的e﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽上了解的e﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽﷽上了解

**数据资产管理平台**

**数据应用分析报告**

**产品需求文档**

|  |  |  |  |
| --- | --- | --- | --- |
| 作 者 | 九昊 | 初稿时间 | 2016/12/26 |
| 审 核 |  | 终稿时间 |  |
| 审 批 |  | 审批时间 |  |
| 编号 |  | 文档密级 | 内部文档 |

1 修订记录 3

2 现状和问题 3

3 目标和价值 4

4 项目组成员 4

5 产品约定 4

5.1 名词和术语 4

5.2 公共约定与处理 5

5.2.1 适用范围 5

5.2.2 权限管理 5

5.2.3 用户使用流程 5

6 产品功能 5

6.1 数据埋点采集方案 6

6.2 数据应用注册 7

6.3 应用元数据模型 8

6.4 数据应用分析报告 8

6.4.1 应用概览 9

6.4.2 应用质量 11

6.4.3 应用产出 13

6.4.4 应用效能 14

6.5 应用列表 15

6.6 应用模块360 16

6.6.1 链路分析 16

6.6.2 产出分析 17

6.6.3 访问明细 17

6.7 方案风险点 18

6.7.1 产品信息注册 18

6.8 权限控制 18

6.9 其他辅助功能 19

6.9.1 重大事件标识 19

6.9.2 解读可配置化 19

6.10 产品I期项目计划安排 19

6.11 BUG LIST 19

# 修订记录

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 更新时间 | 备注 | 作者 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# 现状和问题

当前已经完成业务源、数据存储、数据计算等元数据的采集，对于数据价值有直接体现的用户端服务或产品元信息的采集缺失，导致资产管理面向数据全链路时环节缺失；集团数据服务形式主要以分析类产品和可视化工具为主，收集用户端元数据以及行为数据，上溯至服务端、ODPS端，形成端端互通意义重大。为上游生产调度、影响分析、稳定性保障 提供参考信息。

# 目标和价值

* 本期实现DT内部数据产品注册开发流程设计及埋点方案
* 完成应用分析报告，为数据产品提供质量监控、问题诊断等功能
* 丰富OneMeta体系，为端端互通和数据ROI提供基础数据参考

# 项目组成员

|  |  |
| --- | --- |
| 角色 | 人员 |
| PD | **九昊** |
|  |  |
|  |  |

# 产品约定

## 名词和术语

核心/非核心模块:以数据资产等级切分，暂定D1、D2等级为核心模块，D3、D4为非核心模块

计算终端链路: 数据应用使用的表节点上溯至无并列分支的节点，最终链路是唯一服务于目标表的。

## 公共约定与处理

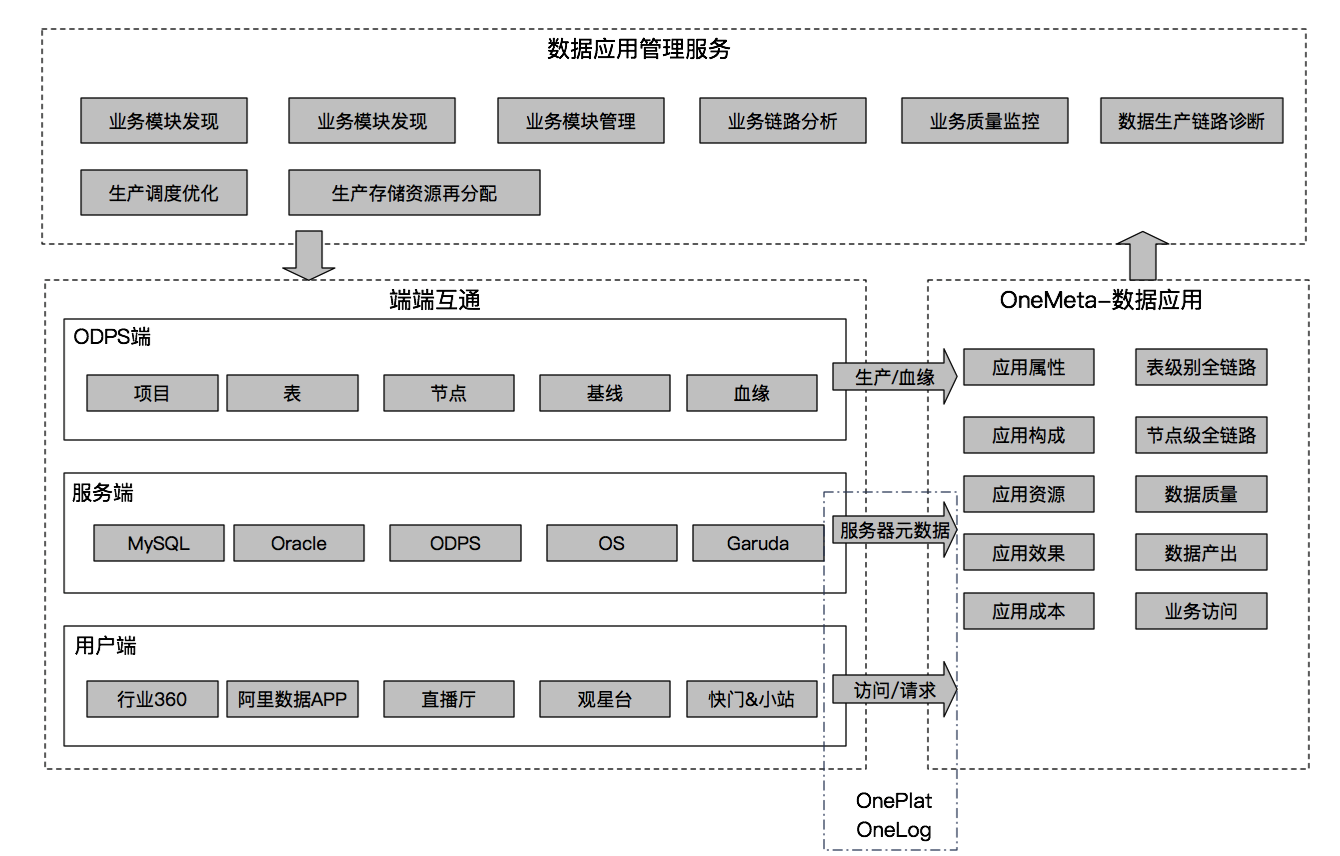
### 适用范围

### 权限管理

### 用户使用流程

# 产品功能

产品架构：



本期项目主要分为四大模块：数据埋点方案、数据应用注册、应用元数据模型、数据应用分析报告。

**数据埋点方案**通过在数据产品前后端引入SDK，采集用户的数据使用行为，上报服务端，沉淀到数据应用访问模型中。**数据应用注册**通过PD手动注册形式，将数据产品元信息丰富到OneMeta元数据体系中。**应用元数据模型**集成应用信息、数据生产、用户属性信息、用户访问行为 等数据以应用为核心的模型。**数据应用分析报告**通过可视化方式，对应用质量、生产、效能等进行监控和诊断，并提供面向问题的排查工具，旨在优化数据质量、提高生产效率和价值ROI。



## 数据埋点采集方案

涉及产品类型：

|  |  |  |
| --- | --- | --- |
| 产品线 | 团队 | 备注 |
| 生意参谋 | 山越团队 |  |
| A+等Apy产品 | 天池团队 |  |
| Freework产品 | 天池团队 |  |
| Java改造产品 | 天池团队 |  |

数据产出：

DT产品数据映射关系-主表

|  |  |
| --- | --- |
| 字段名称 | 字段含义 |
| Aone产品ID | 为每个注册数据数据产品分配一个唯一ID |
| PD负责人 | 产品Owner |
| 开发负责人 | 技术Owner (同一产品不同模块是否可能不同) |
| 产品名称 | 产品全称，如 生意参谋1688版 |
| 功能/页面名称 |  |
| 父级功能/页面名称 |  |
| SpmID | 关联到页面的Spm |
| 访问URL/JSON | 产品形态或交互方式不同，关注的数据访问方式不同 |
| OneService表名 | OS中注册的逻辑表 |
| 链路类型 | 实时或离线 |
| ETL表名 | ？？ |
| AppID |  |
| 用户访问次数 |  |

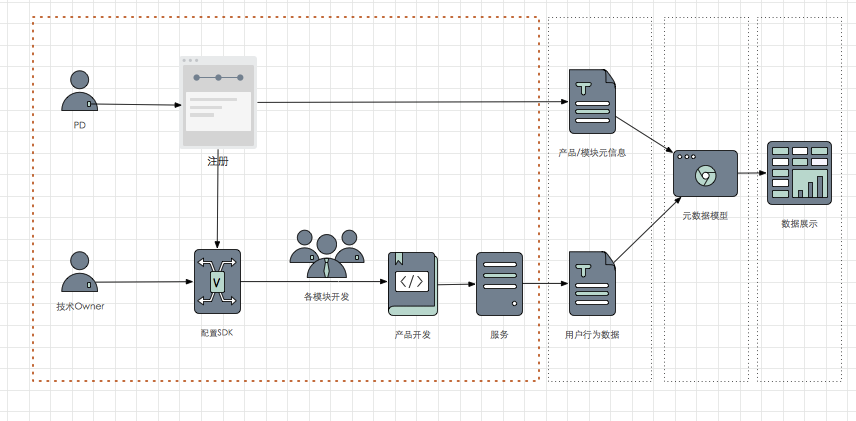
产品数据映射关系-明细表

|  |  |
| --- | --- |
| 字段名称 | 字段含义 |
| 产品名称 |  |
| 功能/页面名称 |  |
| 父级功能/页面名称 |  |
| SpmID |  |
| 访问URL/JSON |  |
| OneService表名 |  |
| 链路类型 |  |
| 访问用户ID |  |
| 鹰眼ID |  |
| 执行SQL |  |
| 备注 |  |

主要采集内容是 用户端与服务端之间的关系 以及 用户端访问行为

## 数据应用注册

数据应用注册是获取产品元信息的基础，基本流程为



|  |  |
| --- | --- |
| **功能块** | **文案or 展示逻辑** |
| PD注册 |  |
| 生成SDK |  |
| 埋点 |  |
| 采集 |  |

备注：由于资源限制，一期不能实现以上流程，而是采用无注册手动分配code直接进行打点的方式将数据反馈，模型端依赖日志对模块信息和模块层级关系进行统计计算，缺少维表。

## 应用元数据模型

参考 @卫成 《端端互通数据模型设计文档.docx》

## 数据应用分析报告

数据应用分析报告展示形式为web页面，分为四个模块：

应用概览: 概括展示应用在质量、产出、效能方面的核心指标，判断应用产品的健康状况

应用质量: 从DQC覆盖、指标预警、质量事件三个方面描述应用数据质量状况

应用产出: 提供模块级产出时间预警以及产出延迟问题诊断

应用效能：关注的应用成本消耗和应用产生价值，分析类数据产品以访问数据评估价值

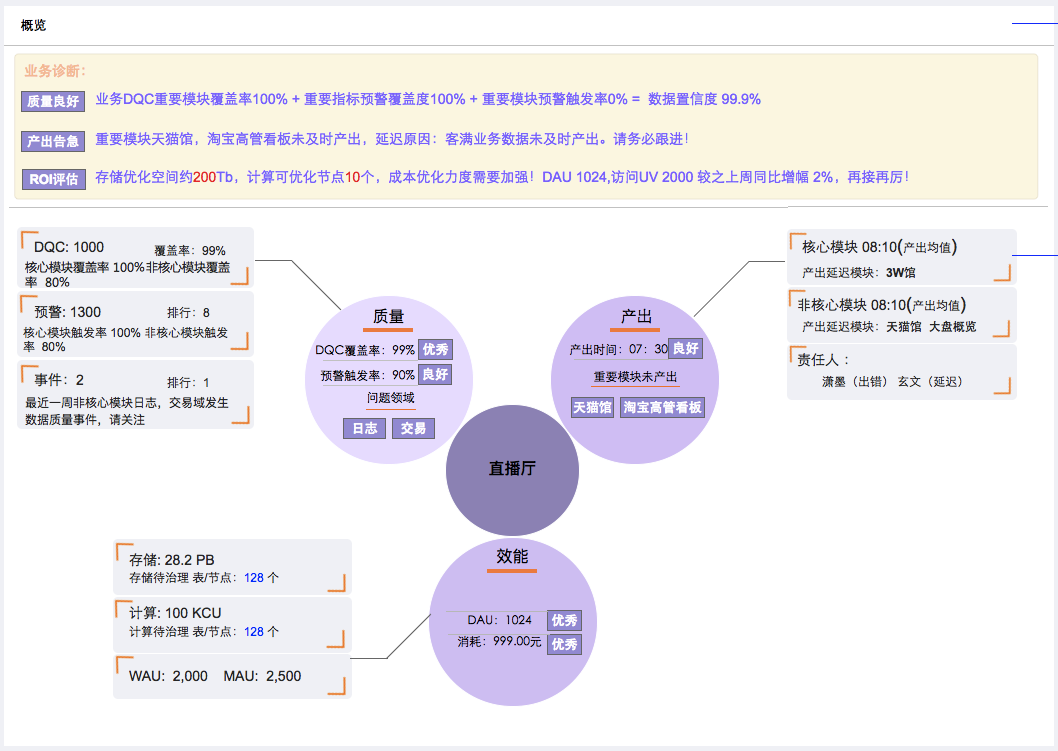
模块筛选方式：分为两级

一级为筛选数据产品名称，需要注意产品名称来源(来源是否是@晋远提供维表，保证产品名称可以统一管理)

二级为模块+页面筛选，二者在下拉列表中有层级区别，但是两个级别都会作为数据汇总的粒度。

### 应用概览

展示样式：



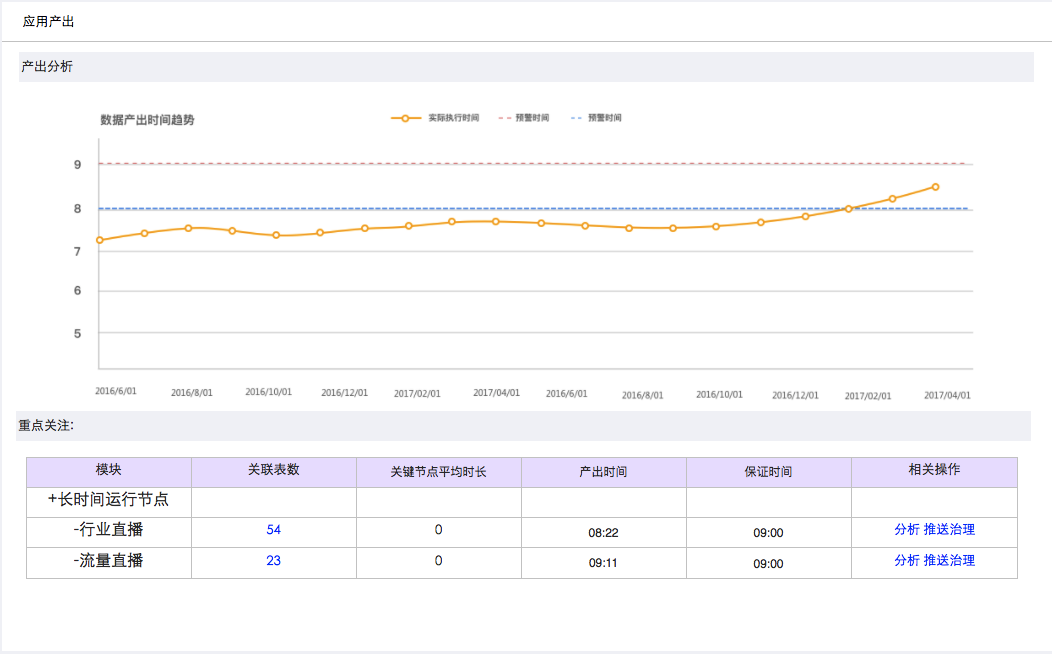
| **功能块** | **文案or 展示逻辑** |
| --- | --- |
| 业务诊断 | 【展现】  业务诊断模块：   1. 标识出数据“质量状况”：优秀、良好、差 (对应规则01) 2. 标识出数据“产出状况”：准时、及时、延迟 (对应规则02) 3. ROI评估：不展示具体ROI数值，展示消耗的成本和应用的访问情况 |
| 应用质量 | 【指标】  DQC配置数  DQC覆盖率  核心/非核心模块覆盖率  预警触发数  核心/非核心模块触发率  预警事件数  问题领域 |
| 应用产出 | 【指标】  产出时间  核心/非核心模块产出时间  产出延迟模块  产出问题责任人 |
| 应用效能 | 【指标】  数据存储  数据计算(终端链路，参考5.1名词和术语)  成本  存储待治理表/节点数  计算待治理表/节点数  UV  WAU  MAU |
| 统计规则 | 1. DQC配置数：产品或者模块关联的末端表上溯全部链路表对应的DQC配置数目 2. DQC覆盖率：产品或者模块关联的末端表上溯全部链路表中，有DQC配置的表占比 3. 核心/非核心模块覆盖率：核心模块关联的表上溯全部链路表中，有DQC配置的表占比（核心模块需要通过手动维护维表进行管理标记） 4. 预警触发数：通过瓦力配置的指标波动预警被出发的数量 5. 核心/非核心模块触发率：核心模块关联的表上溯全部链路表中，预警被触发的指标数占全部数据指标的比 6. 预警：产品或者模块关联的末端表上溯全部链路表，已完成配置的指标数 7. 事件数：通过GOC追踪的评定的质量事件（，存在问题是GOC中存在自动识别和手动录入的事件，手动录入事件无法直接关联到产品，更加无法关联到产品的模块或者页面） 8. 问题领域：产品或模块关联表全链路的治理诊断，归并到数据域上(，存在问题是非公共层数据没有数据域概念，BU内可以归并，如果归并出的问题数据域过多展示相关节点最多的4项 ) 9. 产出时间：末端表导出时间，准实时展示当天数据 10. 核心/非核心模块产出时间：模块关联的末端表最迟产出时间 11. 产出延迟模块：发生模块关联末端表产出延迟的模块 12. 产出问题责任人：产出延迟的模块关联末端表Owner 13. 数据存储：末端链路表对应的全部存储占用 14. 数据计算(末端链路，参考5.1名词和术语)：末端链路节点对应的全部计算量 15. 成本：末端链路计算和存储消耗费用加上数据成本分摊费用之和（待定，末端表数据成本是否包含全部上游成本分摊） 16. 存储待治理表/节点数：模块关联末端表上溯的全部表和节点存储问题待治理的部分 17. 计算待治理表/节点数：模块关联末端表上溯的全部表和节点计算问题待治理的部分 18. UV：昨日(或者当日)访问用户数 19. WAU：近7天访问用户数 20. MAU：近30天访问用户数 |
| 补充：规则 | 规则01:  规则02: 产出准时：为触及数据产出预警线(非基线)；产出及时：触及数据产出预警线，已发布警报但是为触及实际基线；产出延迟：产出时间破线的； |

### 应用质量



| **功能块** | **文案or 展示逻辑** |
| --- | --- |
| 应用质量模块 | 【展现】   1. 分为三个模块：DQC配置、指标预警、质量事件，模块之间以Tab形式切换展示详情信息 2. DQC模块分为：责任人分析、质量分及DQC趋势分析 3. 指标预警模块分为：模块预警分布、模块预警趋势分析 4. 质量事件模块分为：质量事件原因分析、质量事件趋势分析   【指标】  DQC配置：   1. 配置DQC表数 2. 模块关联中表数 3. DQC覆盖率 4. 部门DQC覆盖率均值 5. 责任人分布(待定) 6. 模块质量均分与模块DQC覆盖率趋势(近30天)   指标预警：   1. 触发预警数 2. 配置预警数 3. 预警触发率 4. 部门预警触发率均值 5. 模块指标阈值设置，预警触发值 6. 预警触发次数趋势 (30天)   质量事件：   1. 最近7天质量事件数(故障认定的，数据来源于) 2. 财年内质量时间数 3. 质量事件原因分析：责任人归属分布 和 事件原因分析 4. 质量事件趋势：近30天数据，每日质量事件数 |
| 统计规则 | 1. 配置DQC表数： 模块关联的末端表上溯全部链路表中有配置DQC的表数量 2. 模块关联总表数： 模块关联的末端表上溯全部链路表数量 3. DQC覆盖率：模块关联的末端表上溯全部链路表中有配置DQC的表数占全部表数量的比 4. 部门DQC覆盖率：该应用所属BU中DQC覆盖率 5. 责任人分布(待定)： 6. 模块质量均分：模块关联末端表上溯全部链路表质量分平均分 7. 模块DQC覆盖率趋势(近30天)：模块关联末端表上溯全部链路表中有DQC配置的表占比 8. 触发预警数：模块关联的末端表**上溯**全部链路中，指标波动预警平台当日触发的报警数(预警发出的配置波动预警的指标数) （需要确认数据是准实时还是T-1的） 9. 配置预警数：模块关联的末端表**上溯**全部链路中，已配置指标波动的指标数量 10. 预警触发率：触发预警数/配置预警数 11. 部门预警触发率均值：该应用所属BU预警触发率 12. 模块指标阈值设置，预警触发值：(数据来源瓦力) 13. 预警触发次数趋势 (30天)：模块或者页面关联的触发预警数30天的趋势 14. 最近7天质量事件数(故障认定的，数据来源于) 15. 财年内质量事件数 16. 质量事件原因分析：责任人归属分布 和 事件原因分析 17. 质量事件趋势：近30天数据，每日质量事件数 |
| 重点关注 | 【指标】  ~~指标数:无法提供~~  重点关注列表逻辑需要通过初始化是否重点数据来手动完成，现在没有产品信息注册无法流程化实现  关联表数：当前模块或者页面关联的数据表数量  未配置DQC表数：当前模块或者页面关联的数据表数中未配置DQC的表数量  平均质量分：模块或表关联的表上溯全部链路全部表节点的质量分平均分 |

### 应用产出



| **功能块** | **文案or 展示逻辑** |
| --- | --- |
| 应用产出 | 【展现】   1. 产出分析：模块对应的预警时间、保证时间 和 实际产出时间 日趋势 2. 重点关注：表格形式，展示待产出问题模块，以产出相关的治理项组织   【指标】   1. 模块产出时间趋势 2. 关联表数（可索引到详情） 3. 关键节点平均执行时长：（当前仅仅使用关键节点平均时长不能准确描述问题！） 4. 准实时产出时间 5. 保证产出时间 |
| 统计规则 | 1. 关联表数（可索引到详情） 2. 关键节点平均执行时长：（当前仅仅使用关键节点平均时长不能准确描述问题！） 3. 产出时间：T-1数据，该模块整体的产出时间 4. 保证产出时间：需要手动维护的数据，直接从日志中拿不到 |
| 统计规则 |  |
| 备注 |  |

### 应用效能



| **功能块** | **文案or 展示逻辑** |
| --- | --- |
| 成本消耗分析 | 【展现】   1. 数据成本日趋势，成本是否可以归一为货币形式 2. 趋势展示近30天成本消耗   【指标】   1. 当前模块成本日消耗 |
| 产品模块访问分析 | 【展现】   1. 默认展示产品模块DAU，可选指标为 WAU、MAU、日PV   【指标】   1. DAU 当日访问用户数 2. PV : 当日访问浏览量 3. WAU: 近7天访问用户数(7天去重，包含周末) 4. MAU：近30天访问用户数(30天去重) |
| 重点关注 | 【展现】   1. 展现形式为表格，且根据治理项组织   【指标】   1. 关联表数 2. 最近30天访问UV 3. 最近30天访问PV |

## 应用列表



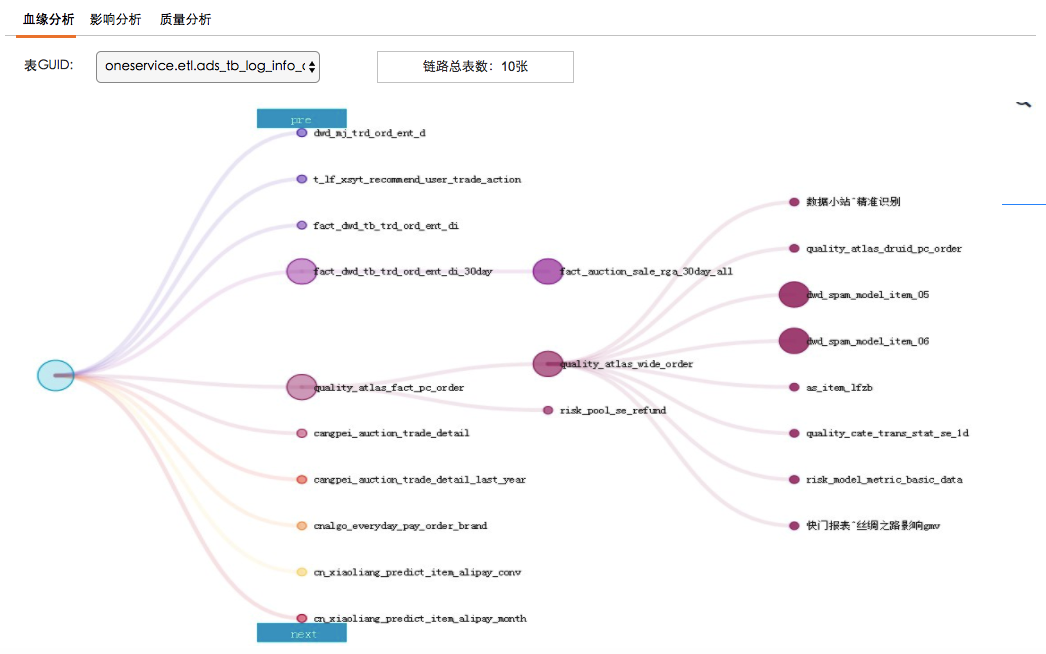
应用列表展示已经在端数据采集中注册的产品信息，包含产品名称、唯一编码、模块名称、页面名称、产品责任人、关联SPM、所属BU等内容，数据按照流程设计来源于6.2 应用注册 部分，但是由于资源问题，一期来源为日志数据中聚合汇总所得。

支持的筛选项为：产品责任人、产品名称、产品编码、所属BU、资产等级（待定）、技术责任人、产品类型、产出要求、关联SPM

## 应用模块360

### 链路分析

链路分析包含三个方面：血缘分析、影响分析、质量分析



血缘分析：

通过产品/模块关联的末端表追溯生产节点或者上游表，用以生产问题或展示数据血缘。

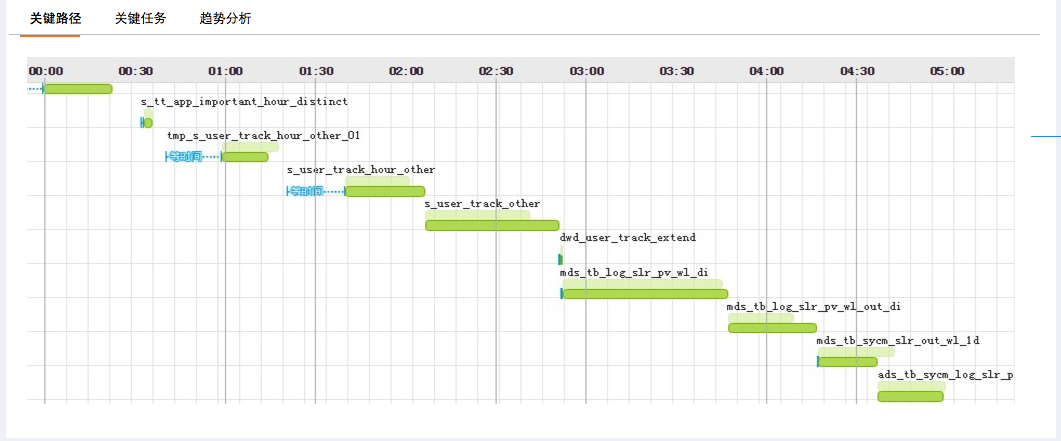
在交互上，固定一个末端表/节点，构建其上游依赖表/节点。根节点为顶部选择的模块或页面。

影响分析：

从字段级血缘入手，依赖字段级血缘，

### 产出分析

产出分析包含三个方面：关键路径分析、关键任务、(产出)趋势分析



**关键路径**指的是模块和页面关联的全部上游链路中数据生产最长路径，链路的构建以 模块或者页面为基础。**关键任务**则是关键路径上的任务节点详情列表，并且通过**任务ID跳转**到任务调度系统进行具体的问题排查和解决。

### 访问明细



表格形式展示 模块/页面 来访用户的行为数据，展示字段有 花名、所属部门、PV、所选日期内访问天数、所属BU.

由于生意参谋主要为外部用户，该部分不包含“生意参谋”。

## 方案风险点

### 产品信息注册

按照完善方案，产品信息是通过PD注册生成包含token秘钥的SDK，并由开发同学在各个模块调用，同时同步信息到产品信息中，注册信息作为维表提供给元数据模型层，保证数据完整性。由于资源不足，当前方案没有PD注册产品信息开发补充模块信息这一步骤，数据主要依靠日志采集聚合获得，可能会导致没有被访问到的模块不能纳入到监控中。

**对于数据质量存在风险。**

01.10更新：确定@晋远 公告平台可以输出产品层次结构。

## 权限控制

ACL权限平台接入，该页面权限设置为 中级

## 其他辅助功能

### 重大事件标识

### 解读可配置化

## 产品I期项目计划安排

## BUG LIST