数仓公共开发规范

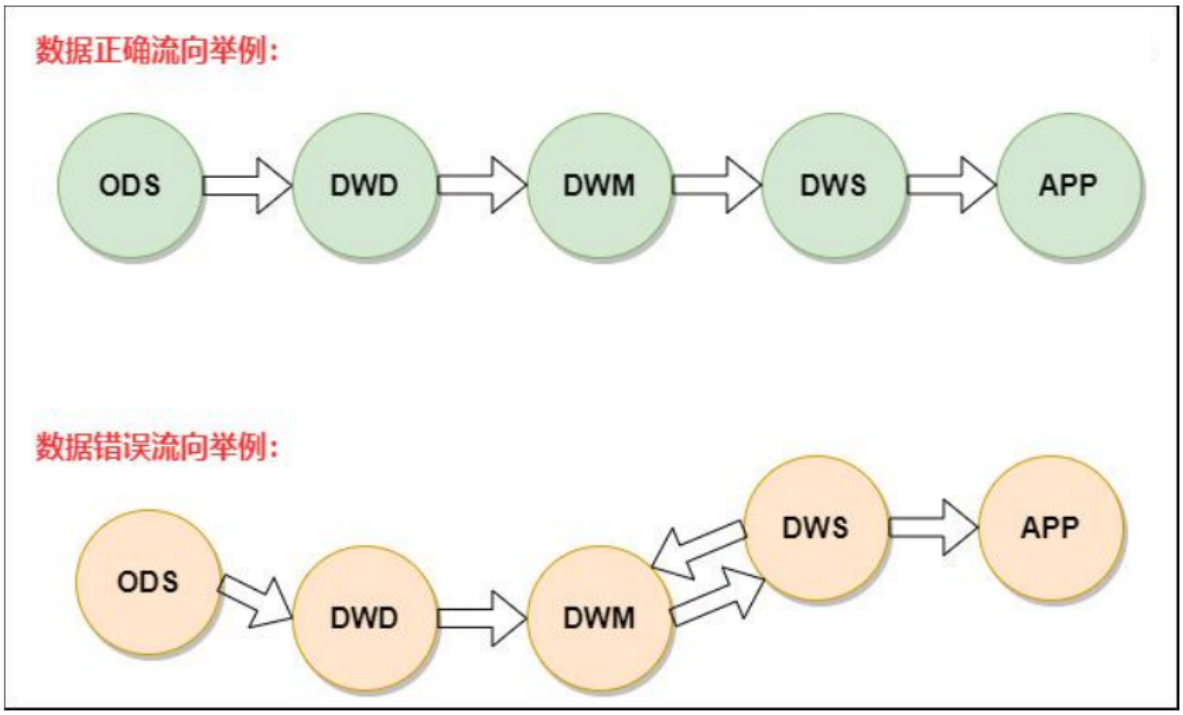
1. 层次调用规范

稳定业务 按照标准的数据流向进行开发，即 ODS –> DWD –> DWS –> APP。 非稳定业务 或探索性需求，可以遵循 ODS -> DWD -> APP 或者 ODS -> DWD -> DWM->APP 两个模型数据流。

在保障了数据链路的合理性之后，也必须保证模型分层引用原则：

- 正常流向：ODS -> DWD -> DWM -> DWS -> APP，当出现 ODS -> DWD -> DWS-> APP 这种关系时，说明主题域未覆盖全。应将 DWD 数据落到 DWM 中，对于使用频度非常低的表允许 DWD -> DWS。
- 尽量避免出现 DWS 宽表中使用 DWD 又使用（该 DWD 所归属主题域）DWM的表。
- 同一主题域内对于 DWM 生成 DWM 的表，原则上要尽量避免，否则会影响ETL 的效率。
- DWM、DWS 和 APP 中禁止直接使用 ODS 的表， ODS 的表只能被 DWD 引用。
- 禁止出现反向依赖 ， 例如 DWM 的表依赖 DWS 的表。

举例：



2. 数据类型规范

需统一规定不同的数据的数据类型，严格按照规定的数据类型执行：

1. 金额： double 或使用 decimal(11,2) 控制精度等， 明确单位是分还是元 。
2. 字符串： string。
3. d id 类： bigint。
4. 时间： string。
5. 状态： string

3. 数据冗余规范

宽表的冗余字段要确保：

1. 冗余字段要使用高频， 下游 3 3 个或以上使用 。

2. 冗余字段引入 不应造成本身数据产生过多的延后 。
3. 冗余字段 和已有字段的重复率不应过大， 原则上不应超过 60%，如需要可以选择 join 或原表拓展。

4. NULL 字段处理规范

- 对于维度字段，需设置为-1
- 对于指标字段，需设置为 0

5. 指标口径规范

保证主题域内，指标口径一致，无歧义 。
通过数据分层，提供统一的数据出口，统一对外输出的数据口径，避免同一指标不同口径的情况发生。

1) 指标梳理

指标口径的不一致使得数据使用的成本极高，经常出现口径打架、反复核对数据的问题。在数据治理中，我们将需求梳理到的所有指标进行进一步梳理，明确其口径，如果存在两个指标名称相同，但口径不一致，先判断是否是进行合并，如需要同时存在，那么在命名上必须能够区分开。

2) 指标管理

指标管理分为原子指标维护和派生指标维护。
原子指标：

- 选择原子指标的归属产线、业务板块、数据域、业务过程
- 选择原子指标的统计数据来源于该业务过程下的原始数据源
- 录入原子指标的英文名称、中文名称、概述
- 填写指标函数
- 系统根据指标函数自动生成原子指标的定义表达式
- 系统根据指标定义表达式以及数据源表生成原子指标 SQL

派生指标：

- 在原子指标的基础之上选择了一些维度或者修饰限定词。

6. 数据表处理规范

1) 增量表

新增数据，增量数据是上次导出之后的新数据。

1. 记录每次增加的量，而不是总量；
2. 增量表，只报变化量，无变化不用报；
3. 每天一个分区。

2) 全量表

每天的所有的最新状态的数据。

1. 全量表，有无变化，都要报；
2. 每次上报的数据都是所有的数据（变化的 + 没有变化的）；
3. 只有一个分区。

3) 快照表

按日分区，记录截止数据日期的全量数据。

1. 快照表，有无变化，都要报；
2. 每次上报的数据都是所有的数据（变化的 + 没有变化的）；
3. 一天一个分区

4) 拉链表

记录截止数据日期的全量数据。

1. 记录一个事物从开始，一直到当前状态的所有变化的信息；
2. 拉链表每次上报的都是历史记录的最终状态，是记录在当前时刻的历史总量；
3. 当前记录存的是当前时间之前的所有历史记录的最后变化量（总量）；
4. 只有一个分区。

7. 表的生命周期管理

这部分主要是要通过对历史数据的等级划分与对表类型的划分生成相应的生命
周期管理矩阵。

1) 历史数据等级划分

主要将历史数据划分 P0、Pl、P2、P3 四个等级，其具体定义如下：

- P0： 非常重要的主题域数据和非常重要的应用数据，具有不可恢复性，如交易、日志、集团 KPI 数据、 IPO 关联表。
- Pl： 重要的业务数据和重要的应用数据，具有不可恢复性，如重要的业务产品数据。
- P2： 重要的业务数据和重要的应用数据，具有可恢复性，如交易线 ETL 产生的中间过程数据。
- P3： 不重要的业务数据和不重要的应用数据，具有可恢复性，如某些 SNS产品报表。

2) 表类型划分

1. 事件型流水表（增量表）
   事件型流水表（增量表）指数据无重复或者无主键数据，如日志。
2. 事件型镜像表（增量表）
   事件型镜像表（增量表）指业务过程性数据，有主键，但是对于同样主键的属性会发生缓慢变化，如交易、订单状态与时间会根据业务发生变更。
3. 维表
   维表包括维度与维度属性数据，如用户表、商品表。
4. e Merge
   Merge 全量表包括业务过程性数据或者维表数据。由于数据本身有新增的或者发生状态变更，对于同样主键的数据可能会保留多份，因此可以对这些数据根据主键进行 Merge 操作，主键对应的属性只会保留最新状态，历史状态保留在前一天分区 中。例如，用户表、交易表等都可以进行 Merge 操作。
5. ETL 临时表
   ETL 临时表是指 ETL 处理过程中产生的临时表数据，一般不建议保留，最多 7天。
6. TT 临时数据
   TT 拉取的数据和 DbSync 产生的临时数据最终会流转到 DS 层，ODS 层数据作为原始数据保留下来，从而使得 TT&DbSync 上游数据成为临时数据。这类数据不建议保留很长时间，生命周期默认设置为 93 天，可以根据实际情况适当减少保留天数。

7. 普通全量表

很多小业务数据或者产品数据，BI 一般是直接全量拉取，这种方式效率快，对存储压力也不是很大，而且表保留很长时间，可以根据历史数据等级确定保留策略。通过上述历史数据等级划分与表类型划分，生成相应的生命周期管理矩阵，如下表所示：

		P0	P1	P2	P3
ODS层	事件型流水表（增量表）	永久保留	3年	365天	180天
	事件型流水表（增量表）	永久保留	3年	365天	180天
	维表（全量表）	33天+极限存储	33天+极限存储	33天+极限存储	33天+极限存储
	Merge 全量表	2天	2天	2天	2天
	普通全量表	3年	3年	3年	3年
	新同步全量表	3天	3天	3天	3天
DWD层	事件型流水表（增量表）	永久保留	3年	365天	180天
	事件型流水表（增量表）	永久保留	3年	365天	180天
	维表（全量表）	33天+极限存储	33天+极限存储	33天+极限存储	33天+极限存储
	普通全量表	3年	365天	365天	180天
DWS层	各粒度数据	永久保留	3年	3年	3年
临时存储区	ETL临时表	7天	3天	3天	3天
	TT临时表	7天	7天	7天	7天
应用层	运营报表	永久保留	——	——	——
	对外数据	7年	——	——	——
	内部产品	3年	——	——	——