败原因
- 获取结果输出状态
- 获取任务结果下载地址
- 下载任务结果
- 根据 request_id 获取任务 id
- 获取用户信息

实际传递时，请移除下方示例中的全部注释(#灰色部分)!!!

实际传递时，请移除下面示例中的全部注释(#灰色部分)!!!

获取我有权限的服务器列表

获取我有权限的服务器下拉列表。

位置：[查询统计] -> [SQL统计] -> [左侧第一个 tab “我的仓库”版块]

方法：GET

网址：/servers

示例：http://**127.0.0.1:8080**/exapi/servers?apikey=xxxxxx

结果：

```
{
  "code": 0,
  "data": [
    {
      "id": 24, /* 服务器 id */
      "name": "医疗服务部", /* 服务器名称 */
      "type": "hive" /* 服务器类型：Hive, MySQL */
    }
  ]
}
```

3、转转数据中台构建过程：DaaS

➤ DaaS 落地的关键点-1

- 数仓统一可落地的流程规范，统一认知：

- 层次明确合理：规则、层次、划分、依赖清晰 → 不做不定项选择，质量控制和运维
- 流程机制约束：审批+巡检 → 先污染后治理



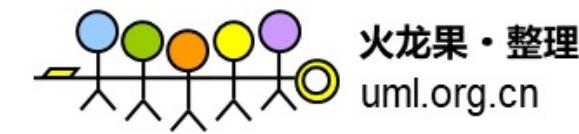
3、转转数据中台构建过程：DaaS

➤ DaaS 落地的关键点-2

- 业务与数据增长，海量数据、报表、标签是服务能力的象征，但会带来哪些问题？
 - 信息过载，数据沼泽 → 负资产
 - 寻找数据、理解数据、信任数据、使用数据 → 矛盾凸显？



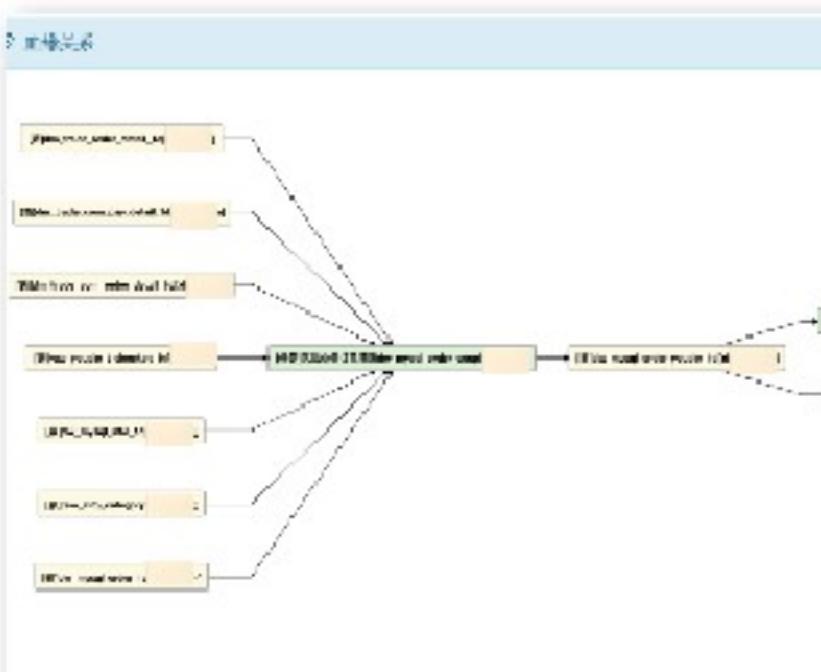
3、转转数据中台构建过程：DaaS



➤ DaaS 落地的关键点-2

- 数据资产管理

- 元数据管理
 - 生命周期管理
 - 性能优化
 - 权限管理



模块功能		功能说明	
序号	功能名称	操作按钮	操作说明
1	新增	新增	添加新客户信息，包括客户基本信息、客户工作和家庭情况。审核通过的数据将显示在客户列表中。
2	修改	修改	修改客户信息，包括客户基本信息、客户工作和家庭情况。审核通过的数据将显示在客户列表中。
3	删除	删除	删除客户信息，包括客户基本信息、客户工作和家庭情况。审核通过的数据将显示在客户列表中。
4	导出	导出	将客户信息导出为Excel文件，方便数据的进一步处理。
5	导入	导入	将Excel文件导入系统，快速添加客户信息。
6	搜索	搜索	根据客户姓名、手机号码或身份证号进行模糊搜索。
7	统计	统计	查看客户总数、客户分布（按地区）、客户年龄分布等统计信息。
8	报告	报告	生成客户分析报告，包含客户画像、客户行为趋势等。

3、转转数据中台构建过程：DA

- DA：转转数据应用层
- 目标
 - 数据业务化，价值输出，形成完整的数据闭环 → 数据能力共享、赋能
- 产品矩阵：
 - What (BI报表、Skyeye、画像)
 - Why (根因分析、Report)
 - How (智能Push、A/B Test 、API...)



3、转转数据中台构建过程：DA

- 数据智能：数据科学之路
- 目标：Hindsight → Insight → Foresight
 - 数据是DT时代的“石油”，但价值需要被进一步的提炼和挖掘
 - 广告投放 / 根因分析 / 智能告警
 - 用户挖掘 / 付费提醒 / 流失预警

