|  |
| --- |
| 成都丽维家科技有限公司 |
| 数据平台架构 |
| V0.3 |

|  |
| --- |
| James Lei  2019-12-6 |

目录

[1 业务现状 4](#_Toc26543649)

[1.1 分销系统Sales 4](#_Toc26543650)

[1.2 订单管理系统OMS 5](#_Toc26543651)

[1.3 K3 ERP 5](#_Toc26543652)

[1.4 工厂管理系统FMS 5](#_Toc26543653)

[1.5 物流管理系统TMS 6](#_Toc26543654)

[1.6 安装平台安达人 6](#_Toc26543655)

[2 业务系统调研 7](#_Toc26543656)

[2.1 功能模型 7](#_Toc26543657)

[2.2 业务操作 8](#_Toc26543658)

[2.3 数据域 10](#_Toc26543659)

[3 数据平台需求 14](#_Toc26543660)

[3.1 BI功能需求 14](#_Toc26543661)

[3.2 ML功能需求 14](#_Toc26543662)

[4 数据平台功能 15](#_Toc26543663)

[4.1 管理中心 15](#_Toc26543664)

[4.2 规划 15](#_Toc26543665)

[4.3 研发 16](#_Toc26543666)

[4.3.1 开发 17](#_Toc26543667)

[4.3.2 数据处理 19](#_Toc26543668)

[4.3.3 即席查询 19](#_Toc26543669)

[4.3.4 萃取 19](#_Toc26543670)

[4.3.5 发布 20](#_Toc26543671)

[4.3.6 运维 20](#_Toc26543672)

[4.3.7 权限 20](#_Toc26543673)

[4.4 资产 21](#_Toc26543674)

[4.4.1 全景 21](#_Toc26543675)

[4.4.2 地图 21](#_Toc26543676)

[4.4.3 安全 21](#_Toc26543677)

[4.4.4 治理 22](#_Toc26543678)

[4.4.5 服务 22](#_Toc26543679)

[4.5 服务 23](#_Toc26543680)

[4.5.1 服务API 23](#_Toc26543681)

[4.5.2 个人后台 23](#_Toc26543682)

[4.5.3 平台管理 24](#_Toc26543683)

[5 数据平台架构 24](#_Toc26543684)

[5.1 数据同步 25](#_Toc26543685)

[5.2 数仓建模 25](#_Toc26543686)

[5.3 计算 27](#_Toc26543687)

[5.4 数据源 27](#_Toc26543688)

[5.5 数据API 27](#_Toc26543689)

[5.6 数据可视化 27](#_Toc26543690)

[6 数据服务架构 28](#_Toc26543691)

[6.1 数据拉取服务架构 28](#_Toc26543692)

[6.1.1 元数据模块 28](#_Toc26543693)

[6.1.2 主处理模块 28](#_Toc26543694)

[6.1.3 其他模块 29](#_Toc26543695)

[6.2 数据推送服务架构 29](#_Toc26543696)

[7 数据平台开发规划 30](#_Toc26543697)

[7.1 近期规划 30](#_Toc26543698)

[7.2 中期规划 31](#_Toc26543699)

[7.3 远期规划 31](#_Toc26543700)

[8 数据平台部署 31](#_Toc26543701)

[8.1 业务系统数据量 31](#_Toc26543702)

[8.2 业务系统带宽 32](#_Toc26543703)

[8.3 数据平台存储 32](#_Toc26543704)

[8.4 数据平台计算 33](#_Toc26543705)

[8.5 数据平台网络 33](#_Toc26543706)

[8.6 数据平台ECS配置 33](#_Toc26543707)

[9 数据管理开发 34](#_Toc26543708)

[9.1 数据管理 34](#_Toc26543709)

[9.2 数据任务开发 34](#_Toc26543710)

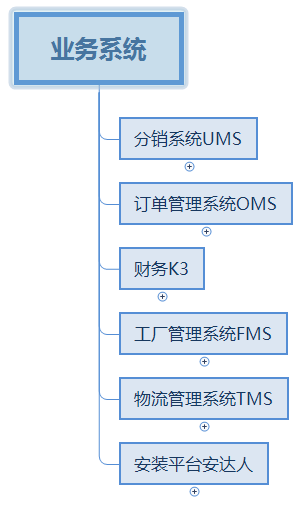
修改记录

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 文件名 | 版本 | 修改内容 | 修改人 | 修改时间 |
| 数据平台架构 | V0.1 |  | James Lei | 2019-8 |
| 数据平台架构 | V0.2 | 添加部署章节 | James Lei | 2019-10 |
| 数据平台架构 | V0.3 | 添加章节4数据平台架构、章节6数据服务架构、章节7数据平台开发规划 | James Lei | 2019-12-6 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# 业务现状

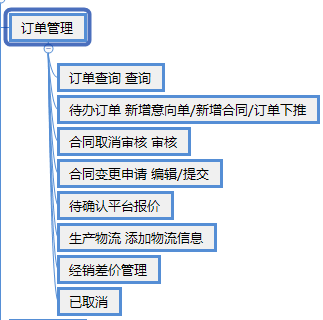
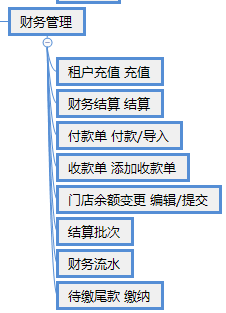
目前有6大业务系统：分销系统，订单管理系统OMS，ERP为金蝶K3，工厂管理系统为联思科技，物流管理系统和安装平台安达人都为自己开发的系统。联思FMS和金蝶K3都为MS SQL Server，其他都为MySQL。

分销系统从小程序、自建等方式得到合同，确认报价后下推订单管理系统，订单管理系统确认付款后下推到K3和OMS系统，工厂管理系统收到订单后拆单，拆单成功将状态同步到OMS和Imos系统，工厂生产后订单入库，完成拣配后工厂发货，订单进入物流管理系统，当实物到达客户现场后，通过安装平台安达人下派安装师傅，完成订单。



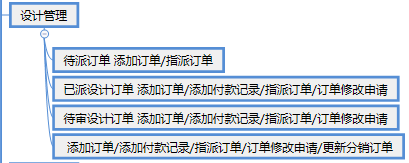
## 分销系统Sales

分销系统主要模块有订单管理、财务管理等，订单管理模块的主要功能有合同审核、下推等操作，财务管理模块的主要功能扣款、结算操作。

## 订单管理系统OMS

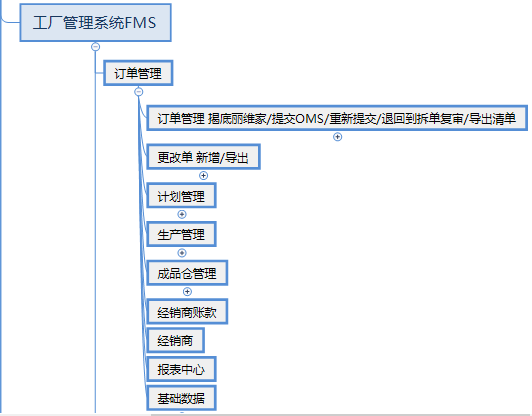
订单管理系统如下图所示，主要由设计管理模块、财务管理模块、拆单管理模块、产品元数据管理模块构成。



## K3 ERP

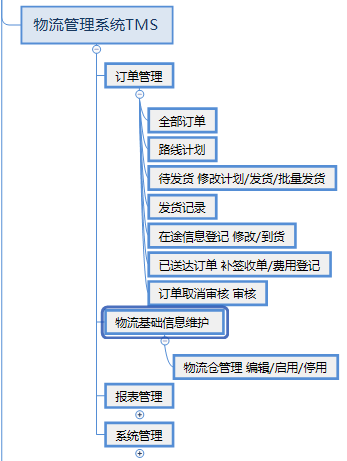
## 工厂管理系统FMS

工厂管理系统主要要订单管理、系统设置两个主模块，主要的功能模块为订单管理、计划管理、生产管理等。



## 物流管理系统TMS

物流管理系统主要功能模块为订单管理。



## 安装平台安达人

安达人主要功能模块为订单管理、售后管理、财务管理、入驻管理、产品管理等。



# 业务系统调研

## 功能模型

粗粒度业务调研表如下：

| 系统 | 功能模型/业务线 | 原材料 | 成品 | 定制 | 软装 | 智能 |
| --- | --- | --- | --- | --- | --- | --- |
| 分销系统Sales | 订单管理 |  | √ | √ | √ | √ |
| 财务管理 |  | √ | √ | √ | √ |
| 营销管理 |  | √ | √ | √ | √ |
| 门店管理 |  | √ | √ | √ | √ |
| 分销市场 |  | √ | √ | √ | √ |
| 小程序 |  | √ |  | √ | √ |
| 订单管理系统OMS | 信息查询 |  | √ | √ | √ | √ |
| 反补订单管理 |  | √ | √ | √ | √ |
| 销售管理 |  | √ | √ | √ | √ |
| 设计管理 |  |  | √ | √ | √ |
| 财务管理 |  | √ | √ | √ | √ |
| 产品元数据管理 |  | √ |  | √ | √ |
| 拆单管理 |  | √ | √ | √ | √ |
| 外购订单管理 |  | √ | √ | √ | √ |
| 供货管理 |  | √ | √ | √ | √ |
| 仓库管理 |  | √ | √ | √ | √ |
| 安装订单管理 |  | √ | √ | √ | √ |
| 用户管理 |  | √ | √ | √ | √ |
| 部门管理 |  | √ | √ | √ | √ |
| 权限管理 |  | √ | √ | √ | √ |
| 订单业务数据管理 |  | √ | √ | √ | √ |
| 订单维护管理 |  | √ | √ | √ | √ |
| 内部资源管理 |  | √ | √ | √ | √ |
| K3 ERP | 对账审计 |  | √ | √ | √ | √ |
| 物料计划 | √ | √ | √ | √ | √ |
|  |  | √ | √ | √ | √ |
| 工厂管理系统FMS | 订单管理 |  | √ | √ | √ | √ |
| 更改单 |  | √ | √ | √ | √ |
| 计划管理 |  | √ | √ | √ | √ |
| 生产管理 |  | √ | √ | √ | √ |
| 成品仓管理 | √ | √ | √ | √ | √ |
| 物流管理系统 | 全部订单 |  | √ | √ | √ | √ |
| 路线计划 |  | √ | √ | √ | √ |
| 待发货 |  | √ | √ | √ | √ |
| 发货记录 |  | √ | √ | √ | √ |
| 在途信息登记 |  | √ | √ | √ | √ |
| 已送达订单 |  | √ | √ | √ | √ |
| 订单取消审核 |  | √ | √ | √ | √ |
| 安装平台安达人 | 订单管理 |  | √ | √ | √ | √ |
| 售后管理 |  | √ | √ | √ | √ |
| 财务管理 |  | √ | √ | √ | √ |
| 入驻管理 |  | √ | √ | √ | √ |
| 用户管理 |  | √ | √ | √ | √ |
| 属性管理 |  | √ | √ | √ | √ |
| 区域管理 |  | √ | √ | √ | √ |
| 产品管理 |  | √ | √ | √ | √ |
| 考试管理 |  | √ | √ | √ | √ |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

## 业务操作

业务过程

| 系统 | 功能模型 | 操作 |
| --- | --- | --- |
| 分销系统Sales | 订单管理 | 查询、新增意向单、新增合同、订单下推、审核、合同变更、确认平台报价、添加物流信息、经销差价管理 |
| 财务管理 | 租户充值、财务结算、付款单、添加收款单、结算批次、财务流水、待缴尾款 |
| 营销管理 | 优惠卷添加审核 |
| 门店管理 | 门店充值 |
| 分销市场 | 加入购物车、支付订单 |
| 小程序 | 入驻审核、提现审核、门店管理、推荐商品 |
| 订单管理系统OMS | 信息查询 | 订单查询、反补订单、客户付款、订单附件、订单流程日志 |
| 返补订单管理 | 补件申请、待品审补、下推设计、追责、返补设计订单、返补拆单管理、返补责任人、返补定责 |
| 销售管理 | 待发布销售订单添加、修改、指派、更新，已发布销售订单添加、补缴定金、添加付款记录、修改、指派，补缴定金管理 |
| 设计管理 | 待派订单添加、指派，已派设计订单添加订单、添加付款记录、指派订单、订单修改，待审设计订单添加、添加付款记录、订单修改 |
| 财务管理 | 客户付款记录审核、提醒，客户付款同步K3，确认账单完结，特殊审核完结账单，账单导入，业务类型管理、支付方式管理，结算方式管理，结算方式管理，银行账号管理，POS终端管理，浙农金服还款记录，退回销售定金 |
| 产品元数据管理 | 分销管理，二级分销管理，供应商管理，计量单位管理，产品属性定义管理，产品类型管理，产品管理 |
| 拆单管理 | 工艺审核，已确认打印，待拆订单指派、退回，待确认决算金额，待核定决算金额，已拆订单指派，已分订单，拆单模板 |
| 外购订单管理 | 待询价订单、回复，待确认订单，已确认订单 |
| 供货管理 | 待回复订单，待确认备货订单，备货中订单，已发货订单 |
| 仓库管理 | 仓库信息，待入库，已入库，待出库，已出库 |
| 安装订单管理 | 待指派安装订单，待安装订单，待验收安装订单，已完成安装订单，返补安装订单 |
| 用户管理 | 普通用户管理，员工用户管理，区域编码管理，缓存管理 |
| K3 ERP | 物流 |  |
| 资金流 |  |
|  |  |
|  |  |
| 工厂管理系统FMS | 订单管理 | 提交OMS、重新提交、退回拆单复审 |
| 更改单 | 新增、导出，物料清单导入、导出，更改明细 |
| 计划管理 | 未排产订单、已排产订单、产能统计，生产批次 |
| 生产管理 | 生产订单，工序跟踪，生产计件 |
| 成品仓管理 | 入库扫描，库位管理，成品管理，发货计划，装车发货 |
| 物流管理系统 | 全部订单 |  |
| 路线计划 |  |
| 待发货 | 修改计划，发货，批量发货 |
| 发货记录 |  |
| 在途信息登记 | 登记、修改、到货 |
| 已送达订单 | 补签收单、费用登记 |
| 订单取消审核 | 审核 |
| 安装平台安达人 | 订单管理 | 新订单创建、重置，取消订单，派单，重置状态、调换人工，补差申请，补件审核，补件无需安装服务、添加补件，补差申请，签收订单，评价订单，待审核订单，待终审订单，取消审核 |
| 售后管理 | 售后工单添加、编辑、回访、售后服务、申请赔偿、商家退款，赔偿审核，退款审核 |
| 财务管理 | 工人结算，商家结算，付款单付款、导入审、审核、付款，收款单添加、审核，结算，资金流水，平台银行账户添加 |
| 入驻管理 | 商家入驻审核，工人入驻审核，项目经理入驻审核，线下辅材店入驻审核，商家管理添加、编辑记账额度、绑定工人、解绑工人、绑定商家、解绑商家、编辑、启用、停用，合伙人添加、绑定商家、解绑商家、绑定工人、解绑工人、编辑启用停用，工人添加、编辑、启用、停用，未入驻工人添加，项目经理添加、编辑、信用，线下辅材店管理添加、编辑、启用、停用 |
| 用户管理 | 个人用户管理编辑、重置密码，员工用户管理同步、编辑、绑定组织机构 |
| 属性管理 | 工人技能属性添加、编辑、启用、停用、删除，数据字典添加、编辑、删除，评论添加、编辑、删除 |
| 区域管理 | 运营区域管理开放运营、关闭运营 |
| 产品管理 | 基础产品添加、启用、停用、删除、修改、下架、删除，产品分类添加、修改删除，区域价格管理，产品单位管理添加、编辑、删除，产品类目超类管理 |
| 考试管理 | 标签管理，试题管理，试卷生成规则管理，成绩管理 |
|  |  |  |
|  |  |  |
|  |  |  |

## 数据域

数据域分类：

| 数据域 | 业种过程举例 | 数据域编码 |
| --- | --- | --- |
| 分销域 |  | sales |
| 订单域 |  | oms |
| 财务域 |  | K3 |
| 生产域 |  | fms |
| 物流域 |  | tms |
| 安装域 |  | adr |
|  |  |  |

| 系统 | 功能模型 | 操作 | 数据域 |
| --- | --- | --- | --- |
| 分销系统DMS | 订单管理 | 查询、新增意向单、新增合同、订单下推、审核、合同变更、确认平台报价、添加物流信息、经销差价管理 | 分销域 |
| 财务管理 | 租户充值、财务结算、付款单、添加收款单、结算批次、财务流水、待缴尾款 | 分销域 |
| 营销管理 | 优惠卷添加审核 | 分销域 |
| 门店管理 | 门店充值 | 分销域 |
| 分销市场 | 加入购物车、支付订单 | 分销域 |
| 小程序 | 入驻审核、提现审核、门店管理、推荐商品 | 分销域 |
| 订单管理系统OMS | 信息查询 | 订单查询、反补订单、客户付款、订单附件、订单流程日志 | 订单域 |
| 返补订单管理 | 补件申请、待品审补、下推设计、追责、返补设计订单、返补拆单管理、返补责任人、返补定责 | 订单域 |
| 销售管理 | 待发布销售订单添加、修改、指派、更新，已发布销售订单添加、补缴定金、添加付款记录、修改、指派，补缴定金管理 | 订单域 |
| 设计管理 | 待派订单添加、指派，已派设计订单添加订单、添加付款记录、指派订单、订单修改，待审设计订单添加、添加付款记录、订单修改 | 订单域 |
|  |  | 订单域 |
| 财务管理 | 客户付款记录审核、提醒，客户付款同步K3，确认账单完结，特殊审核完结账单，账单导入，业务类型管理、支付方式管理，结算方式管理，结算方式管理，银行账号管理，POS终端管理，浙农金服还款记录，退回销售定金 | 订单域 |
| 产品元数据管理 | 分销管理，二级分销管理，供应商管理，计量单位管理，产品属性定义管理，产品类型管理，产品管理 | 订单域 |
| 拆单管理 | 工艺审核，已确认打印，待拆订单指派、退回，待确认决算金额，待核定决算金额，已拆订单指派，已分订单，拆单模板 | 订单域 |
| 外购订单管理 | 待询价订单、回复，待确认订单，已确认订单 | 订单域 |
| 供货管理 | 待回复订单，待确认备货订单，备货中订单，已发货订单 | 订单域 |
| 仓库管理 | 仓库信息，待入库，已入库，待出库，已出库 | 订单域 |
| 安装订单管理 | 待指派安装订单，待安装订单，待验收安装订单，已完成安装订单，返补安装订单 | 订单域 |
| K3 ERP | 物流 |  | 财务域 |
| 资金流 |  | 财务域 |
| 信息流 |  | 财务域 |
| 工厂管理系统FMS | 订单管理 | 提交OMS、重新提交、退回拆单复审 | 生产域 |
| 更改单 | 新增、导出，物料清单导入、导出，更改明细 | 生产域 |
| 计划管理 | 未排产订单、已排产订单、产能统计，生产批次 | 生产域 |
| 生产管理 | 生产订单，工序跟踪，生产计件 | 生产域 |
| 成品仓管理 | 入库扫描，库位管理，成品管理，发货计划，装车发货 | 生产域 |
| 物流管理系统 | 全部订单 |  | 物流 |
| 路线计划 |  | 物流 |
| 待发货 | 修改计划，发货，批量发货 | 物流 |
| 发货记录 |  | 物流域 |
| 在途信息登记 | 登记、修改、到货 | 物流域 |
| 已送达订单 | 补签收单、费用登记 | 物流域 |
| 订单取消审核 | 审核 | 物流域 |
| 装平台安达人 | 订单管理 | 新订单创建、重置，取消订单，派单，重置状态、调换人工，补差申请，补件审核，补件无需安装服务、添加补件，补差申请，签收订单，评价订单，待审核订单，待终审订单，取消审核 | 安装域 |
| 售后管理 | 售后工单添加、编辑、回访、售后服务、申请赔偿、商家退款，赔偿审核，退款审核 | 安装域 |
| 财务管理 | 工人结算，商家结算，付款单付款、导入审、审核、付款，收款单添加、审核，结算，资金流水，平台银行账户添加 | 安装域 |
| 入驻管理 | 商家入驻审核，工人入驻审核，项目经理入驻审核，线下辅材店入驻审核，商家管理添加、编辑记账额度、绑定工人、解绑工人、绑定商家、解绑商家、编辑、启用、停用，合伙人添加、绑定商家、解绑商家、绑定工人、解绑工人、编辑启用停用，工人添加、编辑、启用、停用，未入驻工人添加，项目经理添加、编辑、信用，线下辅材店管理添加、编辑、启用、停用 | 安装域 |
| 用户管理 | 个人用户管理编辑、重置密码，员工用户管理同步、编辑、绑定组织机构 | 安装域 |
| 属性管理 | 工人技能属性添加、编辑、启用、停用、删除，数据字典添加、编辑、删除，评论添加、编辑、删除 | 安装域 |
| 区域管理 | 运营区域管理开放运营、关闭运营 | 安装域 |
| 产品管理 | 基础产品添加、启用、停用、删除、修改、下架、删除，产品分类添加、修改删除，区域价格管理，产品单位管理添加、编辑、删除，产品类目超类管理 | 安装域 |
| 考试管理 | 标签管理，试题管理，试卷生成规则管理，成绩管理 | 安装域 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

订单管理业务过程

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据域 | 一致性 | | | | | |
| 业务过程 | 新增意向单、新增合同 | 订单下推 | 审核、合同变更 | 确认平台报价 | 添加物流信息 | 经销差价管理 |
| 分销域 | √ | √ | √ | √ | √ | √ |
| 订单域 |  |  |  |  |  |  |
| 财务域 |  |  |  |  |  |  |
| 生产域 |  |  |  |  |  |  |
| 物流域 |  |  |  |  |  |  |
| 安装域 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

财务管理业务过程：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据域 | 一致性 | | | | | |
| 业务过程 | 租户充值 | 财务结算 | 付款单、添加收款单 | 结算批次 | 财务流水 | 待缴尾款 |
| 分销域 | √ | √ | √ | √ | √ | √ |
| 订单域 |  |  |  |  |  |  |
| 财务域 |  |  |  |  |  |  |
| 生产域 |  |  |  |  |  |  |
| 物流域 |  |  |  |  |  |  |
| 安装域 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

# 数据平台需求

## BI功能需求

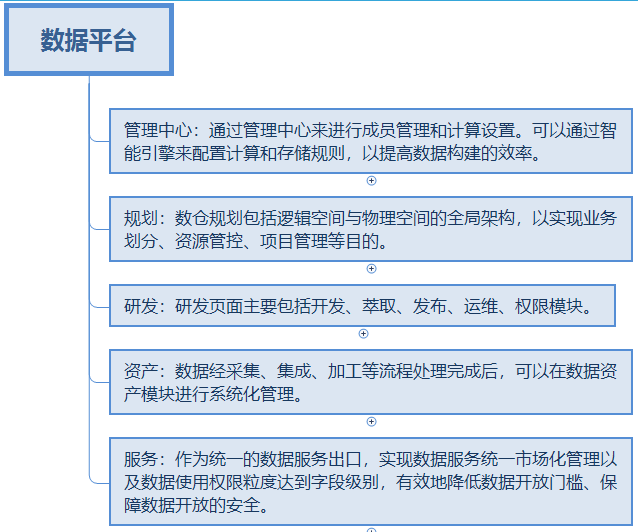
1. 合同（订单）信息统计跟踪：跟踪合同全生命周期中各个阶段状态。
2. 供应链管理，物料需求计划。根据订单统计物料需求，根据历史订单预测。
3. 财务管理，统计各个系统财务数字，核对各个系统财务数字差别。
4. 云仓库数据指标：按年、月统计云仓销售金额、销售数量、物流情况（海运、陆运），显示云仓视频。
5. 云工厂数据指标：按年、月统计各云工厂产值、订单数量，显示云工厂视频。
6. 安达人数据指标：按年、月、区域统计订单金额。

## ML功能需求

自动拆单：学习历史数据，得到拆单模型，将拆单模型应用到实际拆单业务过程中。

板材采购量预测：根据各种相关的历史数据如订单数量，预测最优的板材采购量。

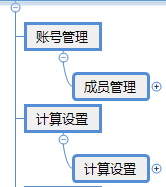
# 数据平台功能



数据平台包括5大部分功能：管理中心、数仓规划、研发、资产、服务，涵盖从资源管理、项目管理，到业务规划、指标定义，到数据治理、资产管理，到数据服务，到数据运维监控报警的全业务过程。

## 管理中心

管理中心用于管理平台的用户和计算资源。

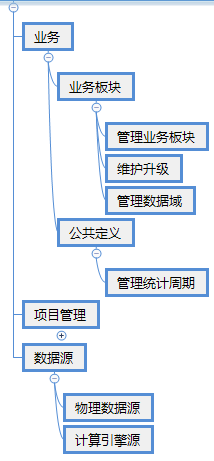


* 账号管事：管理中心的功能对所有系统成员开放展示，但设置、增删操作仅对超级管理员开放。
* 计算设置：计算设置包括选择计算引擎类型、设置集群及元数据。计算引擎类型可选Flink。

## 规划

数仓规划包括逻辑空间与物理空间的全局架构，以实现业务划分、资源管控、项目管理等目的：

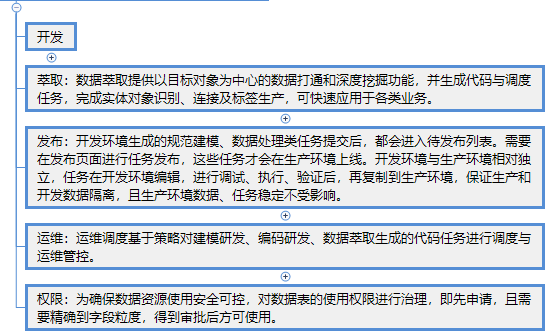
* 逻辑空间：基于业务特征划分命名空间及其核心对象，包括业务板块、数据域、公共定义等模块。
* 物理空间：基于开发协作管理需求划分物理项目，包括项目管理、数据源管理等模块。



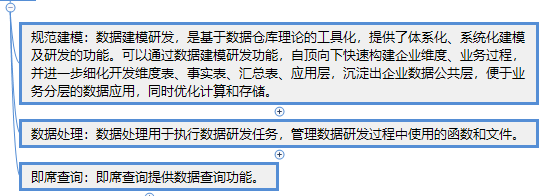
* 业务板块：业务板块是逻辑空间的重要组成部分，是基于业务特征划分的命名空间。在同一个业务板块中可能包含多个不同的项目，业务板块与项目的关系为1：N。
* 数据域：数据域主要用于存放同一业务板块内不同概念的指标。例如，您可以划分出商品域、交易域、会员域等，用于存放不同意义的指标。
* 公共定义：公共定义用于定义系统级概念对象，以保证全局概念统一、通用化引用。公共定义包括定义统计周期，省市区等。统计周期即统计的时间范围，例如最近7天、最近30天等（类似于SQL中where后的时间条件）。
* 项目管理：项目是一种物理空间上的划分，便于用户在数据中台建设过程中对物理资源及开发人员进行隔离化管理。一个业务板块可以包含多个项目，每个系统成员可以加入多个不同的项目。
* 数据源：数据源分为物理数据源和计算引擎源，其中物理数据源：存储数据的物理数据库即物理数据源。物理数据源通常是数据建设的数据来源或基础，用于数据引入的数据同步环节。计算引擎源是具备计算能力的数据库。

## 研发

研发主要包括开发、萃取、发布、运维、权限模块。

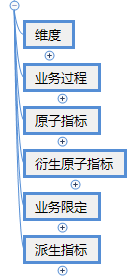


### 开发



* 规范建模：包含规范定义、逻辑表、数据开发、建模引擎4部分。
* 数据处理：包括代码任务、同步任务、函数管理、资源管理。
* 即席查询：逻辑表模型以业务板块为命名空间，所以针对逻辑表，引用逻辑表的方式为：业务板块.逻辑表。

#### 规范定义



* 维度：维度是人们观察事物的角度，是指一种视角，是确定事物的多方位、多角度、多层次的条件和概念。维度即进行数据统计的对象。通常情况下，维度是实际存在、不因事件发生而存在的实体，例如时间维度、地区维度、产品维度等。创建维度，即从顶层规范业务中实体（或称主数据）的存在性及唯一性。
* 业务过程：业务过程即业务活动中所有的事件，为不可拆分事件。创建业务过程，即从顶层视角，规范业务活动中事件的内容类型及唯一性。
* 原子指标：原子指标是对指标统计口径、具体算法的抽象。设计即开发的理念，指标定义同时也明确了设计统计口径（即计算逻辑），提升了研发效率，并保证了统计结果的一致性。
* 衍生原子指标：
* 业务限定：原子指标是计算逻辑的标准化定义，业务限定则是条件限制的标准化定义。为保障所有统计指标统一、标准、规范地构建，业务限定在业务板块内唯一，并唯一归属于一个来源逻辑表，计算逻辑也以该来源逻辑表模型的字段为基础进行定义。由于一个业务限定的定义可能来自于归属不同数据域的多个逻辑表，因此一个业务限定可能归属于多个数据域
* 派生指标：派生指标即常见的统计指标，派生指标=原子指标+业务限定+统计周期+统计粒度。为保证统计指标标准、规范、无二义性地生成，将派生指标抽象为如下四部分：

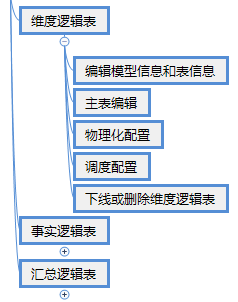
原子指标：明确统计口径，即计算逻辑。

业务限定：统计的业务范围，筛选出符合业务规则的记录（类似于SQL中Where后面的条件，不包括时间区间）。

统计周期：统计的时间范围，例如，最近一天、最近30天等（类似于SQL中Where后面的时间条件）。

统计粒度：统计分析的对象或视角，定义数据需要汇总的程度，可理解为聚合运算时的分组条件（类似于SQL中的Group By的对象）。粒度是维度的一个组合，例如某个指标是某个卖家在某个省份的成交额，则粒度就是卖家、地区这两个维度的组合。基于上述定义组合机制，即可批量、无重复地快速创建派生指标。同时，可以保证命名和计算逻辑明确且不重复，业务人员也可以自主定义生产。派生指标作为与字段同级的概念，在同一个统计粒度内唯一。

#### 逻辑表



* 维度逻辑表：维度逻辑表与维度一一对应，是通过丰富维度中的属性信息构建而成的。维度发布成功后，系统会自动创建对应的维度逻辑表，并支持对维度逻辑表进行添加属性、关联维度、添加子维度、物理化配置等操作。编辑维度逻辑表的核心目标是丰富维度之上的相关属性信息。需要注意的是：抽象提炼维度自身具备且相对持久的属性，无需冗余非维度关联的对象信息。子维度中不需要重复定义已发布的父维度中相同属性字段，系统会自行引用。
* 事实逻辑表：归属于某个业务过程的事实逻辑表，可以描述具体业务过程所对应事务的详细信息。创建事实逻辑表即完成公共的事务明细数据沉淀，以便提取业务中事务相关的明细数据。
* 汇总逻辑表：汇总逻辑表是由一个特定的分析对象（如会员）及其相关的统计指标组成的。组成一个汇总逻辑表的统计指标都具有相同的统计粒度（如会员）。汇总逻辑表面向用户提供了以统计粒度（如会员）为主题的所有统计数据（如会员主题集市），是支持的主题式数据查询工具。组成汇总逻辑表的统计指标有两种来源：系统按照相同统计粒度，自动汇聚。通过非派生指标创建方式创建的普通物理表字段，根据同主键原则归一至相应的汇总逻辑表。

#### 建模引擎

逻辑表开发、提交和调度运行，会自行根据系统内置规则对逻辑表模型进行校验、物理化，提供查询使用。标准规范、可靠稳定及高效的逻辑表研发过程依赖于智能建模引擎服务。

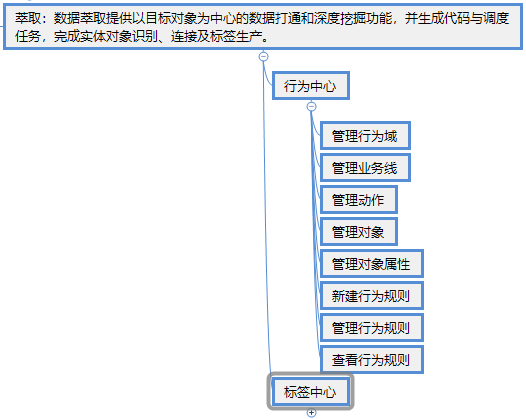
### 数据处理

数据处理用于执行数据研发任务，管理数据研发过程中使用的函数和文件。数据处理包含任务管理和文件管理两个模块：任务管理用于执行数据研发任务，包括代码任务和同步任务。文件管理用于管理数据研发过程中使用的文件，包括函数管理和资源管理。

### 即席查询

即席查询提供数据查询功能。计算引擎类型下支持SQL。系统会根据您所设置的计算类型自动识别切换SQL类型。

### 萃取



* 行为中心：行为元素管理功能提供行为数据的标准化定义与归类。行为元素分为行为域、业务线、动作、对象、对象属性，可以对行为元素进行归类管理。

行为域：聚合业务含义一致的行为数据，如电商域、文娱域。

业务线：基于行为域将行为数据进一步细分，各业务线之间相对独立。

动作：行为主体发出的操作，如购买、浏览。

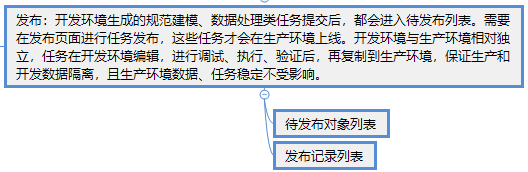
对象：行为主体操作的具体事物，如商品、电影。

对象属性：对象的描述性信息，如名称、品牌、年份。

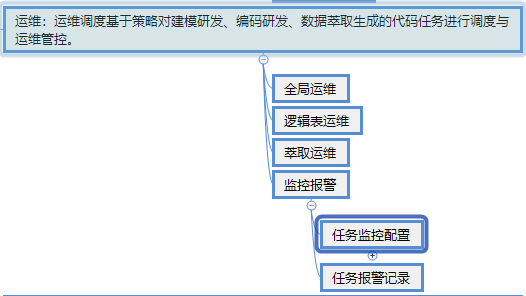
* 标签中心：支持标签可视化配置并自动生成标签调度任务，对不同开发状态的标签任务进行管理。

### 发布

开发环境生成的规范建模、数据处理类任务提交后，只有发布成功，任务才会同步到生成环境进行调度。



### 运维

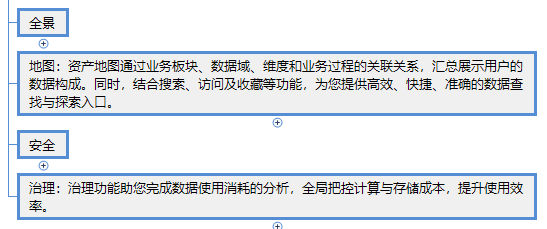


* 全局运维：可以分为周期任务、周期实例、手动任务、手动实例，手动补数实例。
* 逻辑表运维：可以分为逻辑表任务、逻辑表实例。
* 萃取运维：标签补数据、编辑标签和查看标签实例。
* 监控报警：可以分为任务监控配置、任务报警记录。

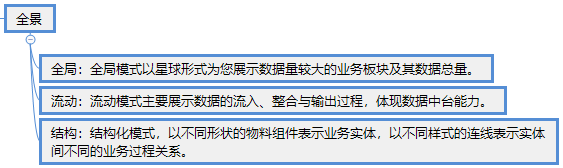
### 权限



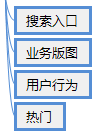
## 资产



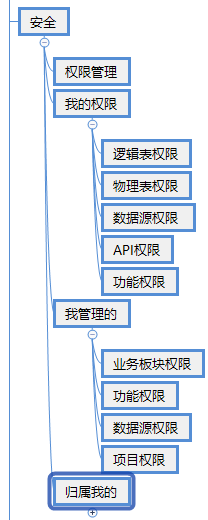
### 全景



### 地图



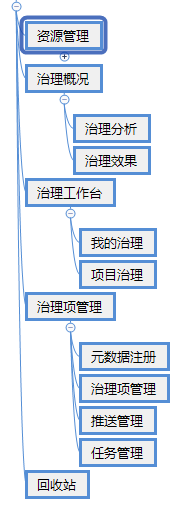
### 安全



* 权限管理：为确保数据使用安全可控，使用数据表需要先申请权限。
* 我的权限：包括逻辑表权限、物理表权限、数据源权限、API权限、功能权限。
* 我管理的：包括业务板块权限、功能权限、数据源权限、项目权限。

### 治理

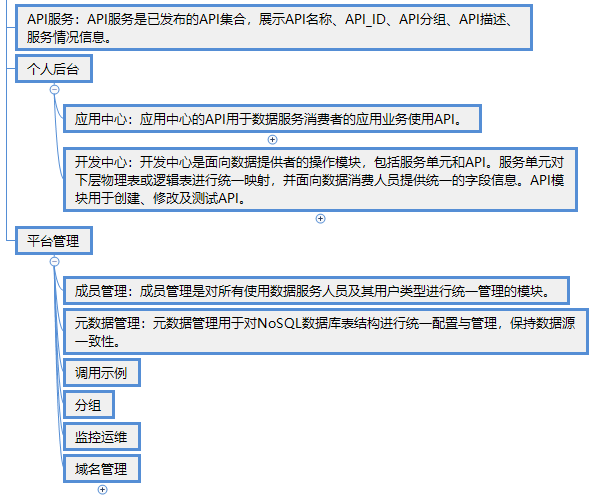
治理功能助您完成数据使用消耗的分析，全局把控计算与存储成本，提升使用效率。治理包括资源管理、治理概述、治理工作台、治理项管理、回收站：



* 资源管理：主要用于资源分析。资源分析是从全局视角为您展示资源消耗、资源消耗增速、资源消耗分布及资源治理概况。
* 治理概览：由治理分析和治理效果组成。治理分析针对全局进行问题诊断、待治理问题点的分析。治理效果助您分析治理后效果，评定治理情况、推动治理优化。
* 治理工作台：由我的治理和治理项目组成。我的治理面向当前登录用户，根据其所参与和所负责的项目，为您提供高效快捷的治理入口和能力。项目治理面向当前登录用户，根据其所参与和所负责项目的信息进行管理。
* 治理项管理：由元数据注册、治理项管理、推送管理、任务管理组成。可以自定义创建治理项、发起推送、查看任务执行，也可以使用系统内置的通用治理项。
* 回收站：在治理工作台中删除或下线的表格暂时存留起来，以防止对数据的误操作。

### 服务

数据服务（OneService），是基于Dataphin建设数据中台的最后一步，作为统一的数据服务出口，实现数据服务统一市场化管理以及数据使用权限粒度达到字段级别，有效地降低数据开放门槛、保障数据开放的安全。数据服务由 API服务、 个人后台、 平台管理组成。



* API服务：操作包括查看API信息、申请API权限。
* 个人后台：包括应用中心、开发中心个人过程中的全过程。开发中心基于数据服务提供者的开发业务场景需求，分为服务单元和API模块。应用中心基于数据服务消费者的应用业务场景需求，分为API、应用和应用分组模块。
* 平台管理：包括成员管理、元数据管理、调用示例、分组、监控运维、域名管理。实现统一数据服务监控及管理。

## 服务

### 服务API

* 查看API信息：
* 申请API权限：

### 个人后台

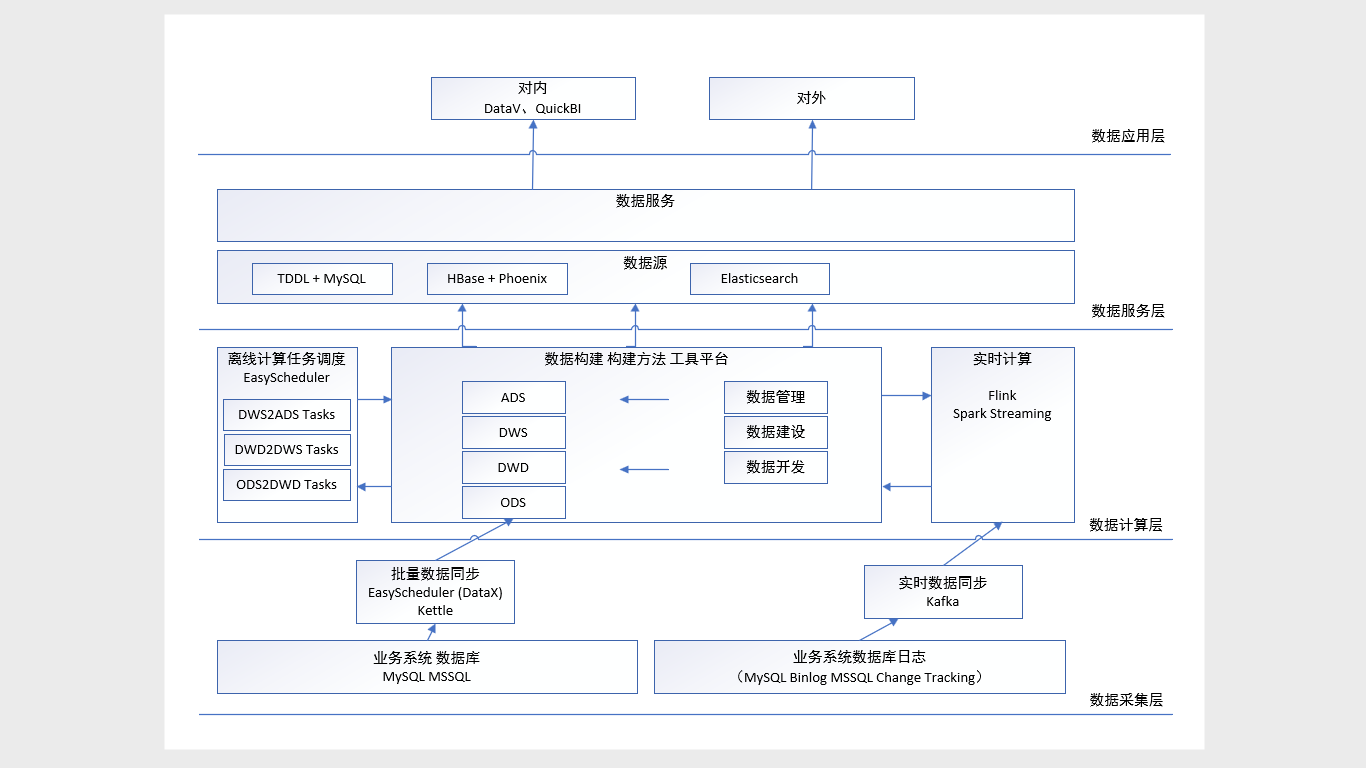
* 开发中心：服务单元可以为单个或多个同源/异构物理表，以及逻辑表（支持单物理表、多张同源/异构物理表）。同一个服务单元可以对接多个API，即服务单元与API的关系是1：N。管理API：模板向导模式下修改API、自定义SQL模式下修改API、测试API、发布API、下线API、删除API、转让API管理员。
* 应用中心：管理API：查看API、申请API、编辑API权限、解除API权限、调试API。管理分组：分组是指对应用进行分组管理。管理应用：应用是数据服务消费者应用业务使用API的主体，同一个应用可以对接多个API，即应用与API的关系是1：N。管理应用成员。

### 平台管理

* 成员管理：对所有使用数据服务人员及其用户类型进行统一管理的模块。
* 元数据管理：用于对NoSQL数据库表结构进行统一配置与管理，保持数据源一致性。
* 调用示例：调用SDK指南。
* 分组：用于对服务单元进行分组管理。
* 监控运维：用于对当天生产环境的API运行情况进行监控，包括平均响应时长、调用次数、错误率、offline百分比。
* 域名管理：域名管理分为二级域名管理和独立域名管理。

# 数据平台架构

数据平台覆盖从数据ETL到最终可视化的全部流程。数据可以分为实时数据与历史数据。同步支持从业务数据库及业务服务同步数据抽取日志，可以使用Datax等实现。历史数据抽取后存入数据仓库，数据仓库分为4层：ODS、DWD、DWS、ADS，ODS为原始数据存储，从业务系统同步过来的数据不做任何处理存入ODS表，ODS数据经过清洗转换（ET，ID-mapping）等存入DWD数据仓库明细表，ET等计算采用开源计算框架或者python，DWD数据经过聚合计算计算不同维度数据形成DWS数据仓库汇总表，DWS再经过计算形成ADS应用数据存储，所有计算均采用计算框架实现。实时数据抽取自业务系统实时信息，采集Flink流式计算不落盘。数据分析师通过开发平台可以在线或者离线开发，一些简单的SQL可以使用Zeppelin和Jupter Notebook开发，开发一些机器学习代码需要用到PyChram，离线开发测试完成后提交到平台部署。各层数仓库和实时数据通过RESTful对外提供API，前端可视化一般可以使用各种自助式BI工具，对于一般BI不能实现的效果可以使用Antd、EChart等控件开发前端应用。



## 数据同步

业务系统数据库可以采集各种数据同步工具进行同步。批量数据同步有两个工具可以选择，Datax为阿里开源的异构数据库批量同步工具，支持主流关系型数据库，用来做不再更新的历史数据同步，开始时可以通过命令行手动进行全量同步，之后通过条件过虑只同步前天的增量数量，通过json配置同步任务，通过命令行调用，工具本身没有调度引擎，可以采集EasyScheduler分布式调度引擎执行DataX的数据同步任务。还可以使用kettle工具，本身有作业调度功能和丰富的工具可以方便的完成数据同步。具体选用哪一个工具，可以根据需求确定。

通过解析业务系统日志（如MySQL Binlog、MSSQL CDC）可以实时同步数据，将数据解析后推到Kafak消息队列，通过各种connect可以方便的连接多种类型数据库。

当然，阿里云也提供BDS服务用于RDS MySQL到HBase的数据同步。由于RDS MSSQL CDC需要付费才能，可以使用Change Tracking功能自己开发Kafka connector。在ESC在自建SQLServer服务，可以开启CDC功能后，通过Kafka connector连接将数据同步到HBase。

## 数仓建模

数据仓库按业务需要可以分为离线数据仓库和实时数据仓库，如架构图所示，数仓离线数分为仓库和实时数据仓库都可以分为4层，ODS、DWD、DWS、ADS。实时数据仓库采用HBase做为存储引擎。离线数据仓库ODS和DWD需要存储大量的数据，也可以使用HBase做为存储引擎，外加Phoenix做为查询引擎，离线数据仓库DWS使用MySQL做为存储查询引擎，离线数据仓库ADS使用Elasticsearch做为存储和查询引擎。

ODS（Operational Data Store）为操作数据存储层，一般为从业务系统同步过来的原始数据，不在同步过程中做ETL，数仓保留一份业务系统数据副本，能够最大限度保留信息，数据平台可以根据后续需求再进行ETL。由于离线数据仓库ODS层集成了6大异构系统的数据，采用HBase做为核心的存储库，而实时数据仓库数量量较小，采用MySQL完全满足存储需求。

DWD（Data Warehouse Detail）为数仓库明细层，ODS数据结合DIM经过ETL计算可以生成粒度最细的明细表，包括事实表和维度表。DWD数据量基本和ODS层保持一致，也需要采用HBase做为存储。

DWS（Data Warehouse Summary）为数仓汇总层，DWD经过聚合计算可以生成汇总表。汇总层需要支持前端BI工具的灵活多变的查询需求，采集MySQL可以满足需求。

DIM（Dimension）为公共维度层，数据可以从ODS层经过处理，生成统一唯一的公共维度表，如地址省市区等。

ADS（Application Data Store）为应用数据服务层，个性化维度汇总层，对于不是特别通用的统计维度数据会放在这一层中，这里计算只有自身业务才会关注的维度和指标。

整个数据仓库采集统一的表名编码格式：

<数仓编号>\_<数据层编码>\_<数据域编码>\_<原始表名>

具体编码如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数仓名 | 数仓编码 | 层名 | 层编码 | 数据域名 | 数据域编码 | 表名 | 表编码 | 表全编码 |
| 离线数仓 | o | ODS | ods | 分销域 | sales | 门店表 | t\_sales\_distributor | o\_ods\_sales\_ t\_sales\_distributor |
| DWD | dwd | 分销域 | sales | 门店表 | t\_sales\_distributor | o\_dwd\_sales\_ t\_sales\_distributor |
| DWS | dws | 分销域 | sales | 门店表 | t\_sales\_distributor | o\_dws\_sales\_ t\_sales\_distributor |
| ADS | ads | 分销域 | sales | 门店表 | t\_sales\_distributor | o\_ads\_sales\_ t\_sales\_distributor |
| 离线数仓 | o | ODS | ods | 生产域 | fms | 订单表 | c\_order | o\_ods\_fms\_ t\_sales\_distributor |
| DWD | dwd | 生产域 | fms | 订单表 | c\_order | o\_dwd\_fms\_ t\_sales\_distributor |
| DWS | dws | 生产域 | fms | 订单表 | c\_order | o\_dws\_fms\_ t\_sales\_distributor |
| ADS | ads | 生产域 | fms | 订单表 | c\_order | o\_ads\_fms\_ t\_sales\_distributor |
| 实时数仓 | r | ODS | ods | 分销域 | sales | 门店表 | t\_sales\_distributor | r\_ods\_sales\_ t\_sales\_distributor |
| DWD | dwd | 分销域 | sales | 门店表 | t\_sales\_distributor | r\_dwd\_sales\_ t\_sales\_distributor |
| DWS | dws | 分销域 | sales | 门店表 | t\_sales\_distributor | r\_dws\_sales\_ t\_sales\_distributor |
| ADS | ads | 分销域 | sales | 门店表 | t\_sales\_distributor | r\_ads\_sales\_ t\_sales\_distributor |
| 实时数仓 | r | ODS | ods | 分销域 | fms | 订单表 | c\_order | r\_ods\_fms\_ t\_sales\_distributor |
| DWD | dwd | 分销域 | fms | 订单表 | c\_order | r\_dwd\_fms\_ t\_sales\_distributor |
| DWS | dws | 分销域 | fms | 订单表 | c\_order | r\_dws\_fms\_ t\_sales\_distributor |
| ADS | ads | 分销域 | fms | 订单表 | c\_order | r\_ads\_fms\_ t\_sales\_distributor |

## 计算

数仓之间的计算可以采集开源计算框架，框架本身提供了一些基础的功能如任务调度等，开发计算任务只需要关注业务逻辑实现。基于实际的业务需求，可以采用Python、shell、SparkSQL等各种计算任务。

## 数据源

数据源这一层可以采用TDDL将各种异构数据库物理表映射为逻辑表。

## 数据API

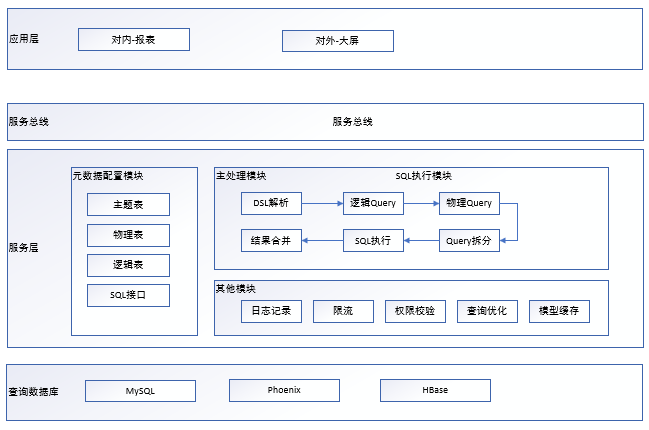
为数据安全和应用调用数据服务的便利，在API层使用网关和身份和权限认证。可视化和其他需要调用数据服务的应用可以通过REST API接口方便集成。

## 数据可视化

大屏展示可以使用阿里DataV，可以很方便的连接到数据库或者API。对公司内部则可以使用自助式的工具Quick BI，方便快捷展示数据。

# 数据服务架构

## 数据拉取服务架构



### 元数据模块

元数据模型，简单来说，就是逻辑表到物理表的映射。

* 数据源：支持跨数据源查询，底层支持接人多种数据源，比如MySQL、HBase等。
* 物理表：物理表是具体某个数据源中的一张表。每张物理表都需要指明主键由哪些列组成，主键确定后即可得知该表的统计粒度。
* 逻辑表：逻辑表可以理解为数据库中的视图，是一张虚拟表，也可以看作是由若干主键相同的物理表构成的大宽表。对用户展现的只是逻辑表，从而屏蔽了底层物理表的存储细节。
* 主题：逻辑表一般会挂载在某个主题下，以便进行管理与查找。

### 主处理模块

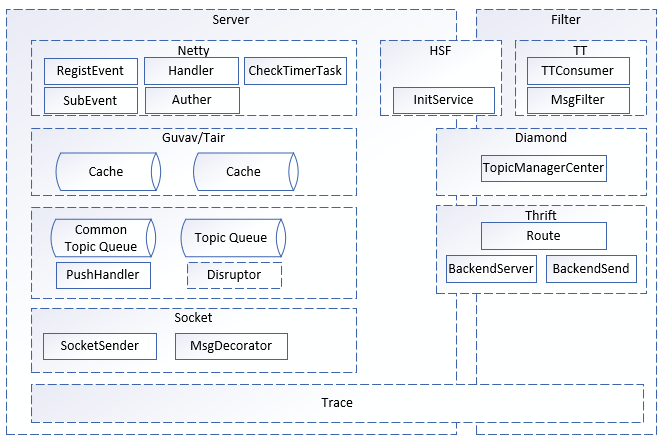
一次查询从开始到结果返回， 一般会经过如下几步。

* DSL解析：对用户的查询DSL 进行语法解析，构建完整的查询树。
* 逻辑Query构建：遍历查询树，通过查找元数据模型，转变为逻辑Query 。
* 物理Query构建：通过查找元数据模型中的逻辑表与物理表的映射关系，将逻辑Query 转变为物理Query。
* Query 拆分：如果该次查询涉及多张物理表，并且在该查询场景下允许拆分，则将Query 拆分为多个SubQuery。
* SQL 执行：将拆分后的SubQuery组装成SQL 语句，交给对应的DB 执行。
* 结果合并：将DB 执行的返回结果进行合井，返回给调用者。

### 其他模块

* 限流：限流有很多种方法，我们采用的是应用内的QP S 保护。针对调用者以及数据源等关键角色做了QPS 阔值控制。也就是说，如果某个调用者的调用量突增，或者对某个数据源的查询流量突增，超过了预设的QPS 阑值，则后续的请求立即失败返回，不再继续处理。通过快速失败，将超出系统处理能力的流量直接过滤掉，保障了系统的可用性。
* 校验权限：UMS配置报表查询的用户角色及角色权限，用户调用此接口时校验用户的权限，如果有权限则执行用户的查询，没有权限则不执行。
* 查询优化：Get 请求与List 请求分别有独立的线程池进行查询，但是一个请求具体是Get 还是List ，则依赖调用者具体调用哪个方法。在很多情况下，调用者调用的方法不一定是最合适的。
* 模型缓存：将解析后的模型（包括逻辑模型、物理模型）缓存在本地。

## 数据推送服务架构



数据推送服务主要提供Web Socket 和long polling 两种方式，其应用场景主要是商家端实时直播。在强制刷新这种情况下long polling 会给服务器带来成倍的压力。而WebSocket 方式，可以在这种场景下，有效地缓解服务器的压力，给用户带来最实时的体验。

数据推送服务主要提供即时任务，是一个面向MetaQ 等不同消息源，通过定制过滤规则，向Web 、无线等终端推送消息的中间件平台。核心服务器端基于高性能异步事件驱动模型的网络通信框架Netty 4实现，结合使用Guava 缓存实现本地注册信息的存储，Filter 与Server 之间的通信采用Thrift 异步调用高效服务实现，消息基于Disruptor 高性能的异步处理框架（可以认为是最快的消息框架）的消息队列，在服务器运行中Zookeeper 实时监控服务器状态，以及通过Diamond 作为统一的控制触发中心。

# 数据平台开发规划

数据平台的功能覆盖从数据仓库规划、研发、部署、运维、管理等大数据开发的全流程，功能比较全，需要长时期的开发才能完成。需要将整个开发计划分期分步开发。功能版块按照数据开发的流程部署开发最好，开发流程依次为：管理中心、规划、研发、资产、服务，需要很长的开发周期才能服务与业务。为了将已经有报表功能转到大数据部这边，需要将服务功能版块提前开发。

## 近期规划

按模块间的相互依赖关系排优先级如下：

* 服务版块的元数据管理
* 服务版块的主处理模块服务
* 服务版块的其他模块

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 编号 | 版块 | 模块 | 任务 | 计划持续时间 | 计划开始时间 |
| 1 |  | 模块 | demo模块 | 前端demo |  |  |
| 2 |  | 模块 | demo模块 | 后端demo |  |  |
| 3 |  | 服务 | 元数据管理 | 前端首页 |  |  |
| 4 |  | 服务 | 元数据管理 | 前端编辑页面 |  |  |
| 5 |  | 服务 | 元数据管理 | 后端服务 |  |  |
| 6 |  | 服务 | 元数据管理 | 后端支持MySQL、HBase元数据 |  |  |
| 7 |  | 服务 | 主处理模块 | 前端SQL管理页面 |  |  |
| 8 |  | 服务 | 主处理模块 | 后端SQL处理逻辑 |  |  |
| 9 |  | 服务 | 服务总线 | 前端API管理页面 |  |  |
| 10 |  | 服务 | 服务总线 | 后端API服务 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

## 中期规划

规划

研发

## 远期规划

资产

# 数据平台部署

## 业务系统数据量

业务系统累积到现在的数据量如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 业务系统 | 数据库 | 网络 | 数据量（截至到2019年8月16日） |
| 分销系统DMS | Aliyun RDS MySQL | Aliyun经典网络 | 30.0GB |
| 订单管理系统OMS |
| 物流管理系统TMS |
| 安装平台安达人 |
| K3 ERP | MSSQL | Aliyun经典网络 | 90.5GB |
| 联思FMS | MSSQL | 优宝家内网（通过VPN访问） | 6.8GB |
| 总数据量 |  |  | 127.3GB |

目前每处理订单量大概100单，每个订单数据量为10KB，订单在业务系统流转6次，每次流转数据量增加2倍，所以可大概计算每天的增量数据量为：

每天增量数据量 = 100 X 10 X 6 X 2 = 12MB

每月增量数据量 = 12 X 31 = 0.4GB

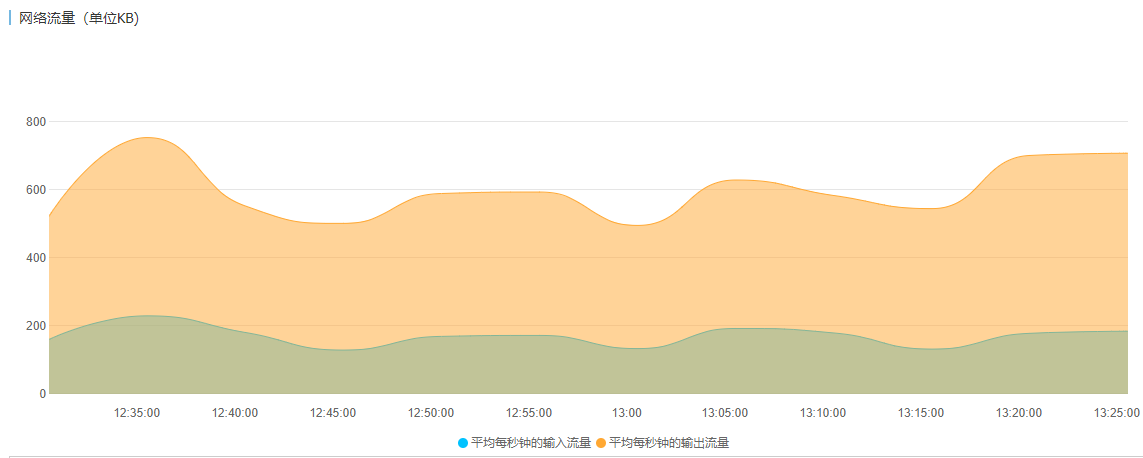
每年增量数据量 = 0.4 X 12 = 5GB

通过RDS备份日志也可以看出每天备份数据量大概10MB左右。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 备份开始/结束时间 | 备份策略 | 备份大小 | 备份方法 | 备份类型 | 状态 |
| 2019-09-23 19:39~2019-09-23 20:02 | 实例备份 | 6.89G | 物理备份 | 全量 | 完成备份 |
| 2019-09-22 19:39~2019-09-22 20:02 | 实例备份 | 6.89G | 物理备份 | 全量 | 完成备份 |
| 2019-09-20 19:39~2019-09-20 20:04 | 实例备份 | 6.88G | 物理备份 | 全量 | 完成备份 |
| 2019-09-18 19:39~2019-09-18 20:04 | 实例备份 | 6.87G | 物理备份 | 全量 | 完成备份 |
|  |  |  |  |  |  |

## 业务系统带宽

RDS的数量流量统计如下：



入口流量基本保持在200KB/s

出口流量基本保持在600KB/s

## 数据平台存储

根据目前的历史数据和增量数据，规划的数据盘为500GB，满足一到两年业务数据增长需求，系统盘为100GB，满足后继增加Spark部署需要。后续业务如果增长加快，再增加相关资源。

## 数据平台计算

根据目前的数据平台需求，大部分BI需求为运行SQL计算，不需要很多算力。小部分ML需求和业务紧密相关的，对实时性要求比较高，需要很多算力。

## 数据平台网络

业务系统数据库大部分上云，运行在阿里云RDS VPC网络环境下。另外FMS联思系统数据库在优宝家内网下，需要通过VPN才能接到到数据，VPN使用开源SoftEther，需要在ECS安装VPN Client程序，需要公网IP。阿里云DataV和QuickBI访问数据平台数据可以使用VPC网络，对外发布数据必须使用公网IP，1MB/s可以满足K3数据同步和发布数据需求。

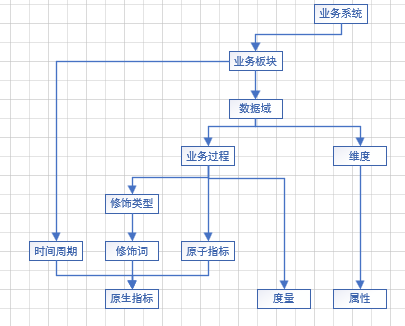
## 数据平台ECS配置

|  |  |  |  |
| --- | --- | --- | --- |
| 资源 | 配置 | 价格（元） | 部署应用 |
| ECS1数据服务 | 基础配置  计费方式 ： 包年包月  地域 ： 华北 2 (北京) / 随机分配  实例 ： 通用型 g6 / ecs.g6.large(2vCPU 8GiB)  购买数量 ： 1 台  镜像 ： CentOS 7.6 64位(安全加固)  系统盘 ： 高效云盘 200GiB  数据盘 ： 1 块 500GiB  网络和安全组  网络 ： 专有网络  公网带宽 ：绑定弹性公网IP | 5000.00  (2019年9月24日报价) | mySQL代理，web server |
| ECS2计算服务器 | 基础配置  计费方式 ： 包年包月  地域 ： 华北 2 (北京) / 随机分配  实例 ： 通用型 g6 / ecs.g6.large(8vCPU 32GiB)  购买数量 ： 1 台  镜像 ： CentOS 7.6 64位(安全加固)  系统盘 ： 高效云盘 200GiB  数据盘 ： 1 块 500GiB  网络和安全组  网络 ： 专有网络  公网带宽 ：绑定弹性公网IP | 11000.00  （2019年9月24日报价） | EasyScheduler、ML、ETL |

# 数据管理开发

## 数据管理

根据前期调研的过程，将整个调研过程用流程固化下来如下图。业务系统下分多个业务板块，业务板块下又可以分为多个数据域，数据域下又可以分为多个业务过程，同时一个数据域里有多个维度，业务过程里有许多原子指标，有了原子指标就可以生成许多衍生指标。将这个数据指标管理的流程用UI页面管理起来。



## 数据任务开发

每个层次之间的数据计算清洗转换，需要开发计算任务或者服务。

1. 业务系统->ODS层数据同步任务：这个任务可以DataX实现，主要需要配置DataX Job，放入EasyScheduler任务调度器里定时执行。Job任务为json格式，可以根据数据管理系统配置的元数据，用脚本或者服务自动将配置的元数据转换为job。
2. ODS->DWD数据清洗任务：这个任务可以用Python实现或者服务实现。
3. DWD->DWS数据聚合任务：这个任务可以用SQL或者Python实现。
4. DWS->ADS数据主题计算任务：这个任务可以用SQL或者Python实现。

前期计算任务开发可以使用Python开发，总结业务，手动提交到Git，手动部署实现。后期可以将这些任务做成服务自动部署到运行系统。