数仓各层开发规范

1、ODS 层设计规范

1）同步规范：

1. 一个系统源表只允许同步一次；
2. 全量初始化同步和增量同步处理逻辑要清晰；
3. 以统计日期和时间进行分区存储；
4. 目标表字段在源表不存在时要自动填充处理。

2）表分类与生命周期：

1. ods 流水全量表：
   - 不可再生的永久保存；
   - 日志可按留存要求；
   - 按需设置保留特殊日期数据；
   - 按需设置保留特殊月份数据；
2. ods 镜像型全量表：
   - 推荐按天存储；
   - 对历史变化进行保留；
   - 最新数据存储在最大分区；
   - 历史数据按需保留；
3. ods 增量数据：
   - 推荐按天存储；
   - 有对应全量表的，建议只保留 14 天数据；
   - 无对应全量表的，永久保留；
4. ods 的 etl 过程中的临时表：
   - 推荐按需保留；
   - 最多保留 7 天；
   - 建议用完即删，下次使用再生成；
5. BDSync 非去重数据：
   - 通过中间层保留，默认用完即删，不建议保留。

3）数据质量

1. 全量表必须配置唯一性字段标识；
2. 对分区空数据进行监控；
3. 对枚举类型字段，进行枚举值变化和分布监控；
4. ods 表数据量级和记录数做环比监控；
5. ods 全表都必须要有注释；

2. 公共维度层设计规范

1) 设计准则

1. 一致性
   共维度在不同的物理表中的字段名称 、 数据类型 、 数据内容必须保持一致 （历史原因不一致，要做好版本控制）
2. 维度的组合与拆分

- 组合原则：
  将维度与关联性强的字段进行组合，一起查询，一起展示，两个维度必须具有天然的关系，如：商品的基本属性和所属品牌。
  无相关性：如一些使用频率较小的杂项维度，可以构建一个集合杂项维度的特殊属性。
  行为维度：经过计算的度量，但下游当维度处理，例：点击量 0-1000,100-1000等，可以做聚合分类。
- 拆分与冗余：
  针对重要性，业务相关性、源、使用频率等可分为核心表、扩展表。
  数据记录较大的维度，可以适当冗余一些子集。

2) 存储及生命周期管理

建议按天分区。

1. 3 个月内最大访问跨度<=4 天时，建议保留最近 7 天分区；
2. 3 个月内最大访问跨度<=12 天时，建议保留最近 15 天分区；
3. 3 个月内最大访问跨度<=30 天时，建议保留最近 33 天分区；
4. 3 个月内最大访问跨度<=90 天时，建议保留最近 120 天分区；
5. 3 个月内最大访问跨度<=180 天时，建议保留最近 240 天分区；
6. 3 个月内最大访问跨度<=300 天时，建议保留最近 400 天分区；

3. DWD 明细层设计规范

1) 存储及生命周期管理

建议按天分区。

1. 3 个月内最大访问跨度<=4 天时，建议保留最近 7 天分区；
2. 3 个月内最大访问跨度<=12 天时，建议保留最近 15 天分区；
3. 3 个月内最大访问跨度<=30 天时，建议保留最近 33 天分区；
4. 3 个月内最大访问跨度<=90 天时，建议保留最近 120 天分区；
5. 3 个月内最大访问跨度<=180 天时，建议保留最近 240 天分区；
6. 3 个月内最大访问跨度<=300 天时，建议保留最近 400 天分区；

2) 事务型事实表设计准则

 基于数据应用需求的分析设计事务型事实表，结合下游较大的针对某个业务过程和分析指标需求，可考虑基于某个事件过程构建事务型实时表；
 一般选用事件的发生日期或时间作为分区字段，便于扫描和裁剪；
 冗余子集原则，有利于降低后续 IO 开销；
 明细层事实表维度退化，减少后续使用 join 成本

3) 周期快照事实表

- 周期快照事实表中的每行汇总了发生在某一标准周期，如某一天、某周、某月的多个度量事件。
- 粒度是周期性的，不是个体的事务。
- 通常包含许多事实，因为任何与事实表粒度一致的度量事件都是被允许的。

4) 累积快照事实表

- 多个业务过程联合分析而构建的事实表，如采购单的流转环节。
- 用于分析事件时间和时间之间的间隔周期。
- 少量的且当前事务型不支持的，如关闭、发货等相关的统计。

4. DWS 公共汇总层设计规范

数据仓库的性能是数据仓库建设是否成功的重要标准之一。 聚集 主要是通过 汇总明细粒度数据 来获得改进查询性能的效果。通过访问聚集数据，可以减少数据库在响应查询时必须执行的工作量，能够快速响应用户的查询，同时有利于减少不同用访问明细数据带来的结果不一致问题。

1) 聚集的基本原则

- 一致性 。聚集表必须提供与查询明细粒度数据一致的查询结果。
- 避免单一表设计 。不要在同一个表中存储不同层次的聚集数据。
- 聚集粒度可不同 。聚集并不需要保持与原始明细粒度数据一样的粒度，聚集只关心所需要查询的维度。

2) 聚集的基本步骤

第一步：确定聚集维度

在原始明细模型中会存在多个描述事实的维度，如日期、商品类别、卖家等，这时候需要确定根据什么维度聚集，如果只关心商品的交易额情况，那么就可以根据商品维度聚集数据。

第二步：确定一致性上钻

这时候要关心是按月汇总还是按天汇总，是按照商品汇总还是按照类目汇总，如果按照类目汇总，还需要关心是按照大类汇总还是小类汇总。当然，我们要做的只是了解用户需要什么，然后按照他们想要的进行聚集。

第三步：确定聚集事实

在原始明细模型中可能会有多个事实的度量，比如在交易中有交易额、交易数量等，这时候要明确是按照交易额汇总还是按照成交数量汇总。

3) 公共汇总层设计原则

除了聚集基本的原则外，公共汇总层还必须遵循以下原则：

- 数据公用性 。汇总的聚集会有第三者使用吗？基于某个维度的聚集是不是经常用于数据分析中？如果答案是肯定的，那么就有必要把明细数据经过汇总沉淀到聚集表中。
- 不跨数据域 。数据域是在较高层次上对数据进行分类聚集的抽象。如以业务
- 区分统计周期 。在表的命名上要能说明数据的统计周期，如 _1d 表示最近 1 天， _td 表示截至当天， _nd 表示最近 N 天。