|  |  |
| --- | --- |
| 产品名称 | 密级 |
| DataOne数据模型 | 机密 |
| 产品版本 | 共 页 |
|  |

DataOne开发手册

|  |  |  |  |
| --- | --- | --- | --- |
| 拟制 |  | 日期 |  |
| 审核 |  | 日期 |  |
| 批准 |  | 日期 |  |



华为技术有限公司

版权所有 侵权必究

（DVP05T04 V2.8 / 仅供内部使用）

修订记录

| 日期 | 修订版本 | CR号 | 修改章节 | 修改描述 | 作者 |
| --- | --- | --- | --- | --- | --- |
| 2016-12-11 | 1.00 |  |  | 初稿完成 | *梁笑* |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**关键字**

目录

[1. DataOne概述 5](#_Toc469335392)

[1.1 背景介绍 5](#_Toc469335393)

[1.2 DataOne模型架构 6](#_Toc469335394)

[1.3 DataOne建模过程 7](#_Toc469335395)

[1.3.1 数据调研 7](#_Toc469335396)

[1.3.2 数据建模 7](#_Toc469335397)

[2.数据调研介绍 8](#_Toc469335398)

[2.1 ID调研 8](#_Toc469335399)

[2.2 BD调研 10](#_Toc469335400)

[3.模型总体介绍 12](#_Toc469335401)

[4.整合层模型介绍 12](#_Toc469335402)

[4.1 数据清洗 12](#_Toc469335403)

[4.2 域模型 13](#_Toc469335404)

[4.2.1参与人域 14](#_Toc469335405)

[4.2.2产品域 15](#_Toc469335406)

[4.2.3 营销域 16](#_Toc469335407)

[4.2.4业务交互域 16](#_Toc469335408)

[4.2.5 参考域 17](#_Toc469335409)

[4.3代码表使用 17](#_Toc469335410)

[5汇总层模型介绍 18](#_Toc469335411)

[7.DataOne环境介绍 19](#_Toc469335412)

[7.1 DataOne集群介绍 19](#_Toc469335413)

[7.2 权限控制 20](#_Toc469335414)

[7.3 TCC使用 20](#_Toc469335415)

[8．DataOne开发规范 20](#_Toc469335416)

[9．DataOne需求受理流程 21](#_Toc469335417)

[9.1 ODS接入数据需求流程 21](#_Toc469335418)

[9.2 CDM表变更流程 21](#_Toc469335419)

[9.3 ADS表变更流程 21](#_Toc469335420)

[9.4 新增TCC变更流程 21](#_Toc469335421)

[10.其他 22](#_Toc469335422)

[10.1 erwin的使用 22](#_Toc469335423)

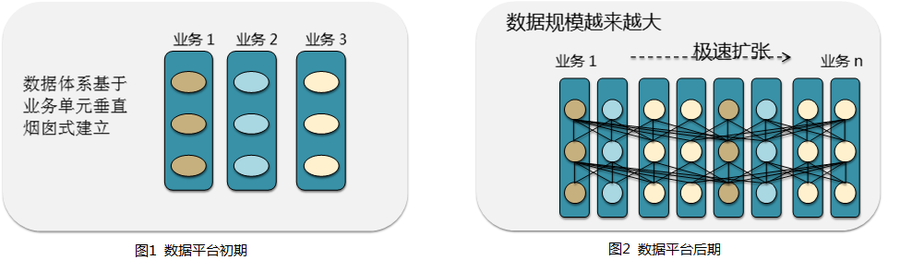
[10.2 拉链表的使用 22](#_Toc469335424)

[10.3 UDF的使用 23](#_Toc469335425)

# DataOne概述

## 背景介绍

大数据平台承接消费者云服务部的大数据建设，目前已接入了包括用户经营、内容经营、开放平台、基础云、家庭云、vmall、软件工程部等8大业务部门的120多个业务的数据，支撑在线报表有650多张，用户画像200多个标签，以及推荐、MAA等20多个系统。随着数据量的累积，总的数据量已达到12P，数据价值越来越突出。但是，大数据平台在建设初期，并没有考虑建立统一的数据模型，需求都是烟囱式实现。



随着需求的增长，带来的问题越来越突显，主要问题有以下方面：

1. 数据一致性问题严重

主要表现为同一指标在不同的报表中，或者在报表与画像系统中，统计结果不相同。

1. 数据冗余

主要表现为相同的指标在多个报表中重复计算，相同的数据在多张表中重复存储。

1. 数据质量差

ODS表进入数据平台后未做数据清洗，表的数据未做统一规范，脏数据占比高。

1. 数据爆炸

ODS表每天的数据增量有2T左右，而集群每天的数据增量达到30T，数据重用率低导致数据量急剧增长。

1. 权限控制

所有表都建在一个库下，用户可直接对ods进行操作，建表比较随意，权限界限不明确。

1. 元数据管理

无元数据管理机制，要了解数据基本通过口口相传，或者看脚本，数据门槛高。

1. 研发效率低

平台中的表复用率低，很多报表开发人员都是从ODS表到最终结果表一路开发下来，开发路径长，研发效率低。

DataOne模型致力于解决上述暴露出的问题，建立统一的数据模型，规范数据质量，提升研发效率。

## DataOne模型架构

DataOne模型采用4层架构。

1. ODS层中各个业务推送的数据，经过ETL进入到ODS，在ods层以统一的文件格式进行数据落地与存储。

2. 整合层是由ODS层数据经过数据清洗，编码映射和转换后，将静态信息归类与整合后，到达数据整合层。

3. 汇总层是对整合层和ODS层数据进行再次处理，根据使用者的要求，对数据进行多维度、多层次存储，支撑应用层数据展现。

4. 应用层是基于需求和应用展现，对本系统内的数据直接提供给业务部门使用。

## DataOne建模过程

### 数据调研

大数据建模首先要做的是了解数据，通过数据调研理解各个业务系统的元数据，打通不同业务系统间的数据，用模型展现基础数据间的天然关系。

数据调研的工作主要包括ID调研和BD调研。ID调研调研各个业务系统的业务数据，了解各个系统的业务含义，有什么样的表，表的数据如何。BD调研即需求调研，调研现有的报表、画像、推荐等各个系统的业务需求，将需求归纳总结为统一的指标或者维度。

### 数据建模

在ID调研和BD调研的基础上，进行数据建模。数据建模的过程可以分为概念模型、逻辑模型、物理模型三个阶段。

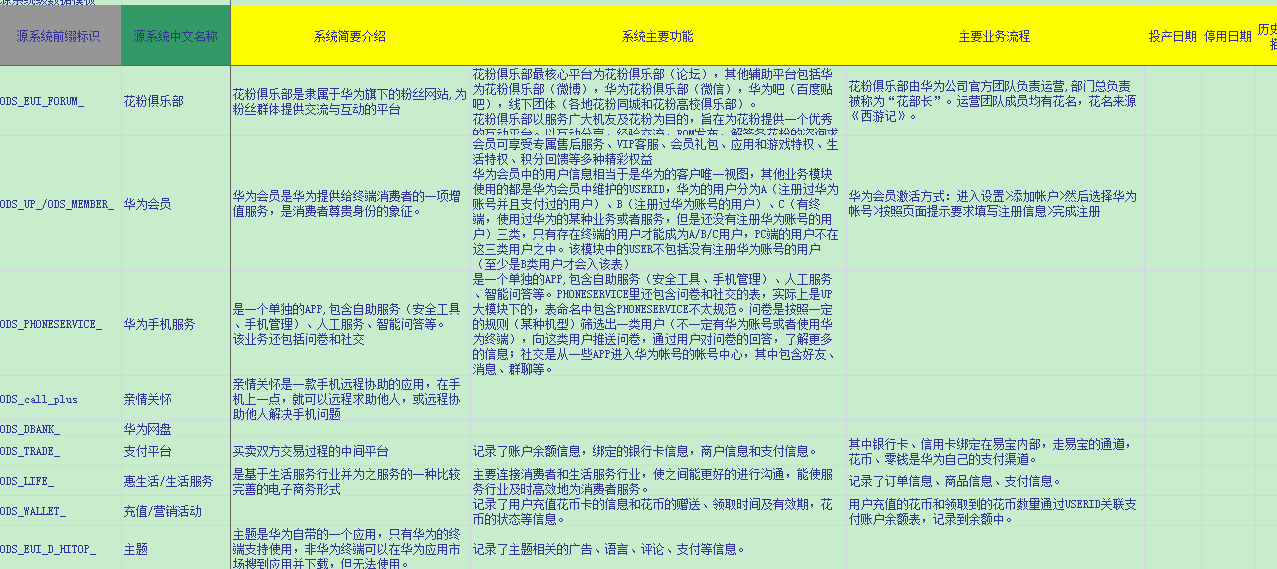


# 2.数据调研介绍

## 2.1 ID调研

ID调研是针对ODS表的原始数据的调研，主要调研内容包括系统级调研，表级调研以及字段级调研，认识数据在系统间流转关系，确定建模的系统范围、表范围和字段范围。

1. 系统级调研主要调研ODS中接入的业务系统有哪些，这些系统的业务含义，系统的主要功能，系统的业务流程以及系统有哪些表等，对数据源进行一个整体的调研。



1. 表级调研的主要调研内容包括表的中文标识，来源系统，业务分类，增量规则，表中有什么样的数据以及跟其他表的关系等。



1. 字段级调研的主要调研内容包括表的主键，字段的类型，字段的中文解释，字段中有什么样的数据，字段中的码值对应等。



ID调研梳理出各个源系统的表之间的实体关系，通过实体关系展现出数据在表之间的流转关系。



通过ID调研，将源系统数据归纳划分为参与人、产品、营销、业务交互以及地址五大主题域。划分好主题域后，将各个ODS表对应到相应的主题域。ID调研的结果为整合层模型提供建模指导。

Id调研的文档可以在以下路径下载：

系统级调研：

表级调研： 

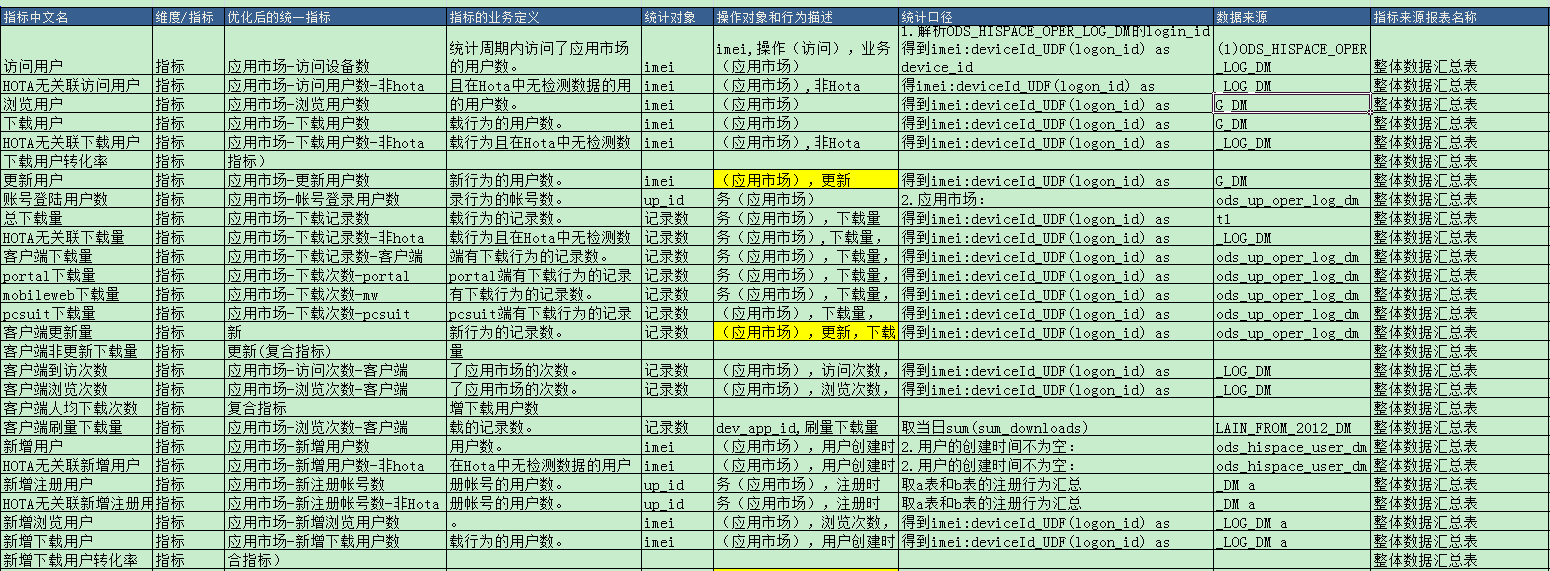
字段级调研：

## 2.2 BD调研

BD调研也就是需求调研，调研的对象包括大数据平台目前对外承接的报表，以及用户画像系统中的标签，调研的内容为各个指标或维度的业务口径、技术口径、数据源加工逻辑等。BD调研的结果为汇总层模型提供建模指导。

BD调研的输出文档有3部分：

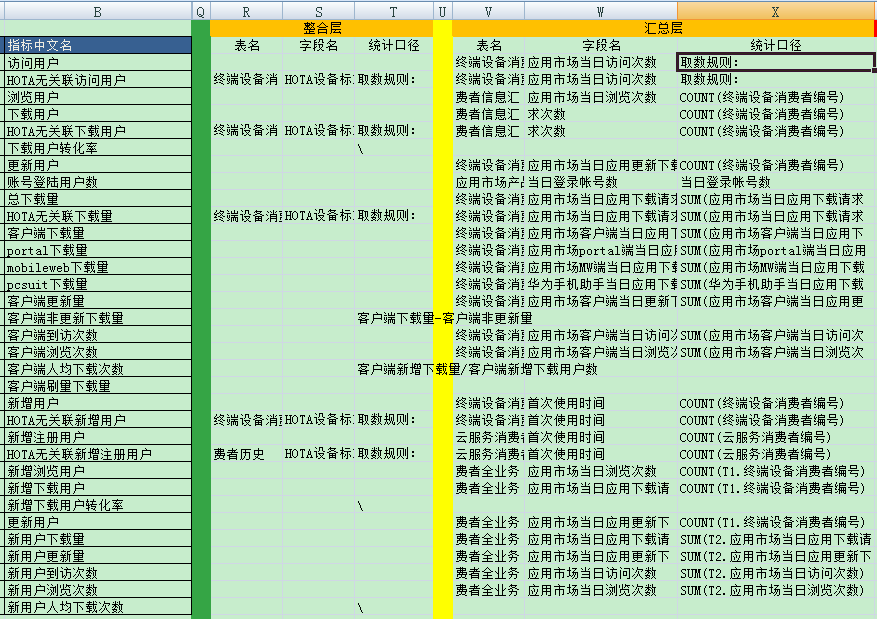
第一部分每张报表的指标口径调研结果。



第二部分将分布在各个报表中的指标汇总后的指标口径汇总，根据指标汇总加工成相应表的字段属性。



第三部分各个报表的指标与模型的映射关系，说明报表如何使用新的模型实现。



以上文档第一部分和第二部分汇总在《BD调研合并.xlsx》：



第二部分汇总在《汇总层指标汇总.xls》：

# 

# 3.模型总体介绍

DataOne模型的总体架构分ODS层，整合层，汇总层，应用层四层。四层在模型架构里面已介绍，此处不在赘述。

模型建模工具采用erwin7.3,在配置库上可以取到数据模型：

# 4.整合层模型介绍

整合层的作用是将ODS数据经过数据清洗，编码映射和转换后，对静态信息做归类与整合。整合层的整合对象为来自各个业务系统的静态数据。整合层将业务数据划分为参与人、产品、营销、业务交互、地址五大主题域。ID调研完成会将各个ODS的表归入五大域。整合层对五大主题域的表进行数据整合。

## 4.1 数据清洗

在ODS的数据进入整合层域模型和汇总层之前，会先进行数据的加解密、数据清洗，大字段拆分等动作。

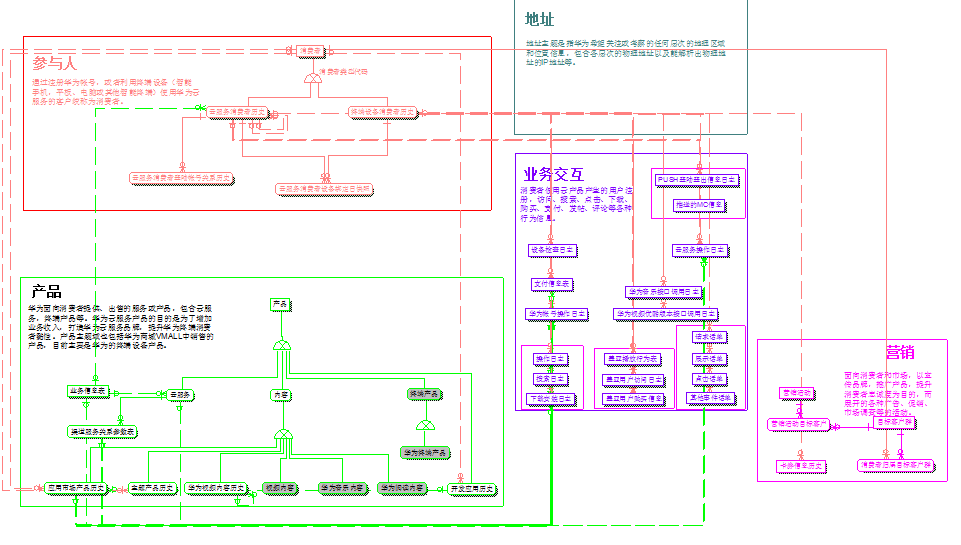
数据清洗表的命名采取的规则为：比如有一张ods表ods\_hispace\_oper\_log\_dm进入整合层后表名为dwd\_hispace\_oper\_log\_dm;

整合层数据清洗规则主要包括：

1. 将不符合业务规则的数据字段值置为空或者删除：比如帐号id规则应该全为数字，不为数字的帐号过滤掉。
2. 统一数据格式。比如日期统一为YYYYMMDD.
3. 字段的加解密操作。统一对字段解密后再用sha256加密。
4. 统一字段命名。比如设备Id统一用字段名imei.
5. 大字段拆分。将大字段拆分成符合业务规则的有实际业务含义的字段。

## 4.2 域模型

整合层将业务数据划分为参与人（CUST）、产品(PROD)、营销(CAMP)、业务交互(EVENT)、地址(ADDR)五大主题域。



整合层表的命名以DWD\_主题域开头。整合层目前有29张数据表。

|  |  |  |
| --- | --- | --- |
| 表名 | 中文表名 | 对应的脚本 |
| dwd\_prod\_hispace\_app\_info\_ds | 应用市场产品历史 | dwd\_prod\_hispace\_app\_info\_ds.sql |
| dwd\_prod\_theme\_class\_rela\_ds | 主题产品分类关系 | dwd\_prod\_theme\_class\_rela\_ds.sql |
| dwd\_prod\_theme\_info\_ds | 主题产品历史 | dwd\_prod\_theme\_info\_ds.sql |
| dwd\_prod\_theme\_class\_info\_ds | 主题分类历史 | dwd\_prod\_theme\_class\_info\_ds.sql |
| dwd\_cust\_up\_info\_ds | 云服务消费者历史 | dwd\_cust\_up\_info\_ds.sql |
| dwd\_cust\_user\_device\_bind\_rela\_dm | 云服务消费者设备绑定关系历史 | dwd\_cust\_user\_device\_bind\_rela\_dm.sql |
| dwd\_cust\_login\_acct\_real\_ds | 云服务消费者登陆帐号关系历史 | dwd\_cust\_login\_acct\_real\_ds.sql |
| dwd\_prod\_dev\_app\_info\_ds | 开发应用历史 | dwd\_prod\_dev\_app\_info\_ds.sql |
| dwd\_cust\_fans\_statistics\_info\_dm | 花粉消费者统计信息日快照 | dwd\_cust\_fans\_statistics\_info\_dm.sql |
| dwd\_cust\_group\_rela\_ds | 消费者群组关系历史 | dwd\_cust\_group\_rela\_ds.sql |
| dwd\_cust\_group\_info\_ds | 消费者群组历史 | dwd\_cust\_group\_info\_ds.sql |
| dwd\_cust\_feature\_info\_ds | 消费者特征历史 | dwd\_cust\_feature\_info\_ds.sql |
| dwd\_cust\_addr\_rela\_ds | 消费者地址关系历史 | dwd\_cust\_addr\_rela\_ds.sql |
| dwd\_prod\_hwmovie\_catalog\_info\_ds | 盖亚栏目历史 | dwd\_prod\_hwmovie\_catalog\_info\_ds.sql |
| dwd\_prod\_hwmovie\_content\_catalog\_rela\_ds | 盖亚栏目内容关系 | dwd\_prod\_hwmovie\_content\_catalog\_rela\_ds.sql |
| dwd\_prod\_hwmovie\_catalog\_spec\_rela\_ds | 盖亚栏目专题关系 | dwd\_prod\_hwmovie\_catalog\_spec\_rela\_ds.sql |
| dwd\_prod\_hwmovie\_content\_info\_ds | 盖亚内容历史 | dwd\_prod\_hwmovie\_content\_info\_ds.sql |
| dwd\_prod\_hwmovie\_content\_spec\_rela\_ds | 盖亚内容专题关系 | dwd\_prod\_hwmovie\_content\_spec\_rela\_ds.sql |
| dwd\_prod\_hwmovie\_spec\_info\_ds | 盖亚专题历史 | dwd\_prod\_hwmovie\_spec\_info\_ds.sql |
| dwd\_cust\_device\_info\_ds | 终端设备消费者历史 | dwd\_cust\_device\_info\_ds.sql(包含18个小脚本) |
| dwd\_prod\_service\_info | 业务信息历史 | dwd\_prod\_service\_info.sql |
| dwd\_prod\_package\_info | 云服务历史 | dwd\_prod\_package\_info.sql |
| dwd\_common\_device\_model\_spec | 终端型号参数表 | dwd\_common\_device\_model\_spec.sql |
| dwd\_common\_info\_code | 代码表 | dwd\_common\_info\_code.sql dwd\_common\_info\_code\_handwork.sql |
| dwd\_common\_region\_code | 行政区划代码表 | dwd\_common\_region\_code.sql |
| dwd\_common\_code\_mapping | 代码转换表 | dwd\_common\_code\_mapping.sql |
| dwd\_common\_device\_mapping | 终端设备消费者唯一编号映射表 | dwd\_common\_device\_mapping.sql |
| dwd\_common\_emui\_device\_secretkey | EMUI主题用户秘钥表 | dwd\_common\_emui\_device\_secretkey.sql |
| dwd\_common\_channel\_service\_mapping | 渠道服务关系参数表 | dwd\_common\_channel\_service\_mapping.sql |

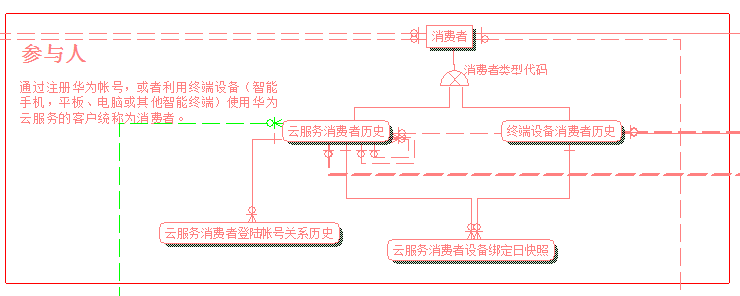
查看这些表的逻辑的方法有两个，一个是看映射文档，通过映射文档可以得到每张表的数据来源：

另一个是看对应的脚本：

https://119.145.15.45:3690/svn/TC\_EMUI\_DataService\_SVN/合作方目录/数据架构/09.开发/整合层二轮测试脚本

### 4.2.1参与人域

参与人目前包含的实际只有消费者，后续考虑增加合作伙伴，竞争者等概念。通过注册华为帐号，或者利用终端设备（智能手机，平板、电脑或其他智能终端）使用华为云服务的客户统称为消费者。

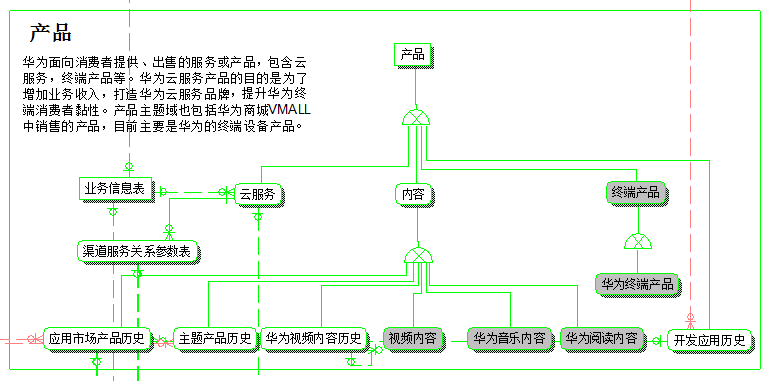


参与人域模型表的元数据在以下文档中进行查看。



### 4.2.2产品域

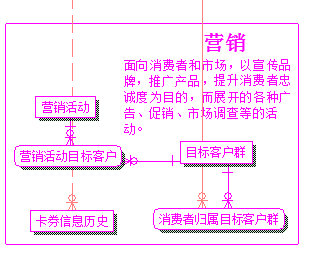
产品域整合的是华为面向消费者提供、出售的服务或产品信息，包含云服务，终端产品等。华为云服务产品的目的是为了增加业务收入，打造华为云服务品牌，提升华为终端消费者黏性。产品主题域也包括华为商城VMALL中销售的产品，目前主要是华为的终端设备产品。



产品域模型表的元数据在以下文档中进行查看。

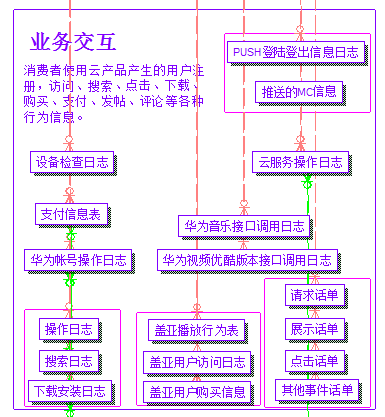
### 4.2.3 营销域

营销域的整合对象是面向消费者和市场，以宣传品牌，推广产品，提升消费者忠诚度为目的，而展开的各种广告、促销、市场调查等的活动。目前整合层营销域只有一张营销的清单信息表。



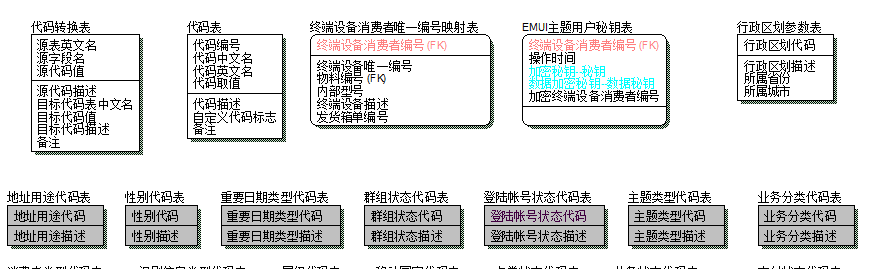
### 4.2.4业务交互域

业务交互域的范围为各类日志表。消费者使用云产品产生的用户注册，访问、搜索、点击、下载、购买、支付、发帖、评论等各种行为信息都可以纳入业务交互域。目前整合层中对日志类整合表主要是大数据采集的整合表。



### 4.2.5 参考域

参考域的表有代码表和参考表。代码表存放的是一些业务含义的编码与码值的对应关系。参考表存放的是公共的标准编码信息。



## 4.3代码表使用

代码表的作用是为各层表中的字段的属性值提供码值翻译。目前的代码表主要有

|  |  |  |
| --- | --- | --- |
| **表中文名** | **表英文名** | **说明** |
| 代码转换表 | dwd\_common\_code\_mapping | 代码转换的目的是将数据源中相同含义不同码值的数据进行转换。比如男女统一转化为0，1 |
| 代码表 | dwd\_common\_info\_code | 提供代码的解释说明 |
| 行政区划代码表 | dwd\_common\_region\_code | 行政区域的代码表 |

代码表的使用方法：

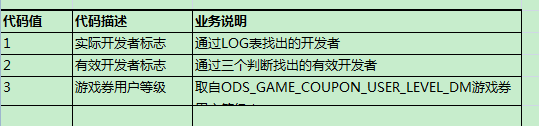
1. 通过文档进行查询。对应文档：《华为数据架构数据标准\_公共代码.xlsx》

比如在DWD\_CUST\_FEATURE\_INFO\_DS中查到Cust\_Feature\_Type\_Cd字段使用了码值CD0010。

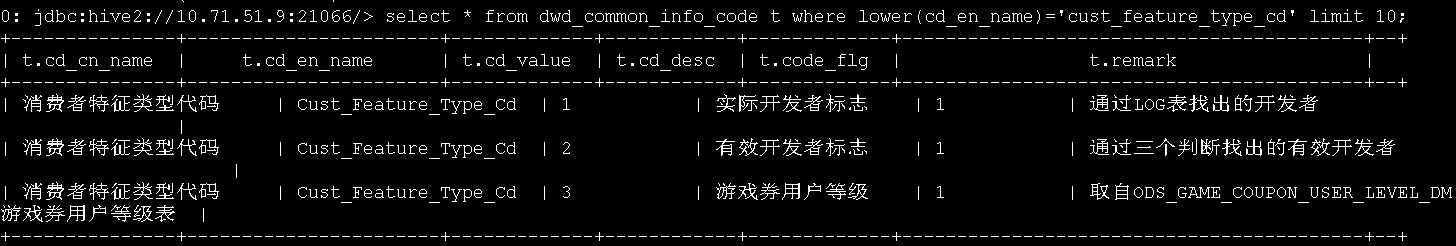


可以在公共代码表中查询该码值对应的业务含义。





1. 在数据表中查询，可通过字段名。比如上面例子中的Cust\_Feature\_Type\_Cd。



# 5汇总层模型介绍

汇总层模型主要汇总业务数据的动态信息，基于需求调研将业务数据进行多维度多层次的加工，汇总成满足业务需求的基础指标。汇总层目前的主要表有：

|  |  |  |
| --- | --- | --- |
| **汇总层表中文名** | **汇总层表英文名** | **说明** |
| 营销任务信息汇总 | dws\_task\_info\_dm | 主要汇总对象为hiad的营销以及画像系统的营销。 |
| 应用市场TOP1000关键字搜索信息 | dws\_hispace\_keywords\_search\_top1000\_dm | 汇总应用市场的关键字搜索常用指标。 |
| 应用市场产品信息汇总 | dws\_hispace\_app\_info\_dm | 对APP相关的常用指标进行汇总。 |
| 云服务消费者信息汇总 | dws\_up\_info\_dm | 对帐号相关的常用指标进行汇总。 |
| 云服务消费者应用产品统计信息 | dws\_up\_app\_statistics\_info\_dm | 对帐号使用游戏app的支付或游戏券类指标进行汇总。 |
| 终端设备消费者信息汇总 | dws\_device\_info\_dm | 对设备相关的常用指标进行汇总。 |
| 终端设备消费者应用产品统计信息 | dws\_device\_app\_statistics\_info | 对设备安装或卸载app的指标进行汇总。 |
| 花粉论坛统计信息汇总 | dws\_fans\_forum\_statistics\_info\_dm | 花粉论坛相关的指标汇总数据。 |
| 华为视频内容信息汇总 | dws\_hwmovie\_content\_info\_dm | 华为视频中视频相关的指标信息汇总。 |
| 用户操作事件日快照 | dws\_user\_oper\_event\_dm | 对常用的日志操作表进行累计汇总。 |
| 云服务消费者全业务统计信息 | dws\_up\_statistics\_info\_dm | 帐号使用各个业务的使用信息汇总 |
| 支付信息日快照 | dws\_trade\_info\_dm | 支付相关的快照信息汇总。 |
| 终端设备消费者全业务统计信息表 | dws\_device\_statistics\_info\_dm | 设备使用各个业务的汇总数据。 |
| 终端设备消费者地域轨迹信息 | DWS\_DEVICE\_AREA\_INFO | 终端设备的地域轨迹信息。 |
| 终端设备消费者地域轨迹日表 | DWS\_DEVICE\_AREA\_DM | 地域轨迹信息的小时表。 |

汇总层指标文档介绍了各个表的各个指标的口径，以及覆盖的报表：



# 7.DataOne环境介绍

## 7.1 DataOne集群介绍

DataOne集群一期规划150台服务器，145个数据节点。数据通道和网关机如下所示：

|  |  |
| --- | --- |
| 数据通道 | 网关机 |
| 10.71.52.11 lf2-bi-d1dw-71-52-11 | 10.71.53.16 lf2-bi-d1sgw-71-53-16 |
| 10.71.52.12 lf2-bi-d1dw-71-52-12 | 10.71.54.16 lf2-bi-d1sgw-71-54-16 |
| 10.71.53.11 lf2-bi-d1dw-71-53-11 | 10.71.55.16 lf2-bi-d1sgw-71-55-16 |
| 10.71.54.11 lf2-bi-d1dw-71-54-11 | 10.71.56.16 lf2-bi-d1sgw-71-56-16 |
| 10.71.55.11 lf2-bi-d1dw-71-55-11 |  |
| 10.71.56.11 lf2-bi-d1dw-71-56-11 |  |

FI的管理界面地址为：<https://10.71.51.198:28443/web/>

登录用户：yarntask/Huawei@123 此用户只有查询权限

壳目录地址：/opt/huawei/BIBigdata/ExecuteShell/

## 7.2 权限控制

DataOne集群主要用户有三个：biods,BICoreData,BIApp三个用户对各个数据层的访问权限控制如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | biods | | BICoreData | | BIApp | |
| R | W | R | W | R | W |
| ODS | | ✓ | ✓ | ✓ | 🗶 | 🗶 | 🗶 |
| CDM | DWD | 🗶 | 🗶 | ✓ | ✓ | ✓ | 🗶 |
| DWS | 🗶 | 🗶 | ✓ | ✓ | ✓ | 🗶 |
| ADS | | 🗶 | 🗶 | ✓ | 🗶 | ✓ | ✓ |

## 7.3 TCC使用

DataOne搭建新的集群，集群地址为：https://10.71.7.30:18443/TaskControlCenter/

TCC除ODS入库任务外新建任务名称统一使用小写。

# 8．DataOne开发规范

这部分规范参考《DataOne数据质量标准与规范.docx》

# 9．DataOne需求受理流程

## 9.1 ODS接入数据需求流程

触发该流程的入口条件：ODS新数据源接入。

（1）新接入数据源首先由业务责任人填写数据说明文档，以及建表脚本，并发起建表申请。审批人：何远洪/梁笑。

（2）申请通过后，由业务责任人将表导入元数据管理系统。

（3）以上条件都具备后，由ODS管理员进行建表。

（4）建表成功后由业务责任人发起TCC变更申请。

## 9.2 CDM表变更流程

触发该流程的入口条件：现有的CDM表不能满足业务需求。

1. 填写CDM需求文档，描述业务需求的业务口径与技术口径，发起变更申请。审批人：CDM责任人。
2. CDM责任人根据业务需求判断是否需要变更。
3. 若投入变更，判断是在原有表的基础上增加字段，还是增加新的表。

## 9.3 ADS表变更流程

（1）填写建表申请，发起变更申请。审批人：何远洪/梁笑。

（2）审批通过后由ADS层运维人员执行变更。

## 9.4 新增TCC变更流程

（1）填写TCC变更申请文档。

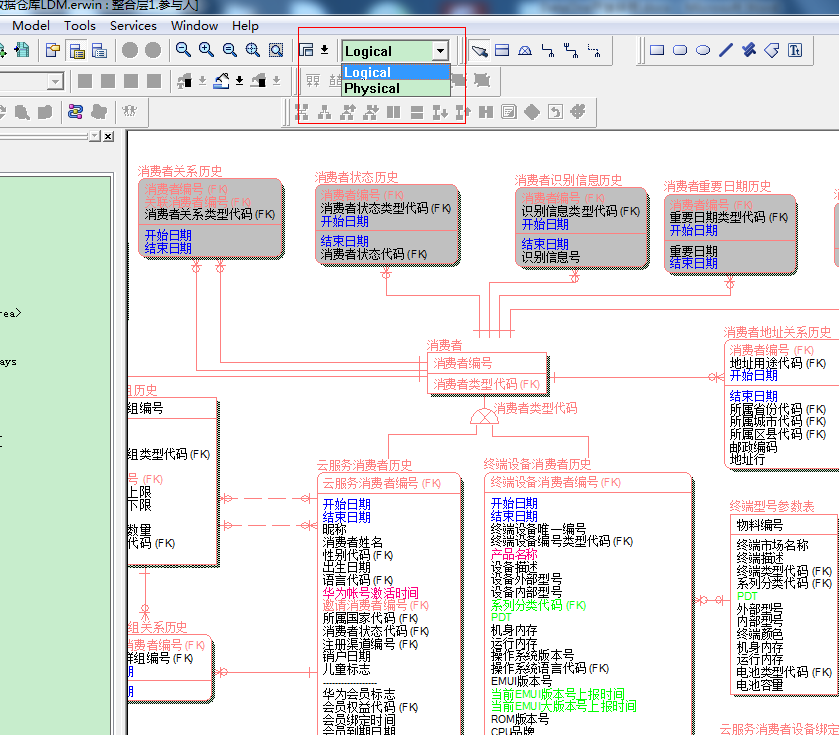
（2）审批通过后有TCC管理员执行变更。

# 10.其他

## 10.1 erwin的使用

1.安装文件：[https://119.145.15.45:3690/svn/TC\_EMUI\_DataService\_SVN/合作方目录/数据架构/04.数据仓库LDM/ERWin 7.3.zip](https://119.145.15.45:3690/svn/TC_EMUI_DataService_SVN/合作方目录/数据架构/04.数据仓库LDM/ERWin%207.3.zip)

2. 逻辑模型与物理模型的切换



## 10.2 拉链表的使用

拉链表以end\_date作为分区，非99991231分区为当日发生变化的数据，99991231为全量分区，存储主键的当前状态。在使用拉链表时，若要查询全量数据的当前状态，限制pt\_d=99991231。为了支持重跑，在脚本中的写法为start\_date<='$ date' AND end\_date>'$ date'。

## 10.3 UDF的使用

DataOne集群对UDF进行了整合，将之前同一JAR包，但是名称创建不同的UDF进行了统一命名，比如将IsEmpty\_udf, IsEmpty, IsEmptyudf统一创建为IsEmpty。老的UDF跟新的UDF的对应关系可在文档中进行查询。文档位置：

