数据治理标准规范

2019-12-06

1 引言

1.1 概述

遵循国家、省、市标准规范要求，结合项目实际情况，制定适应项目数据特征及大数据应用要求的管理及维护体系，确保大数据的灵活可用性，适应未来数据扩展、海量数据增长及大数据发展的趋势，为项目数据资源的连通、共享、交互打好基础。本文是项目数据模型的命名，设计和管理规范。

1.2 文档目标

- 介绍主题模型命名规范
- 介绍主题模型数据类型定义规范
- 介绍主题模型布局规范
- 介绍主题模型注释及版本管理规范

1.3 适用读者

项目的参与成员，包括项目参与人员、客户参与人员、合作伙伴等希望了解本项目主题模型的人员可以参考本文档。

2 术语和定义

- **STG**：stage，数据缓冲区。
- **ODS**：operational data store 数据标准化层。
- **DWD**：data warehouse detail 数据仓库明细层。
- **DWS**：data warehouse subject 数据仓库主题层。
- **ADM**：analysis data market 数据专题层。

3 设计目标

3.1 业务目标

将基础数据作为一个公共服务，为用户提供公共数据服务支撑，帮助数据应用提升获取数据的效率，降低数据加工的深度和复杂度；提升各个产品和应用间数据的一致性。主要包括以下几方面的内容：

- 将业务系统数据同步进入到 ODPS，建立统一、一致、唯一的 ODS 数据层
- 实现通用模型层（DWD、DWS）逻辑的加工和转换。

3.2 技术目标

在满足业务目标的同时，在数据模型设计上，重点关注以下目标：

1、成本：模型设计者必须平衡性能和成本要素对数据模型的影响，尤其是海量数据情况下，在保障业务和性能的前提下，应该使用合理的数据模型方案和存储策略，尽量消除不必要的数据复制和冗余。

2、性能：模型设计者需要兼顾模型刷新性能开销、产出时间和访问性能。

3、数据一致性和数据互通：各个数据模型或者数据表之间保障数据输出的一致性，相同粒度的相同数据项（指标、维度）具有相同的字段名称和业务描述，不同算法的业务指标应显性化区分。

4、数据质量：数据模型需要屏蔽源头垃圾数据源，一方面要保障数据本身的高质量，减少数据缺失、错误、异常等情况发生；另一方面需要保障其对应的业务元数据的高质量，数据有明确的业务含义，为数据使用者提供正确的指引。

5、易用：在保障以上目标的前提下，数据用户能从业务角度出发快速找到所需的数据；能较快的掌握模型的适用场景和使用方法；能相对便捷获取数据。但是，在目标出现冲突时，在通用数据模型并不完全承载用户使用数据的易用性目标要求，数据消费产品和数据应用可以提升数据使用的易用性。

4 设计原则

- **公共处理逻辑下沉及单一：**越是底层公用的处理逻辑更应该在数据调度依赖的底层进行封装与实现，不要让公共的处理逻辑暴露给应用层实现，不要让公共逻辑在多处同时存在。
- **数据可追溯性：**处理逻辑不变，在不同时间多次运行数据结果确定不变。
- **一致性：**相同的字段在不同表字段名相同。
- **命名清晰可理解：**表命名规范需清晰、一致，表名需易于下游理解和使用。
- **成本与性能平衡：**适当的数据冗余换取查询和刷新性能，不宜过度冗余与数据复制。

5 数据分层

数据分为 5 层，每层的作用如下：

- DWD 层（整合数据资源库）：DWD 层按业务过程和业务对象整合数据，并把数据表按一定如人口、法人、办件、政策等数据域进行分类存放。
- ADM 层（应用资源库）:将数据按照分析的专题组织成多为宽表的形式存放，数据主要来源于 DWD 和 DWS 层

6 模型设计规范

6.1 STG 层模型设计规范

6.1.1表命名规范

STG 层表命名规范：stg_{业务库名}_{业务库原始表名}。

6.1.2数据存储及生命周期管理规范

数据表类型	存储方式	最长保存策略
增量表	按日分区	100 天

6.1.3字段集命名规范

字段默认使用源系统字段名称,字段名与maxcompute关键字冲突时处理规则：加一个”_col”后缀，即：源字段名_col。

6.2 ODS层模型设计规范

6.2.1 表设计规范

ODS层数据对STG层数据进行数据全/增量合并以及数据清洗和标准化动作，ODS层有两类数据表：

1、保持原始格式的全量数据表，主要是用于溯源，按照业务主键对STG表数据与前一天ODS清洗全量数据进行合并去重。

2、清洗和代码标准化，代码名称字段扩充后的标准基础数据表，标准化动作：

- 标准代码转换，按行业代码定义标准，对源系统代码进行转换，源系统代码字段及标准代码字段均保留，不能转换成标准代码的代码值在标准代码项内保存成未知值。
- 代码值扩展，对常用代码，将代码对应的名称字段扩展到表内。
- 字符格式转换，进行全半角转换，null/none/空字符串统一转换为null，字段中部分特定字段的清理。
- 时间标准化，扩充按datatime类型的时间字段。
- 身份证格式标准化，统一格式化为18位长度的身份证。

如果部份表的转换或清洗动作较多，为能溯源，可单独保留和 STG 表一致的原始表模型。

6.2.2 表命名规范

标准表：ods_{业务库简写标识}_{业务库原始表名}[_分区标识]。

原始表：ods_{单位简称}_{业务库简写标识}_{业务库原始表名} [_分区标识][_ys]。针对溯源场景使用。

6.2.3 数据存储及生命周期管理规范

数据表类型	存储方式	最长保存策略
全量表	按日分区	100/30/15 天
增量表	按日分区	永久/7200 天

根据存储成本及数据价值的需要动态调整生命周期。初始阶段保障数据的溯源排错，使用较长的生命周期存储策略，成熟阶段和稳定阶段就可以使用较短的生命周期存储策略。

6.2.4 字段集命名规范

1、字段默认使用源系统字段名称，即 stg 层字段命名。

2、有进行代码转换的字段，新增标准代码字段在原代码字段基础上增加 c_ 前缀。

3、扩展增加的代码名称属性字段在原数据项名称的基础上加上 mc 后缀。

4、标准格式化后扩充的日期字段，在原字段基础上加上_dt 后缀。

6.3 DWD 层模型设计规范

6.3.1 表设计规范

（一）数据准入

政府数据来源于多个渠道，各渠道提供的数据中，有很多相似性的数据存在，例如民政局的人口基本信息、社保局的人口基本信息、卫计委的人口基本信息等。这些相似的数据会带来使用成本，因为每个使用者都需要处理两份数据，并且对于这两份数据处理的逻辑也会存在差异。数据整合的目的，是为了更好的建立统一的数据视角来描述同一个事实，方便下游数据使用。

政府数据具有丰富性的特点，如人的基本信息、教育信息、社保信息、违法信息、健康信息；企业的注册信息、经营信息、纳税信息、变更信息、处罚信息。这些信息都散落在各个部门，把这些信息汇聚起来，提供统一的数据视图，准确客观的描述城市管理中的各个主体，也是本层需要达到的目标。

数据准入原则如下：

1） 基于 ODS 层，明确哪些数据作为主数据，哪些数据作为补充数据，并建立主数据与补充数据的关系。

2） 去除对于没有明确属性说明的信息表。

3） 去除数据质量差，数据缺失严重，无人维护的死数据。

（二）表结构设计

按照遵循事实，同时兼顾中性共享和灵活可扩展的原则，对数据进行分类合并。

1） 设置 DWD 结构时，综合考虑需要合并的数据表，选择具有业务含义和业务用途的属性，放入 DWD 中。

2） 使用代理键作为数据记录的唯一标识。

3） 记录中保留数据来源系统和数据来源表信息，方便溯源。

（三）数据更新

1） 对于总体在千万级数据量的数据，所有数据每天生成一个全量分区。

2） 对于大于千万级数据量的数据，每日增量按照业务日期放入一个新的分区。

3） 按照业务日期做数据分区

4） 对于源头提供的历史数据，需要根据业务日期提炼所涉及的分区，然后把相关分区数据和历史数据合并，重新根据业务日期建立动态分区，覆盖原有分区。

5） delta 表保留：对于需要直接同步到 ADM 的增量数据，在 DWD 层数据处理中，需要设计成永久表来保留增量数据，专门用于同步增量数据到 ADM，减少同步成本。

6.3.2 表命名规范

dwd_｛数据域｝_<数据子域>[_数据描述][_分区标识]

数据域结合行业经验如按照人口、法人、信用、政策、地理空间等等划分数据域，将数据按所属业务类别进行模型构建，数据子域是按对象及业务过程对数据域进行进一步细分。

例如：dwd_rk_jy_xsxjxx_df

- dwd：代表数据模型层次
- rkxx：代表人口信息域
- jy：代表 教育 子域
- xsxjxx：业务描述，示例表示：学生学籍信息
- df：代表每日全量分区

6.3.3 数据存储及生命周期管理规范

数据表类型	存储方式	最长保存策略
全量表	按日分区	100/30/15 天
增量表	按日分区	永久/7200 天
delta 表	按日分区	100 天

根据存储成本及数据价值的需要动态调整生命周期。初始阶段保障数据的溯源排错，使用较长的生命周期存储策略，成熟阶段和稳定阶段就可以使用较短的生命周期存储策略。

6.3.4 字段集命名规范

字段命名采用汉字拼音首字母命名。示例如下：姓名：xm；性别：xb，如遇到冲突情况时，例如杭州：hz，护照：hz，分别取冲突字段的前两个字母，杭州：haz，护照：huz。

6.3.5 例外处理规范

对于数据中出现的异常业务日期，例如 2086-9-28 日，会影响正常的分区，需要在数据处理过程中，把这类信息统一放入到 19000101 分区中。数据质量检查的作业需要每日分析出现的错误数据，并需要人工关注和审核。

6.4 DWS 层模型设计规范

6.4.1 表设计规范

DWS 层在 DWD 层的基础上进行对象的融合及汇总计算，主要包含三种类型的数据表：

1. 全局抽象的业务实体及汇总型事实表.

   DWS 全局抽象的业务实体是整合 DWD 中所有业务数据中存在的同类数据对象信息，是所有对象实例的一个全集，形成维度实体数据的主数据信息。如公安行业 DWS 层中人的实体，需要整合户籍登记过程中产生的人，也需要整合来自于出行活动中登记的人。形成一个全量的人的信息。

   汇总型事实表是对事实进行抽象分类，把同一分类下的各类明细事实进行合并，提练通用的属性和指标，如人的行为数据，会提练代表人的证件类型，证件号码，行为发生的时间，地点，行为性质等属性。

2. 面向主题的通用业务指标统计表。

   通用业务指标表是面向主题业务计算需求，按维度或维度组合对一些度量及指标进行统计计算，如人的通迅联络表统计两个人的联系次数，联系时长等。

   - 关系主题
   - 行为主题
   - 轨迹主题

3. 业务标签表

业务标签表是结合业务需求，按设定规则或通过算法进行计算，从各类基础数据中归纳出一些对象或事实的衍生属性/指标。如人的财富状况，是否昼伏夜出等。

6.4.2 表命名规范

dws_{主题域}_<主题分类>[_数据及业务描述]_[分区标识]

表命名中包括主题域，主题分类，数据及业务描述，分区标识等各部份的缩略词，如

dws_jy_jyss_xxfb_df

- dws：代表数据模型层次
- jy：代表 教育 主题域
- jyss：代表 教育设施 主题分类
- xsxjxx：业务描述，示例表示：学校分布
- df：day_full，代表每日全量分区

6.4.3 数据存储及生命周期管理规范

数据表类型	存储方式	最长保存策略
全量表	按日分区	100/30/15 天
增量表	按日分区	永久/7200 天
delta 表	按日分区	100 天

根据存储成本及数据价值的需要动态调整生命周期。初始阶段保障数据的溯源排错，使用较长的生命周期存储策略，成熟阶段和稳定阶段就可以使用较短的生命周期存储策略。

6.4.4 字段集命名规范

- 字段默认使用 DWD 层模型字段名称。
- 字段使用 ODS 层时，字段命名采用汉字拼音首字母命名。字段命名冲突时，取冲突字段的前两个字母。

6.5 ADM 层模型设计规范

6.5.1 表设计规范

在 DWD/DWS 基础数据的上进行加工汇总形成的指标数据存储分析型和加工汇总型数据。来源于标准化的各源系统的汇总、报表数据，是基础数据经过加工按一定维度汇总的指标，或分析数据。加工汇总层的数据需求来源于应用的一些共同性指标，可以是一些中间数据，这些指标的存在，可以大大提高应用系统的处理效率。

6.5.2 表命名规范

表命名规则:adm_{专题域}_[_专题分类]〈应用数据描述>[_自定义业务标签]_[分区标识]。

表命名中包括专题域，应用数据描述及自定义业务描述，分区标识等各部份的缩略词，如

adm_fr_qy_qcyj_df

- adm：代表数据模型层次
- fr: 代表 法人 专题域
- qy：代表 企业 专题分类
- qcyj：业务描述，迁出预警
- df：day_full，代表每日全量分区

6.5.3 数据存储及生命周期管理规范

数据表类型	存储方式	最长保存策略
全量表	按日分区	100/30/15 天
增量表	按日分区	永久/7200 天

根据存储成本及数据价值的需要动态调整生命周期。初始阶段保障数据的溯源排错，使用较长的生命周期存储策略，成熟阶段和稳定阶段就可以使用较短的生命周期存储策略。

6.5.4 字段集命名规范

- 字段默认使用 DWD/DWS 层模型字段名称。
- 字段使用 ODS 层时，字段命名采用汉字拼音首字母命名。字段命名冲突时，取冲突字段的前两个字母。

6.6 其他通用规范

6.6.1 域命名规范

根据不同行业领域模型的命名策略，域命名可为数据域名称的中文拼音首字母拼音或英文单词缩写。

6.6.2 数据类型规范

ODS 层的数据类型基于源系统数据类型转换，转换规则如下：

表 1 Mysql 与 Odps 数据类型映射

Mysql 数据类型	ODPS 数据类型
TINYINT/SMALLINT/ MEDIUMINT/ INTEGER / BIGINT	Bigint
FLOAT/ DOUBLE/ DECIMAL	Double
LONGTEXT/TEXT/VARCHAR/ CHAR	String
DATE/ DATETIME	String

表 2 Oracle 与 Odps 数据类型映射

Oracle 数据类型	ODPS 数据类型
numeric	ID 转换为 bigint，根据实际数据，如果是浮点数则使用 double，默认使用 bigint。
VARCHAR2/VARCHAR	String
DATE	String
CLOB	String

表 3 Sqlserver 与 Odps 数据类型映射

sqlserver 数据类型	ODPS 数据类型
INT/TINYINT/SMALLINT/INTEGER / BIGINT	Bigint
REAL/FLOAT / DOUBLE / DECIMAL /NUMERIC	Double
TEXT/VARCHAR/ CHAR	String
SMALLDATE/ DATETIME /TIMESTAMP	String

DWD 层如果是引用 ODS 层数据，默认使用 ODS 层字段数据类型。

衍生加工数据字段类型按以下标准执行：

- 字符类数据：string
- 标识类和计算求和类：bigint/double
- 时间类型数据：datetime

6.6.3 公共字段

公共字段规范：

- sjly 数据来源，填写来源表信息，多个以逗号分隔。
- yxzt 数据状态，默认为有效写入“I”，删除“D”，更新“U”。
- dw_rksj 处理时间，格式为 14 位的时间 STRING 类型。

6.6.4 唯一记录标识

DWD 及 DWS 采用唯一记录标识，ODS 不需要

- 统一对 dwd、dws 增加“唯一记录标识”字段，使用 MD5 对业务主键创建唯一 ID。
- 唯一记录标识统一命名规范：dwd 表 dwd_zjid，dws 表 dws_zjid。
- 用途：数据去重。

6.6.5 分区命名规范

- 分区字段：所有层次数据表常用时间分区为日分区，字段均命名为 dt，格式为 yyyymmdd。
- 其他时间类型分区标识如下表，时间类型格式为字符串：

分区标识	命名规范	格式类型
不分区	all	/
增量处理表	delta	yyyymmdd
小时增量	hi	yyyymmddhh
小时全量	hf	yyyymmddhh
日增量	di	yyyymmdd
日全量	df	yyyymmdd

周增量	wi	yyyymm[01-04]
周全量	wf	yyyymm[01-04]
月增量	mi	yyyymm
月全量	mf	yyyymm
季增量	qi	yyyy[01-04]
季全量	qf	yyyy[01-04]
年增量	yi	yyyy
年全量	yf	yyyy

- 其他非时间类型的分区字段命名为 fq_[分区字段描述]，字符控制在 5 个以内。

6.6.6 时间修饰规范

中文名	时间维度全称	时间维度缩写	描述
最近 1 天	1day	_1d	最近 1 天
最近 3 天	3day	_3d	最近 3 天
最近 7 天	1week	_1w	最近 7 天
最近 14 天	2week	_2w	最近 14 天
最近 30 天	1month	_1m	最近 30 天
最近 60 天	2month	_2m	最近 60 天
最近 90 天	3month	_3m	最近 90 天
最近 180 天	6month	_6m	最近 180 天
180 天以前	before 6month	_b6m	180 天以前
自然周	calendar week	_cw	自然周
自然月	calendar month	_cm	自然月
自然季度	calendar quarter	_cq	自然季度
自然年	calendar year	_cy	自然年
历史截至当日	start to day	_std	历史截至当日
自然年初截至当日	year to day	_ytd	自然年初截至当

			日
自然季度初截至当日	quarter to day	_qtd	自然季度初截至当日
自然月初截至当日	month to day	_mtd	自然月初截至当日
自然周初截至当日	week to day	_wtd	自然周初截至当日

6.6.7 代码表

1．表命名规范 dim_{业务系统}_{业务描述}

2．统一不做分区，只保留一份永久数据

3．对下游同步使用时，删除原有的数据，保留一份最新数据

4．设计维表时，遵循易用的原则，如设计有父子关系的表结构时，使用宽表设计。

5．代码表的更新规则是全量更新，不与上一个周期的数据做合并

6.6.8 临时表测试表命名规范

对于测试类的表统一在项目下创建文件名为姓名首字母的文件夹，并在下面创建测试文件夹，临时表命名规范如下：

1．中间表命名规则：mid_表名_账期(可选)

2．临时表命名规则：tmp_名字全拼_表名_账期(可选)

3．测试表命名规则：test_姓名全拼_表名

4．测试业务流程命名规则：test_姓名全拼_业务流程名称

5．测试节点命名规则：test_姓名全拼_节点名称

6．不用的测试表、临时表、业务流程和节点及时删除