**DATAONE需求调研设计操作流程**

**更新情况**

|  |  |  |
| --- | --- | --- |
| 创建(修改)人 | 创建（修改）时间 | 重要更新内容 |
| 李慧东 | 20170531 | 更新数据平台相关人员表格 |
| 李慧东 | 20170609 | 基本完成该表的初稿 |

# 1.简介

业务将源表推送至大数据平台成为ODS表，调研人员根据该表数据字典以及和业务人员的沟通，了解该表用途，解决数据问题，最终设计成DWD表，提交开发。

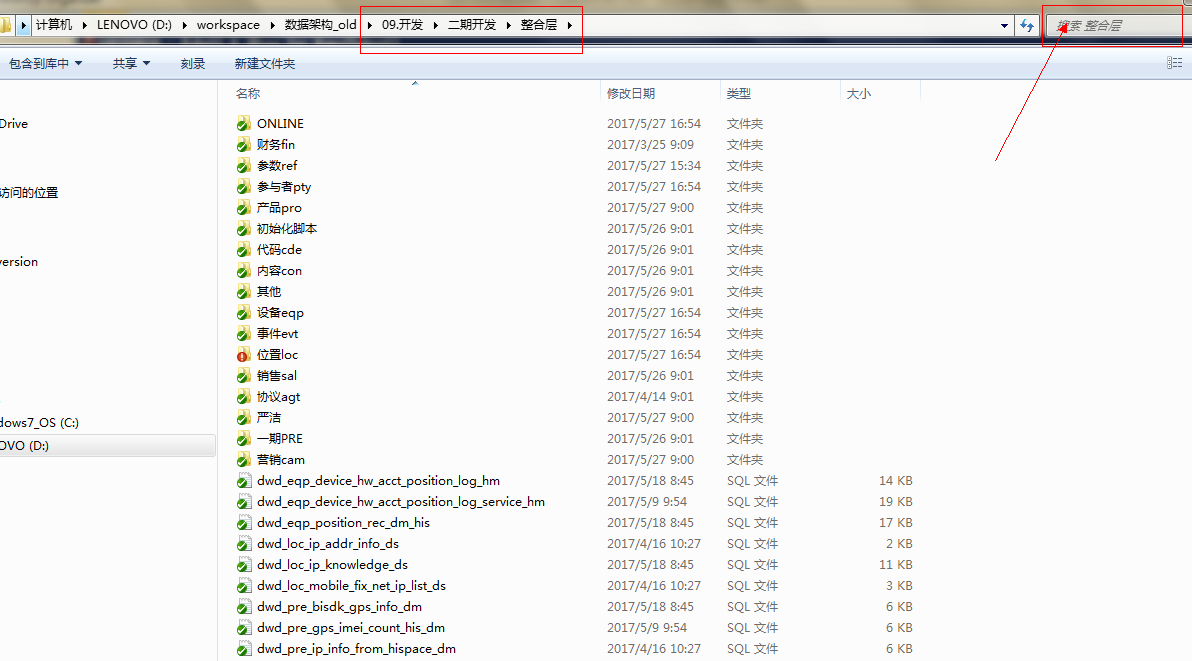
# 2.调研流程

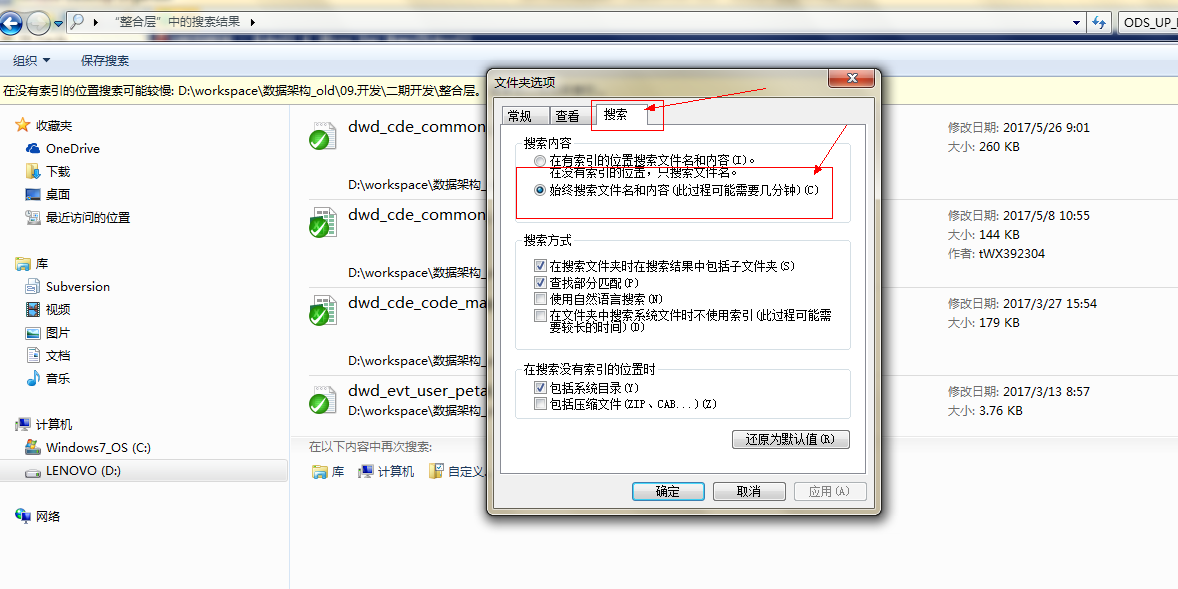
## 2.1 查看是否已进入开发

**第一种方法：**

在09.开发文件夹中搜索该表是否已经进入开发，已经进入则无需在调研了

点击浏览器左上 组织-文件夹和搜索选项

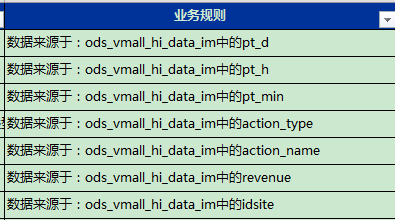




只要有dwd的sql出现，大致看一下脚本内容就可以了



第2种方法：在数据仓库整合层-元数据2.0Excel中相关主题域（如果不太好把握就把所有主题域翻看一边）的业务规则中查看。

****

## 2.2 获取接口文档

在SVN\06.数据架构\01.数据源\数据源接入\已完成文件夹中获取需要调研表的数据字典以及其他相关文件

数据字典样式：



## 2.3 找到业务对接人

拿到文档后，准确定位所属业务范围，找到业务对接人即数据代表，不明确的业务信息，他们无法解答的，让他们协助拉源业务的人员帮忙解答；

## 2.4 数据核对

通过XSHELL，进入数据库查看数据情况

1）Desc查询表结构，查看目前库中字段名称和接口文档中的是否一致。

2）首先需确认数据状态，即是否正常更新，许久未更新数据需要核实该表是否已经废弃，进而确定是否需要调研，如果是天分区可按照pt\_d排序核查近一个月数据。如果某段时间数据量突然大幅变动要问清楚原因。

3）Select 全字段查看数据情况（由于数据较多，可用limit 限制数量），切记如有分区的数据一定要在语句中限定分区。

4）Select 字段,count 查询每个字段的情况

## 2.5表级调研

### 2.5.1 常见问题

1）（对于非日志表的）一定要问清楚主键，并核查清楚，先询问业务方，在通过如下方法验证：

select count(字段),count(distinct(字段)) from 某表，如数据量大可增加分区限制

在设计DWD表时要注意主键字段一定非空，如果ODS的记录中该字段出现空值，那么该条记录就无法进入DWD，此时就要咨询需求方，是否需要这些数据。如果不需要则把该字段设置成主键、如果需要则看下是否还有其他可以设置主键的字段，如果没有那么DWD表可开发成无主键的表。

2）调研清楚该表数据的产生场景、增量还是全量、用途等。

3）表和其它表的关联关系

### 2.5.2 维护文档

https://szxsvn06-ex:3690/svn/TC\_EMUI\_DataService\_SVN/02.项目目录/03.BigData Platform项目/06.数据架构/02.ID调研/02tableLvlTask.xls

|  |  |
| --- | --- |
| **表格名** | **填制方法** |
| 数据平台源表主干名 | 需求中会列明 |
| 表编号 | 自动生成 |
| 源系统业务标识名 | 一般根据ODS后面的字母，再去业务拉通中机型比较核对找到 |
| 源系统表名 | 根据数据字典提供填写，业务侧表名。目前一般写成了数据平台源表主干名 |
| 源表中文名 | 咨询数据代表或业务人员 |
| 所属业务子模块 | 目前已用vlookup自动生成 |
| 业务大类 | 目前已用vlookup自动生成 |
| 表类型 | 见下方 |
| 近源层加载算法类别代码 | 全量或增量，数据字典中会提供 |
| 是否入整合模型层标志 | Y |
| 是否属于一期范围 | 目前默认N |
| 归属整合模型层主题 | 添加归属的主题域 |
| ODS入数规则 | 自己从库中查询的增量还是全量数据，方法见下方。 |
| 关联关键字 | 外键 |
| 调研结果及备注 | 1.数据产生场景：  2.业务数据生成/上报周期：  3.推送到大数据平台周期：  4.每周期文件个数：  5.字段数：  6.每周期数据条数：  7.每周期文件大小：  8.数据接入时间：  9.最小天分区：  10.近一月是否正常更新：  11.近一个月每日数据量：  12.样例： |
| 问题及解答 | 写需要调研的问题及对方的解答 |
| 备用字段5(当前责任人) | 调研人员 |
| 调研咨询人 | 数据代表或业务人员 |

增量全量判断方法：

select

pt\_d,count(\*)

from biods.ods\_push\_dev\_login\_new\_dm

where pt\_d>='20170510'#可查最近一段时间

group by pt\_d;

则跑出20170510及其以后的数据量，看每天的数据量是逐天增加（全量数据）还是忽大忽小（增量数据）

## 2.6 字段级调研

### 2.6.1 常见问题

1）用户ID确认：核查清楚是up\_id还是业务编号，确定好长度，如有加密需解密确认。

2）对于日期、金额、重量等相关字段要调研清楚单位。

3）flag指示器类字段要验证清楚。比如0、1分别指代什么；如果还有其他数字比如2、3、4等则该字段应该是非flag字段。

4）主外键字段调研清楚。

5）调研过程中如遇到加密信息（比如up\_id,imei），需问清楚加密及解密方式。

6）字段名和字段储存实际内容要核对清楚，比如：版本和版本号可能会容易混淆。

7）遇到一些专业名词，自己要先查一下其含义，再进行调研。

8）码值一定要调研清楚。

9）表每个字段的详细说明  
10）如果是枚举值字段提供全部的对应关系  
11）是否涉及个人数据，是否有敏感数据，那些是，是否需要特殊处理。个人数据可见2.6中维护方法  
12）那些字段可以为空，那些字段不可以为空

13）查询每个字段最大长度，核查问题数据，为后续设计做准备

### 2.6.2维护文档

https://szxsvn06-ex:3690/svn/TC\_EMUI\_DataService\_SVN/02.项目目录/03.BigData Platform项目/06.数据架构/02.ID调研/03fieldTask.xlsx

|  |  |
| --- | --- |
| **表格名** | **填制方法** |
| 数据平台源表主干名 | 和表调研保持一致 |
| 数据平台源字段名 | 库中ODS表字段名称 |
| 字段序号 | 从1开始的自然数 |
| 源系统标识名 | 和表调研“所属业务子模块”保持一致 |
| 源系统表名 | 和表调研保持一致 |
| 数据平台表中文名 | 和表调研保持一致 |
| 源系统字段名 | 通过desc进入库中查询表字段情况 |
| 数据平台字段中文名 | 根据数据字典填写 |
| 数据平台源系统数据类型 | 通过desc进入库中查询 |
| 允许NULL标志[N/空] | 1）先根据数据字典填写（如数据字典没有写则需要询问业务人员）需要询问业务人员  2）如数据字典写的是允许为空，则填Y  3）如数据字典写的是不允许为空，则需要进数据库查询实际数据情况 |
| 默认值 | 字段中是否有默认值，如有则填；如没有则空。 |
| 唯一索引或主键字段[UI1,UI2/空] | 字段数据代表或业务人员、再通过数据库核对 |
| 调研结果及备注 | 自己调研出的一些有用信息可填此处 |
| 问题及解答 | 整理好每个字段的问题，同业务人员沟通 |
| 是否属于个人数据 | 根据个人数据定义填写 |
| 个人数据类型 | 根据个人数据定义填写 |
|  |  |

## 2.7 个人数据维护

### 2.7.1 定义：

使用该数据或结合其他数据可以定位到某个自然人的数据，就是个人数据，我们要关注如下列表的个人数据，同时，**务必注意IMEI、IMSI、用户行为数据、用于画像的数据**，这些都会受影响。

|  |  |  |
| --- | --- | --- |
| **分类** | **个人数据类型** | **备注** |
| 通信相关 | 电话号码 | 电话号码、传真号码以及包含电话号码的其他号码，如SIP号码、移动台国际用户识别码（MSISDN） |
| 邮件地址 | 电子邮件地址，语音邮箱地址 |
| 通讯录 | 联系人列表（群组名称和成员）、好友群组、黑名单群组、地址薄等 |
| 通信类应用账号 | 即时通讯应用账号、社交应用账号、实名制BBS账号及对应口令 |
| 终端设备识别号 | 能唯一的确定一个设备（例如计算机、网卡、手机、智能卡），例如MAC地址、国际移动用户识别码（IMSI）、国际移动设备识别码（IMEI） |
| 原始通信内容 | 通信双方（只要其中一方涉及自然人）之间的实际通信内容，包括语音类、短信/彩信类、传真类、数据业务（如即时消息、Email、视频通信、网页浏览等）类等 |
| 个人身份 | 权威社会识别号 | 身份证号码、社会/保险保障号码（社会号码）、驾照、护照号码、签证授权编号等 |
| 生辰信息 | 出生日期、出生地点、年龄以及其他可以计算出生日的信息（例如星座和生肖）另外：出生信息包括公历和其他常见历法计算的出生信息 |
| 家庭信息 | 婚姻状态、家庭成员信息（姓名）、家庭构成 |
| 其他 | 姓名、性别、签名、年金、个人特点、工作类别、工作地点、活动类型、专业背景和成就、行为偏好、照片、声音、地理标记数据、交通数据、种族、国籍、民族、政治观点、宗教信仰、哲学或道德准则、工会成员关系、政治党派、宗教组织、个人识别号码、健康状态（健康诊断信息、残障信息）、医疗记录、遗传数据、生物特征数据（指纹、虹膜等）、性取向、犯罪记录、入狱记录、纪律处分、兵役等数据 |
| 位置和拓扑 | 逻辑地址 | IPv4和IPv6地址、基站或小区ID（CELL ID）、WLAN无线接入点ID（WLAN AP ID）等逻辑地址及其他开放端口 |
| 个人位置信息 | GPS坐标、GPS轨迹 |
| 地址及其编码 | 家庭地址、临时地址、工作地址、邮政编码、账单地址 |
| 资产消耗 | 资产状态 | 银行存款、信用额度、信用记录、账户当前余额及历史余额 |
| 资产账号 | 银行卡账号、网银账号、网络虚拟资产账号（例如支付宝）、信用卡号 |
| 消费和支付信息 | 消费金额、消费明细（账单、信用卡账单）、消费方式、发票ID等 |
| 收入信息 | 收入金额和构成等 |
| 个人财务信息 | 罚款信息、财务状态/数据、个人资产、银行账户信息等 |
| 历史记录 | 私有浏览记录 | 用户浏览URL地址列表、cookies、sessionID、flash cookies、HTTP日志、浏览记录等 |
| 操作习惯和应用的使用频率 | 最常用的app（如某手机用户使用最多的是浏览网页和玩游戏）、每天的通话次数。 |
|  | 注：使用频率本身不是个人数据，但是如果和国际移动用户识别码（IMSI号）、IP地址等信息一起使用，可以识别数据主体，就属于个人数据。收集用户的操作习惯和使用频率通常是用于分析用户的行为，一般都可以对应到具体个人。 |

### 2.7.2 SVN地址

https://szxsvn06-ex:3690/svn/TC\_EMUI\_DataService\_SVN/02.项目目录/03.BigData Platform项目/06.数据架构/02.ID调研/personal\_data/personal\_privacy\_data\_summary.xlsx

# 3.设计流程

设计完毕后交付开发的是以下的映射表

橙色部分根据表级调研内容填写

黄色部分根据PDM内容填写

****

## 3.1 DWD表命名

整合层的表命名为DWD\_域名\_表名称\_结尾。

（1）域名为下表中的主题缩写。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **主题中文名称** | **主题英文名称** | **主题缩写** | **物理表（数量）** |
| 1 | [参与者](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'01参与者'!A1) | Party | PTY | 47 |
| 2 | [产品](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'02产品'!A1) | Product | PRO | 3 |
| 3 | [协议](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'03协议'!A1) | Agreement | AGT | 2 |
| 4 | [设备](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'04设备'!A1) | Equipment | EQP | 11 |
| 5 | [事件](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'05事件'!A1) | Event | EVT | 101 |
| 6 | [位置](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'06位置'!A1) | Location | LOC | 1 |
| 7 | [营销](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'07营销'!A1) | Campaign | CAM | 16 |
| 8 | [内容](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'08内容'!A1) | Content | CON | 32 |
| 9 | [财务](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'09财务'!A1) | Finance | FIN | 1 |
| 10 | [销售](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'10销售'!A1) | Sales | SAL | 12 |
| 11 | [Online](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'11online'!A1) | Online | Onl | 32 |
| 12 | [参数](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'12参数'!A1) | Refference | REF | 2 |
| 13 | [代码](file:///F:\数据架构\04.数据仓库LDM\数据仓库整合层-元数据2.0.xlsx#'13代码'!A1) | Code | CDE | 181 |

一般来说订单归属Sal,广告归属Cam,跟订单和广告相关的表基本都沿用上述主题命名，日志类一般归于Evt，用户的基础信息归于Pty,手机终端相关的产品信息  
归于Eqp，一般大家可以到当前开发目录下参考下，新入的表具体归入哪个主题下。主题后面的二级命名一般沿用业务习惯命名，比如应用市场hispace,视屏video,广告adv,  
Vmall一般也是vmall,二级命名后面一般借鉴源表名能够描述清楚表的意思或用途即可。

（2）结尾分为四类：hm,dm,ds,ds\_his，hm/dm为快照表，保留每小时/每天分区。ds为全量表，保留历史每个月最后一天以及最近一个月的数据。ds\_his表为拉链表，end\_date=’99991231’分区存放全量最新数据，其他分区为历史数据。

目前较多的表结尾是\_ds或\_dm，dm一般指日期增量的表，ds一般指全量表。  
表的全部字长不建议超过40位，如果有超过可适当缩减。

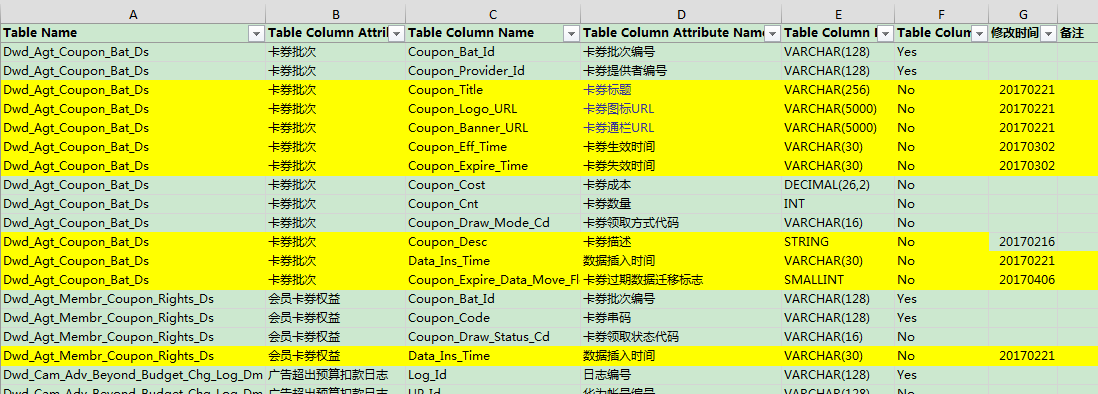
设计时表命名请注意，只要是全量(包含没有天分区全量数据，累全量逻辑，Mapping每天全量直抽,拉链历史表)表后缀应该为\_ds（拉链历史表为\_ds\_his)，而不是\_dm,增量数据直抽没有累全量逻辑的才命名为\_dm

**此处可查询DataOne开发手册2.0整合层建表规范**

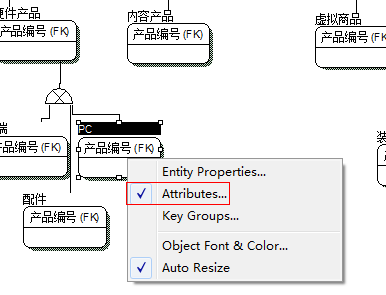
## 3.2 DWD字段命名

字段命名原则上遵循词根，同时字段命名尽量考虑和已入仓的整合层pdm表及字段的关系，同样内容的字段或者同一主题下的表的字段命名尽量一致，简单的来说就是

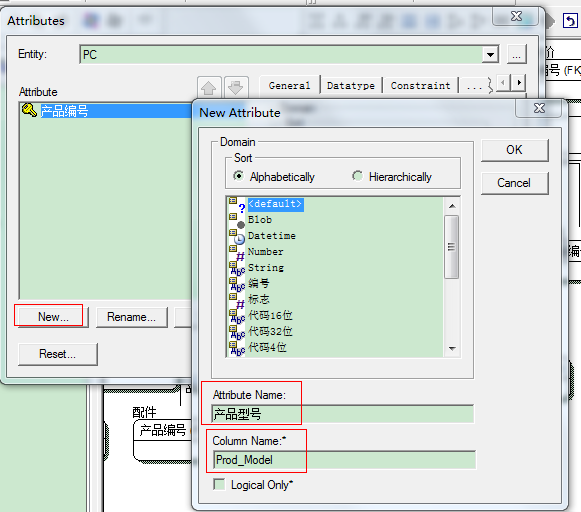
1）先到PDM里找类似的表有没有这样的字段名称，有直接用字段名称，中文注释，字段类型。数据字典中给的中文名和PDM中的不一定相符，所以需要从PDM中选择一个和ODS表意思一致的中文表名。

****

2）没有的话，可以先用erwin里定义进去系统自动生成，具体方法：打开“华为消费者BG云服务大数据数据架构逻辑模型\_内部使用”Erwin Model文件（此文件和AttrC2E\EntityC2E结合使用），点击一个实体，右键单击，选择Attributes，

****

在出现的对话框中点击New，在弹出对的New Attribute的Attribute Name中输入要翻译的中文字段名称，然后按Tab键，Column Name会自动生成英文名。如果没有正常生成英文名，那么说明没有该词根，需要核实是否中文不合适，如果在PDM中没有找到合适的中文名那么就需要增加词根。

****

3）字段命名格式，大致遵循以下规则：  
 '编号类：以“编号”(\_ID) 结尾，指数据项被赋予具有一定规律、易于计算机和人识别处理的符号，形成编号元素，每个编号均对应特定的业务对象，适用于具有特定编号规则的信息项，例如消费者编号、产品编号、设备编号等；

代码类：以“代码”(\_Cd) 结尾，指数据项以对各取值范围进行结构化处理的方式体现，每个代码取值只代表业务对象的一种可能性，适用于有明确分类、分组或多取值的信息项，例如行业代码、语言代码、性别代码、消费者类型代码等；

指示器类：以“标志”(\_Flg) 结尾，指数据项仅包括“是”、“否”两种结果，1 表示是， 0 表示否，例如：外部产品标志，花粉用户标志，会员标志等；

文本类：指数据项以纯文本的形式体现，适用于很难结构化的说明描述类信息，例如客户名称、产品描述等；一般以“描述”，“名称”，“说明”等结尾

金额类：一般以“金额”(\_AMT)，“余额”(\_BAL) 结尾，指数据项以货币金额的形式体现，适用于各类财务信息，例如违约金额、账户余额等；

数值类：指数据项以整数或小数的形式体现，适用于各类以数量反映的信息，例如客户办公面积、流通股数等；

比例类：一般以“率”(\_RAT) 结尾，指数据项以比值的形式体现，适用于各类比率信息，例如利息税率、浮动利率浮动幅度等；

日期类：一般以“日期”(\_Date) 结尾，指数据项以特定日历日的形式体现，适用于各类日期信息，例如开户日期、注册日期等；

时间类：一般以“时间”(\_Time)结尾，指数据项以特定时分秒的形式体现，适用于各类时间信息，例如机构开始营业时间；

日期时间类：一般以“时间”(\_Time)结尾，指数据项以特定日历日加上特定时分秒的形式体现，适用于各类需同时记录日期和时间的信息，例如交易时间。

4）代码字段除特殊情况都要求是VARCHAR(16)，同时要求必须先到通用代码表中通过字段命名和码值内容确认有没有现成的代码内容，有直接使用当前的代码字段  
命名，没有则需要在通用代码表中增加

通用代码表位置：https://szxsvn06-ex:3690/svn/TC\_EMUI\_DataService\_SVN/02.项目目录/03.BigData Platform项目/06.数据架构/13.ETL开发/通用代码表.xlsx

## 其他

1. 和需求人沟通确定增量、全量、累全量
2. 开始跑数据的日期
3. PDM如果有需要更新的内容，在更新后，需要再去维护修改列表。链接地址为：https://szxsvn06-ex:3690/svn/TC\_EMUI\_DataService\_SVN/02.项目目录/03.BigData Platform项目/06.数据架构/12.模型设计/模型文件/修改列表.xlsx