|  |  |
| --- | --- |
| 产品名称 | 密级 |
|  | 机密 |
| 产品版本 | 共页 |
|  |

数据架构方案及与平台工具间接口说明书

*(Title must be in both Chinese and English to facilitate English speaking users' search为方便英语用户检索，此处必须提供中英文版)*

|  |  |  |  |
| --- | --- | --- | --- |
| 拟制 | 何远洪/201904 | 日期 | 2016-05-10 |
| 审核 |  | 日期 | yyyy-mm-dd |
| 批准 |  | 日期 | yyyy-mm-dd |



华为技术有限公司

版权所有 侵权必究

修订记录

|  |  |  |  |
| --- | --- | --- | --- |
| 日期 | 修订版本 | 修改描述 | 作者 |
| 2016-05-10 | 1.0 | 初稿完成 | 何远洪（201904） |
| yyyy-mm-dd | 1.1 |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

目 录

[1 简介 5](#_Toc451951179)

[1.1 范围 5](#_Toc451951180)

[1.2 特性概述 5](#_Toc451951181)

[1.3 假设和约束 5](#_Toc451951182)

[2 数据架构方案 5](#_Toc451951183)

[2.1 数据源接入方案 5](#_Toc451951184)

[2.1.1 数据源接入目标 5](#_Toc451951185)

[2.1.2 数据源管理系统 5](#_Toc451951186)

[2.1.3 数据源属性 6](#_Toc451951187)

[2.1.4 数据源上报 6](#_Toc451951188)

[2.1.5 数据源监控 8](#_Toc451951189)

[2.1.6 数据源可视化 8](#_Toc451951190)

[2.1.7 数据源质量要求 8](#_Toc451951191)

[2.2 核心/业务模型 9](#_Toc451951192)

[2.2.1 核心模型 9](#_Toc451951193)

[2.2.2 业务模型（应用市场） 11](#_Toc451951194)

[2.3 数据架构与各部件间API 12](#_Toc451951195)

[2.3.1 数据流图 12](#_Toc451951196)

[2.3.2 业务场景及信任边界说明 13](#_Toc451951197)

[2.3.3 各部件间接口描述 14](#_Toc451951198)

[3 DFX设计(非功能设计) 15](#_Toc451951199)

[3.1 安全需求设计 15](#_Toc451951200)

[3.1.1 传输安全需求 15](#_Toc451951201)

[3.1.2 存储安全需求 15](#_Toc451951202)

[3.1.3 模型数据权限需求 15](#_Toc451951203)

[3.2 生命周期设计 16](#_Toc451951204)

[3.3 隐私设计 16](#_Toc451951205)

[3.4 法律法规 16](#_Toc451951206)

[4 备注和附录 16](#_Toc451951207)

Keywords 关键词： *Provide Keywords in both Chinese and English to facilitate English speaking users' search. 为方便英语用户检索，此处必须提供中英文版*

Abstract 摘要：*Provide Abstract in both Chinese and English to facilitate English speaking users' search. 为方便英语用户检索，此处必须提供中英文版*

缩略语清单：

| 缩略语 | 英文全名 | 中文解释 |
| --- | --- | --- |
| TCC | Task Center Control | 任务调度中心 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 简介

## 范围

本文描述了数据从接入到上层应用中数据源及核心模型与平台工具间的接口和数据设计方案。

数据源接入及数据模型设计方案覆盖ETL(DataLoder)、ODS(数据源)、核心\业务模型；周边涉及运维人员、业务服务器、监控工具、元数据管理系统、数据质量&生命周期管理系统、TCC(任务调度系统)、上层应用(报表、推荐、画像、等)等内部相关人员及部件。

## 特性概述

本部件涉及数据源的接入、核心/业务模型的设计。

1.数据源运维自助接入(数据推送、质量监控、入库)

2.数据源可视化。

3.核心/业务模型。

## 假设和约束

该部件用户为研发、运维等公司内部人员；

访问该系统时，必须接入公司iaccess或者使用运维vpn；

该部件的数据来源为数据平台的数据仓库，该仓库仅对部署在同一网络环境内的内部系统开放访问权限，不直接对外提供任务服务。

# 数据架构方案

## 数据源接入方案

### 数据源接入目标

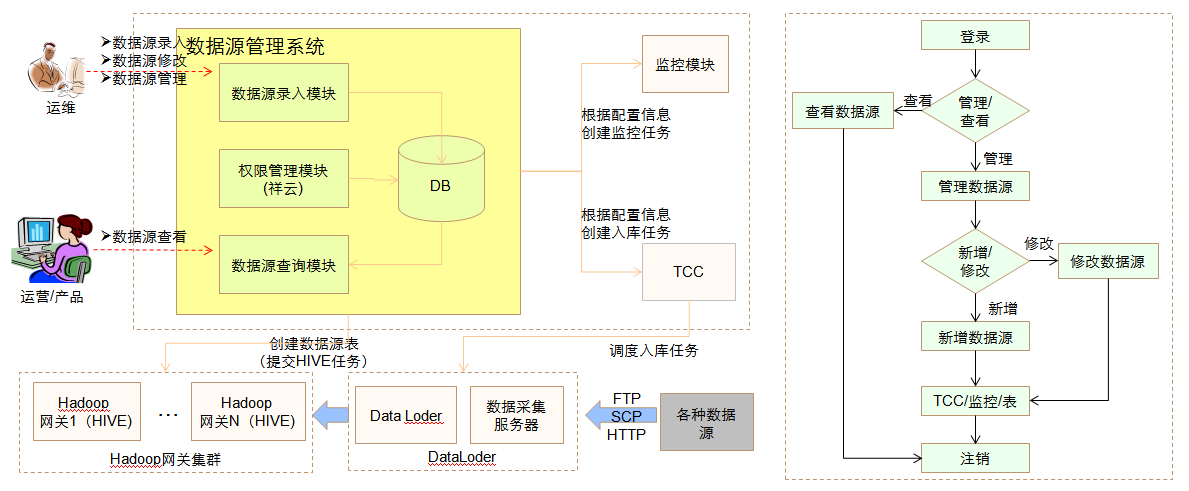
1. 确保接入数据源的完整性
2. 有效管理数据源并提数据源的可视化查询页面

### 数据源管理系统

1. 概述

该系统用于数据源接入BI后的资产管理，通过运维录入数据源相关信息后生成数据源的监控和入库任务，同时业务及产品可以通过该系统查看已接入的数据源。业务数据源变更后也可须通过该系统修改管理。

1. 设计方案



1. 接口（WEBUI）
   1. 数据源录入

通过WEBUI输入数据源信息，具体查看下述章节中数据源属性

* 1. 数据源修改

通过WEBUI页面修改接入的数据源信息

* 1. 数据源查看

查看已接入数据源信息

1. 外部模块交互
   1. 监控模块，生成/修改监控任务
   2. TCC模块，生成/修改调度任务
   3. HIVE集群，创建/修改HIVE数据源表

### 数据源属性

1. 数据来源

用于区分数据来自的业务，便于业务在BI系统中查询已接入BI的数据。

* 1. 包括数据所属部门、业务，数据存储的局点（中国、海外），是否存储在第三方。
  2. 数据来源类型：运维日志(Tomcat)、业务日志、数据库

1. 数据内容

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 字段名/列数据名 | 数据类型 | 是否NULL | 索引 | 默认值/枚举值 | 隐私数据 | 加密数据 | 说明 |
| id | bigint(20) | NO | PRI |  | NO | NO | auto\_increment |

### 数据源上报

1. 上报接口

当前业务上报的数据源主要有四个：云服务的业务服务器、移动终端、网站WEB客户端，跟第三方合作伙伴引入的数据。并通过6类数据采集接口上报到云侧大数据平台，并进行统一清洗、规整、拉通归一、并面向不同应用领域进行洞察分析满足不同大数据应用场景，如下定义：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 接口类别 | 数据采集接口 | 定义 | 用途 | 数据格式 |
| 业务服务器 | 接口1 | 云侧服务器的数据采集接口 | 实时事件流采集接口（含文件推送接口） | 是否压缩，压缩格式(推送接口可配置zip或lzo) |
| 端侧 | 接口2 | 网站web客户端数据采集接口 | JS/URL方式上报接口 | 无 |
| 接口3 | 移动端侧BI SDK数据采集接口 | Bd reporter SDK上报接口  HiAnalytics SDK上报接口 | Zip |
| 接口4 | 移动端侧日志文件上报接口 | 文件上传接口 | 自定义 |
| 第三方 | 接口5 | FTPS接口 | 第三方把数据文件上报到FTPS | 自定义 |
| 接口6 | PULL接口 | 从第三方接口拉取数据 | 无 |

1. 上报数据说明
   1. 数据上报使用的接口
   2. 非(准)实时接口上报周期信息
      1. 上报周期
      2. 上报触发时间点
      3. 上报结束时间点
   3. 数据上报服务器信息
      1. 接口部署机器信息
      2. 非(准)实时接口部署机器信息时
         1. 日志文件所在服务器、目录信息、日志文件名
         2. 数据库及表信息
   4. 对接的BI服务器信息
      1. 对接服务器IP
      2. 非(准)实时接口推送目录信息

### 数据源监控

通过运维录入的数据源信息，生成数据源的文件、任务、指标的监控。数据异常时，短信及邮件通知运维及业务责任人。

### 数据源可视化

通过查询界面，可以查询业务接入的数据的整体概况，然后可向下钻取到具体的字段属性。

### 数据源质量要求

1. 业务对录入的接入数据源信息准确性负责，确保录入的数据源信息准确无纰漏。
2. 大数据平台对数据源录入数据规则进行约束，以输入的规则作为数据清洗的规则，对数据源接入的质量根据业务录入信息进行清洗。
3. 数据平台提供数据源接入的完整性核查机制，根据录入信息检查数据源完整性，监控接入数据与配置信息异常时，发送告警通知业务运维
4. 业务版本升级、接入服务器扩容或下线涉及到数据源变化的场景，业务部门需要在数据源接入管理中修改相关信息，确保数据接入的完整性，对数据接入的完整性负责。

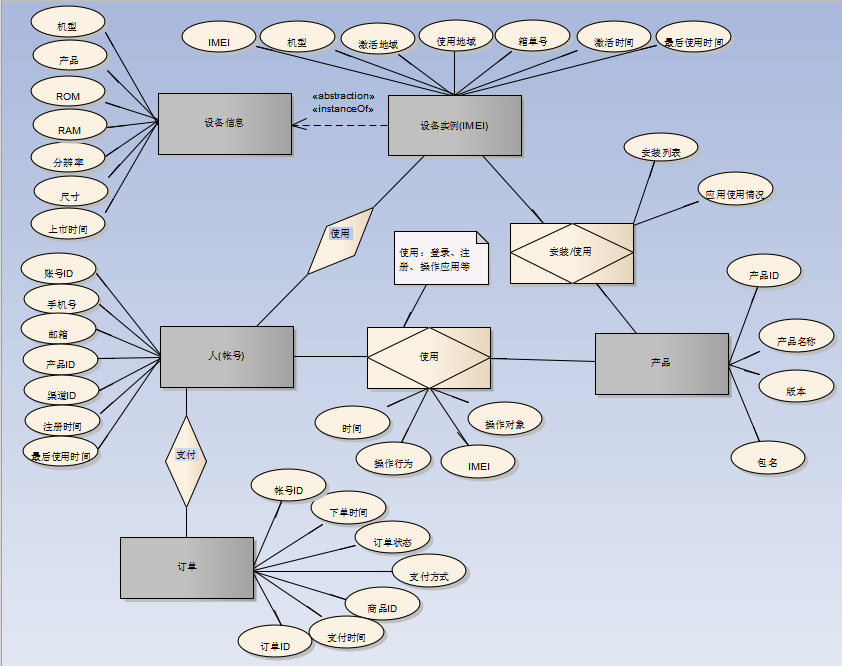
## 核心/业务模型

### 核心模型

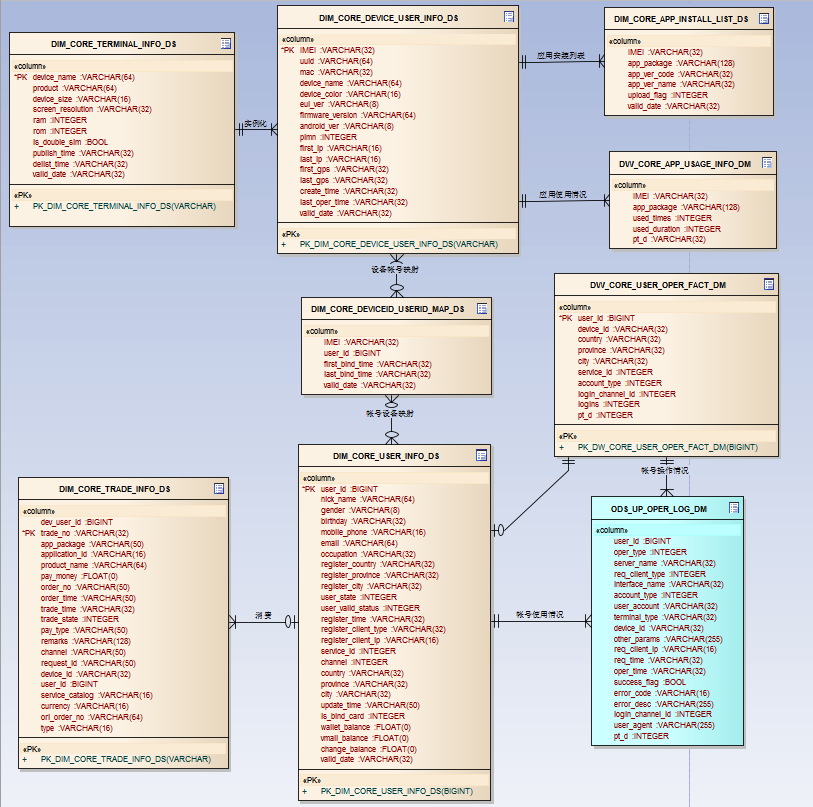
数据平台服务于消费者BG，支撑终端销售，EMUI运营及体验改进。以消费者(设备/帐号)为基础，整合拉通EMUI业务的数据建立核心模型。

本文除把设备(IMEI)和帐号外，把各业务中都使用的通用数据纳入到核心模型，其中包括设备信息，设备状态，产品信息，产品使用情况，订单信息，及帐号的操作行为；未在本文中描述，但需纳入核心模型的表包括人工维护的，所有业务都会使用到的其他映射表，例如：IP地域映射表，PLMN运营商映射表等。

1. 逻辑模型



1. 物理模型



注：蓝色部分为纳入核心模型的关键数据源

涉及关键数据源：

ODS\_UP\_OPER\_LOG\_DM

ODS\_UP\_USER\_DEVICE\_INFO\_DM

ODS\_UP\_USER\_INFO\_DM

ODS\_PUSH\_DEVICE\_TOKEN\_DM

ODS\_PUSH\_TRS\_REQ\_LOG\_DM

ODS\_TRADE\_ACCOUNT\_DM

ODS\_TRADE\_BANKCARD\_DM

ODS\_TRADE\_TRANSACTION\_INFO\_DM

ODS\_HOMECLOUD\_USER\_PREFERENCE\_DATA\_DM

ODS\_HOTA\_UPDATE\_DEVICEINFO\_DM

ODS\_PSI\_5\_IMEI\_DM

ODS\_PSI\_IMEI\_LIST\_DM

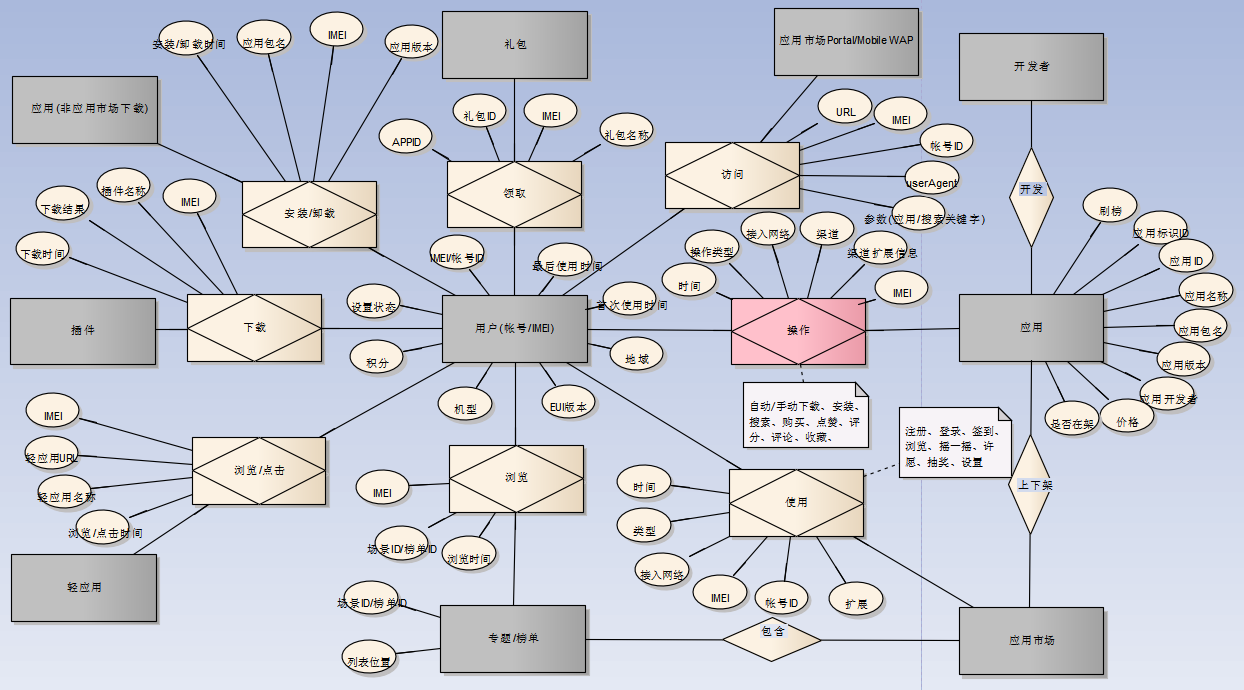
ODS\_PSI\_PACKING\_INFO\_DM

T\_APPA\_HM

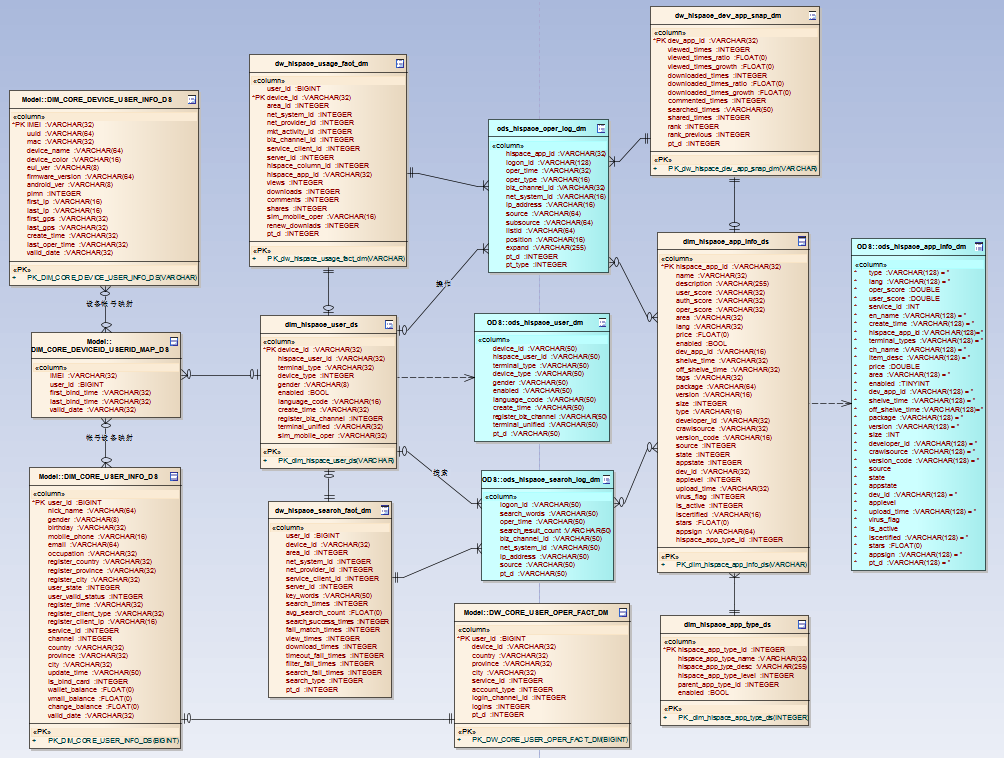
ODS\_EMUI\_HI\_ANALYTICS\_DM

### 业务模型（应用市场）

1. 逻辑模型



1. 物理模型



注：蓝色为纳入业务模型的关键数据源

涉及的业务关键数据源：

ods\_hispace\_oper\_log\_dm

ods\_hispace\_search\_log\_dm

ods\_hispace\_user\_dm

ods\_hispace\_app\_info\_dm

ods\_hispace\_app\_scenewaplink\_dm

ods\_hispace\_score\_log\_dm

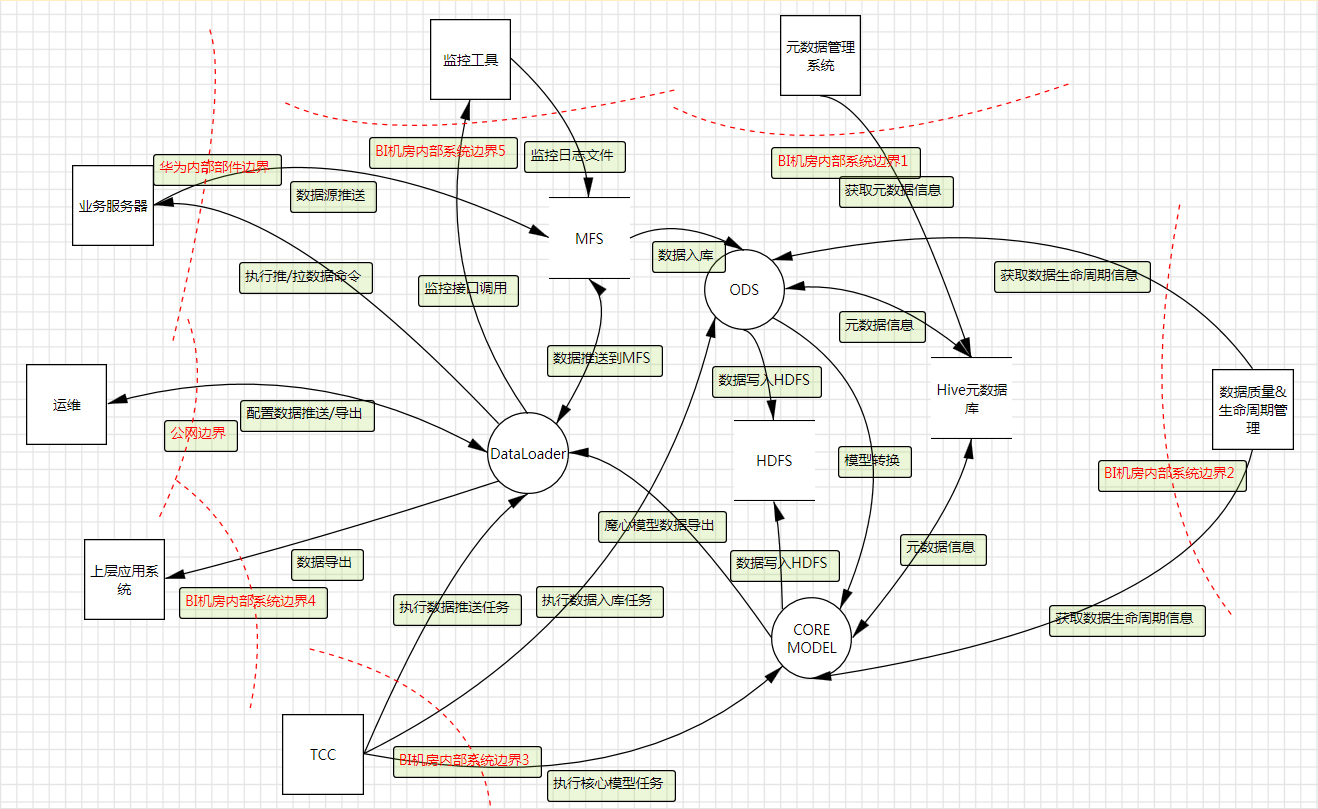
ods\_hispace\_gift\_info\_dm

ods\_hispace\_tab\_id\_dm

ods\_hispace\_app\_renew\_monitor\_dm

## 数据架构与各部件间API

### 数据流图



数据流图元素及说明如下表所示：

|  |  |  |
| --- | --- | --- |
| **元素** | **符号** | **描述** |
| 外部交互方 |  | 能驱动系统业务，但不受系统控制的人和物(如用户，管理员，第三方系统等).通常表示目标系统的输入/输出 |
| 处理过程 |  | 一个过程执行一个任务时的逻辑表示，例如Web Server 、ftp server、LMT server |
| 数据存储 |  | 数据存储表示文件、数据库、注册表项、内存等 |
| 数据流 |  | 数据在系统中的移动方向，如网络通讯、共享内存、函数调用等。 |
| 信任边界 |  | 当数据流穿越不同的信任级别(区域)时，就存在信任边界，例如从用户态到内核态，从客户端到服务端等。 |

### 业务场景及信任边界说明

涉及信任边界的业务场景：

1. 业务运维登录后配置自助数据源接入，BI运维登录后配置统计数据导出
2. ETL(DataLoader)工具从业务服务器推/拉数据，业务服务器把数据源推送到MFS服务器上后入库
3. 监控工具监控推送到MFS的文件（个数，大小，环比，关键指标值）
4. 元数据系统通过数据架构的元数据库(Hive Server元数据库)获取元数据信息
5. 数据质量&生命周期管理系统读取数据源及核心模型中的数据及生命周期信息
6. TCC(任务调度系统)调度数据源接入及模型转换的任务执行
7. 统计数据通过导出工具导出数据到上层应用系统

### 各部件间接口描述

|  |  |  |
| --- | --- | --- |
| 接口名称 | 涉及的模块 | 描述 |
| LRCI | OMA调用CLRC | 此接口完成XXX功能，。。。 |
| 自助数据源接入配置接口 | 运维人员、DataLoader、文件服务器(MFS)、质量监控工具、TCC、业务服务器 | 运维人员登录DataLoder，配置待接入数据源信息后，DataLoder创建数据接入目录，生成TCC(任务调度中心)任务定时触发数据推送，并生成监控任务监控推送的数据源。 |
| 数据源接入接口 | DataLoader、文件服务器(MFS)、TCC | TCC定时执行数据源推送到文件服务器后入库 |
| 监控接口 | 质量监控工具、文件服务器(MFS) | 监控数据源文件个数，大小，环比，关键指标值 |
| 元数据提取接口 | 元数据管理系统、Hive Server 元数据库 | 元数据库系统获取Hive Server元数据库中的元数据 |
| ？？？ | 数据架构、数据质量&生命周期管理系统 | 数据质量&生命周期管理系统从数据架构中获取数据生命周期信息 |
| TCC调度任务接口 | TCC、各部件模块 | TCC调度各部件模块脚本执行 |
| 数据导出接口 | DataLoder、TCC、HIVE、上层应用系统 | 数据导出接口把统计数据从HIVE导出到上层应用系统的数据库中 |

#### 部件间接口参数

1、自助数据源接入配置接口

|  |  |  |
| --- | --- | --- |
| 参数 | 是否必填 | 说明 |
| 业务服务器用户@IP:端口 | Y |  |
| 登录业务服务器证书 | Y |  |
| 业务服务器数据所在目录 | Y |  |
| 日志文件名格式 | Y | 可通过日期匹配后合并 |
| 日志文件关键字 | Y | 用于生产数据源表名(英文) |
| 传输方式 | Y | SCP、FTPS |
| 传输格式 | Y | txt，gz，zip，lzo等 |
| 传输周期 | Y | 小时、天 |
| 传输开始时间 | Y |  |
| 文件大小 |  |  |
| 文件个数 | Y |  |
| 字段数 | Y |  |
| 字段名称 | Y |  |
| 字段描述 | Y |  |
| 关键指标 |  |  |
| 业务 | Y | 业务名称（应用市场等） |

# DFX设计(非功能设计)

## 安全需求设计

### 传输安全需求

所有涉及与周边部件的交互必须确保传输过程是安全的。

1. 本部件与周边部件的交互，必须使用内部ip进行，且配置有对应的白名单。
2. 本部件与用户交互的部分，采用的是web方式，必须采用https传输协议。同时要求用户必须登录运维vpn或者公司iaccess。当用户使用vpn登录时，用户请求直接经过vpn通道达到服务器。

当用户通过iaccess访问时，请求通过iaccess先到达公司研发区域的HIS反向代理，在有反向代理链接北京华为机房的nginx代理，所以要求nginx代理使用白名单方式控制访问请求，仅允许HIS反向代理和vpn通道进行访问。

1. 数据在使用DataLoder工具传输时数据必须使用安全的加密算法加密传输

### 存储安全需求

根据大数据平台与安全组确认的存储方案存储数据

### 模型数据权限需求

模型数据（包括元数据）的权限分配依赖平台解决

## 生命周期设计

## 隐私设计

## 法律法规

# 备注和附录

附录：