大数据开发规范文档

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本号 | 时间 | 作者 | 修改标志 | 备注 |
| V1.0 | 2020-10-21 | Koray | A |  |
|  |  |  |  |  |
|  |  |  |  |  |

A-新增 M-修改 D-删除

## 1.数仓模型层次定义

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **层次** | **命名** | **说明** |
| 1 | 数据贴源层 | ODS | 数据贴源层：   * 只映射COS系统上的数据，不做任何修改，起到备份数据的作用 |
| 2 | 数据仓库层 | DWD | 数据明细层：   * 对ODS层的数据进行清洗和转换，从而得出基础数据表。 |
| DWM | 数据中间层：   * 临时数据层，存储ETL中间过程数据，一般用于中间数据落地、性能调优等临时性数据落地存储。 |
| 3 | 数据应用层 | APP | 数据应用层：   * 以分析的主题对象为建模驱动，基于上层的应用和产品的指标需求，构建主题对象的全量宽表，供后续的业务使用。 |

## 2.命名规范

#### 1.数据库命名规范

规则: 项目名+公共功能

举例:  **lazada\_dw**

#### 2.表命名规范

|  |  |  |
| --- | --- | --- |
| 数据层次 | 表命名规范 | 举例 |
| ODS | ods \_cos/mysql\_主题(模块)\_描述 | ods\_cos\_item\_detail |
| ods\_mysql\_station |
| DWD | dwd\_dim/fact/tmp\_主题(模块)\_描述 | dwd\_dim\_ item\_detail |
|  |
|  |
| DWM | dwm\_主题(模块) \_描述 |  |
| APP | 项目名\_rpt\_主题(模块) \_描述 | lz\_rpt\_item\_review\_goods |

层名\_计算周期&&统计周期\_表类型\_存储类型\_表的业务含义

临时表: tmp\_功能\_表的业务含义

全量快照表: dwd\_ snap\_表的业务含义

注：

1. 表名使用英文小写字母，单词之间用下划线分开，长度不超过30个字符，命名一般控制在小于等于6级。
2. 时间粒度：使用"c"代表当前数据，"h"代表小时数据，"d"代表天数据，"w"代表周数据，"m"代表月数据，"q"代表季度数据， "y"代表年数据。
3. 对象属性，用"t"表示表，用"v"表示视图。

## 3.数据集标准规范

#### 1.数据类型规范

|  |  |  |  |
| --- | --- | --- | --- |
| **数据库** | **类型** | **源数据库数据类型** | **HIVE数据类型** |
| **MYSQL** | 字符 | CHAR | STRING |
| VARCHAR | STRING |
| TEXT | STRING |
| TINYTEXT | STRING |
| BINARY | STRING |
| VARBINARY | STRING |
| MEDIUMTEXT | STRING |
| LONGTEXT | STRING |
| 数值 | TINYINT | DECIMAL(30,8) |
| SMALLINT | DECIMAL(30,8) |
| MEDIUMINT | DECIMAL(30,8) |
| INT | BIGINT |
| BIGINT | BIGINT |
| FLOAT | DECIMAL(30,8) |
| DOUBLT | DECIMAL(30,8) |
| DECIMAL | DECIMAL(30,8) |
| 日期 | DATE | STRING |
| TIME | STRING |
| DATETIME | STRING |
| TIMESTAMP | STRING |
| YEAR | STRING |

#### 2.数据字段规范

* **命名**

小写字母、数字、下划线组成，不同单词之间用下划线分开

* **partition分区列**

Hive partition列在Data中并不存储，这会导致当以文件形式对外提供数据时，数据会有缺失，所以我们要对所有的partition列进行冗余存储。

如：当以分区列为dt字段时，我们多添加一列hp\_dt列即可。（hp：hive partition）

* **列操作**

可以修改列数据类型；新增列只能加到最后；严禁删除列。

* **注释**

注释本着简洁、详实、完整的原则，对于有业务含义的字段，在注释中需要枚举并解释其业务含义，如ods\_loan\_apidata\_order\_info.order\_status 订单状态：1待支付，2支付不成功，3支付成功；

* **类型**

日期时间等格式统一用string类型，字符串也是用string，数值的话会根据字段定义来确定，对于有小数点要求的，比如某些金额、利率，需要用到decimal类型，无小数点要求的用浮点类型double和整数类型(int / bigint）

* **时间字段格式**

Osd层的爬虫原始数据表的时间字段统一采用yyyy-MM-dd HH:ss:mm

时间分区字段的格式用yyyyMMdd形式

App层导出的统计表统一用yyyyMMdd形式

## 4.数据层级设计规范及设计原则

## 4.1贴源层（ODS）

1. 命名规范

* 表结构同步自源系统，包括表、字段和源系统保持一致.
* 数据类型根据数据类型规范要求，转换为标准三类型。

1. 分区规则

* ODS表（如ods\_cos\_item）映射COS上的商品表数据,使用时间分区(dt , hr)加载数据入库；

1. 特点

* 数据原样加载，与COS系统结构一样，数据类型按对应关系修改；
* 增量抽取源系统最新数据，在Hive端采用增量逻辑加工最新数据

(该部分数据，主要是根据源系统增量标识，如保留的Last\_update\_date字段值取每天的增量数据)

* 全量抽取源系统数据

数据量较少的表，或因部分源系统表未设计Last\_update\_date字段，数据采集应采用全量更新方式

1. 设计原则

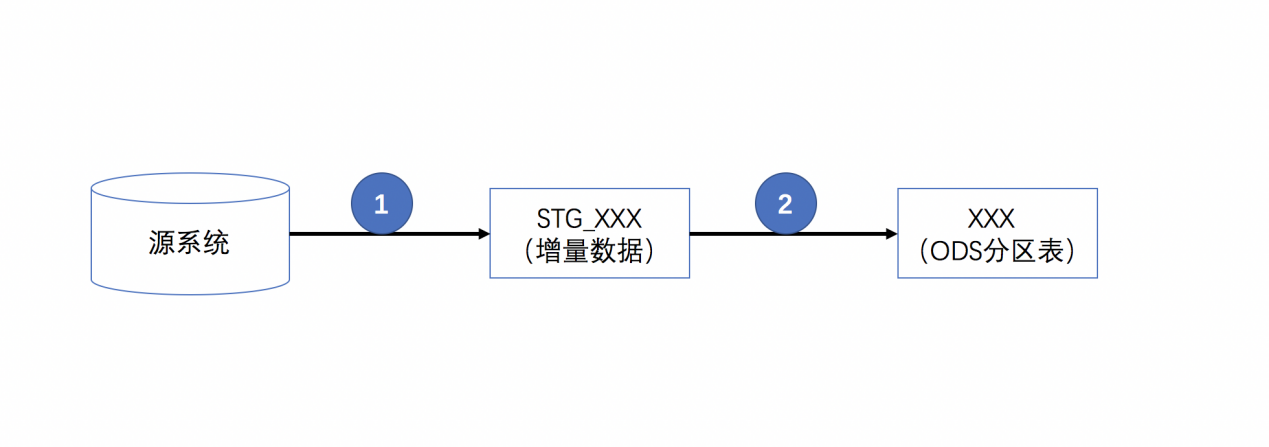
* 源系统分离原则：与源一致，不进行任何整合。
* 访问隔离原则：缓存上游数据，提供对DW关联访问。
* 与其他区域独立原则：

独立加载，不依赖其他区域是否完成，是否有错等

并发抽取控制，根据源系统情况，在调度系统中控制抽取并发数。

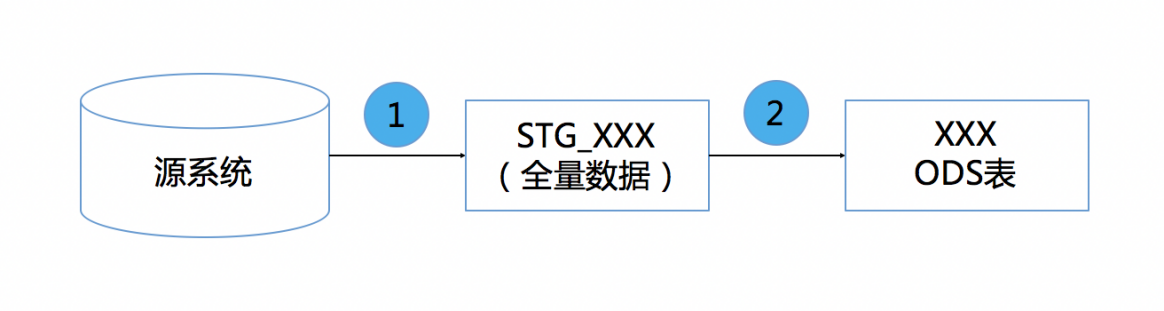
1. 加载策略

* 增量（APPEND）



1. Sqoop采用–hive-overwrite和–where参数按源表时间戳取增量数据抽取到STG\_XXX日增量表
2. STG\_XXX插入到按天分区(跑数时间)的ODS表，并存放最新3个月的分区数据

* 全量（直接覆盖）



1. 通过sqoop脚本将源表全量数据抽到STG\_XXX全量表
2. 将STG\_XXX表每天insert overwrite 到ODS表。

## 4.2仓库层（DW）

1. 命名规范

事实表：DW + 主题(模块) + 业务描述 + 后缀（如需要）（\_D/\_W/\_M/\_HIS）

维度表：DIM + 自定义名称

1. 分区规则

* DW表分区按需定义:

如:小表（全量方式）不考虑分区。

大表（增量方式）优先考虑按数据处理时间日分区（PART\_DAY）,酌情考虑按月分区（PART\_MONTH）

* DIM表原则上不作分区

特点

* 采用主从表模式，轻度关联
* 保留更新历史，增加“是否最新记录”列，方便后续层级使用
* DW表数据采用维度、度量的数据组成
* 包括：事务处理过程、累积快照

1. 设计原则

* 建模原则：事实表以基表为主、桥接表辅助（主要存放不需维度化的信息）
* 自然键保留原则：

维度表，如来源系统，产品编码、条码属性原值保留；

事实表，如来源系统，订单号、收据号退化维信息原值保留；

* 空值映射原则：源数据维度属性空值

事实表中维度化失败，维度表预设特别代理键-1，保证事实表数据完整性；

* 代理键原则：维度表与事实表增加数值主键属性，维度表与事实表中的关联采用代理键
* 维度变更原则：采用缓慢变化维，变更属性增加行，设置开始、结束时间及激活状态属性
* 数据审计原则：统计数据装载过程整理质量、完全性、验证、越界、筛选失败、记录修改时间、时间戳（抽取、清洗、一致性），ETL主要版本、分配版本、当前版本
* 数据延迟原则：事实已到达，维度表无对应维度，则为其创建一个代理键哑元素维（-1）存到事实表，当维度属性到达后，重跑覆盖

1. 加载策略

Append：

* 事实表：

根据源数据自然键信息、或更完整的键值信息，保存到事实表；

维度表：

整合维度信息添加到维度表（保留：编码、合同号、批次号的等自然键信息）。

Upsert：

* 维度表：通过提交属性信息，定位维度记录，更新记录的结束日期为源数据变更日期，激活状态更新为“否”  
  然后新增一条维度记录，开始日期为源数据变更日期，结束日期为 9999-12-31，激活状态为“是”

Delete：

* 事实表：采用覆盖分区，保留历史数据
* 维度表：设置结束日期为源数据删除日期，激活状态更新为“否”

增/全量策略：

* 数据从ODS层，通过增量加载到DW层（按天分区or按月分区），保留记录所有历史数据
* 如全量数据，每天从ODS直接全量覆盖到DW

## 4.3仓库层（DWD）

1. 命名规范

事实表：DW + 主题(模块) + 业务描述 + 后缀（如需要）（\_D/\_W/\_M/\_HIS）

1. 分区规则

* DWD不作分区

1. 特点

* 面向主题，星形数据整合
* **仅保存最新记录**
* 采用维ID、类型、度量的数据组成
* 包括：事务处理过程、周期快照

1. 设计原则

* 同DW层

1. 加载策略

全量加载：

* 数据从DW层，通过最新标识全量加载到DWD层