CDM公共维度层设计规范

更新时间：2020-03-07 16:03:52

[**本页目录**](javascript:void(0))

* [设计准则](https://help.aliyun.com/document_detail/117435.html?spm=a2c4g.11186623.6.917.6fad2bcaEo9PDc#title-s1v-evx-5jr)
* [表命名规范](https://help.aliyun.com/document_detail/117435.html?spm=a2c4g.11186623.6.917.6fad2bcaEo9PDc#title-nw6-1hu-jc3)
* [数据存储及生命周期管理规范](https://help.aliyun.com/document_detail/117435.html?spm=a2c4g.11186623.6.917.6fad2bcaEo9PDc#title-ysy-qhi-nf3)

本文为您介绍CDM公共维度层设计规范。

**设计准则**

* 一致性维度规范

公共层的维度表中相同维度属性在不同物理表中的字段名称、数据类型、数据内容必须保持一致。除了以下情况：

* + 在不同的实际物理表中，如果由于维度角色的差异，需要使用其他的名称，其他名称也必须是规范的维度属性的别名。例如，定义一个标准的会员ID时，如果在一个表中，分别要表示买家ID，卖家ID，那么设计规范阶段就预先对会员ID分别定义买家ID和卖家ID。
  + 如果由于历史原因，在暂时不一致的情况下，必须在规范的维度定义一个标准维度属性，不同的物理名也必须是来自标准维度属性的别名。
* 维度的组合与拆分
  + 组合原则
    - 将维度所描述业务相关性强的字段在一个物理维表实现。相关性强是指经常需要一起查询或进行报表展现、两个维度属性间是否存在天然的关系等。例如，商品基本属性和所属品牌。
    - 无相关性的维度可以适当考虑杂项维度（例如交易），可以构建一个交易杂项维度收集交易的特殊标记属性、业务分类等信息。也可以将杂项维度退化在事实表中处理，不过容易造成事实表相对庞大，加工处理较为复杂。
    - 所谓的行为维度是经过汇总计算的指标，在下游的应用使用时将其当维度处理。如果有需要，度量指标可以作为行为维度冗余到维度表中。
  + 拆分与冗余
    - 对于维度属性过多，涉及源较多的维度表（例如会员表），可以做适当拆分：
      * 拆分为核心表和扩展表。核心表相对字段较少，刷新产出时间较早，优先使用。扩展表字段较多，且可以冗余核心表部分字段，刷新产出时间较晚，适合数据分析人员使用。
      * 根据维度属性的业务不相关性，将相关度不大的维度属性拆分为多个物理表存储。
    - 数据记录数较大的维度表（例如商品表），可以适当冗余一些子集合，以减少下游扫描数据量：
      * 可以根据当天是否有行为，产出一个有活跃行为的相关维表，以减少应用的数据扫描量。
      * 可根据所属业务扫描数据范围大小的不同，进行适当子集合冗余。

**表命名规范**

命名规则：**{project\_name}.dim{业务/pub}{维度定义}[\_{自定义命名标签}]**，其中的pub与具体业务无关，各个业务部都可以共用，例如时间维度。

**数据存储及生命周期管理规范**

CDM公共维度层的表的类型为维度表，存储方式为按天分区。

模型设计者根据自身业务需求设置表的生命周期管理。您可依据3个月内的最大需要访问的跨度设置保留策略，具体计算方式如下：

* 当3个月内的最大访问跨度小于或等于4天时，建议将保留天数设为7天。
* 当3个月内的最大访问跨度小于或等于12天时，建议将保留天数设为15天。
* 当3个月内的最大访问跨度小于或等于30天时， 建议将保留天数设为33天。
* 当3个月内的最大访问跨度小于或等于90天时，建议将保留天数设为93天。
* 当3个月内的最大访问跨度小于或等于180天时， 建议将保留天数设为183天。
* 当3个月内的最大访问跨度小于或等于365天时，建议将保留天数设为368天。

CDM明细层设计规范

更新时间：2020-04-07 15:42:03

[**本页目录**](javascript:void(0))

* [表命名规范](https://help.aliyun.com/document_detail/117437.html?spm=a2c4g.11186623.6.918.435a7ad09KDT5e#title-7ie-2pt-bbt)
* [数据存储及生命周期管理规范](https://help.aliyun.com/document_detail/117437.html?spm=a2c4g.11186623.6.918.435a7ad09KDT5e#title-myr-kqe-34q)
* [事务型事实表设计准则](https://help.aliyun.com/document_detail/117437.html?spm=a2c4g.11186623.6.918.435a7ad09KDT5e#title-o53-w51-zeh)
* [周期快照型事实表](https://help.aliyun.com/document_detail/117437.html?spm=a2c4g.11186623.6.918.435a7ad09KDT5e#title-1uq-w3h-1mj)
* [累计快照事实表](https://help.aliyun.com/document_detail/117437.html?spm=a2c4g.11186623.6.918.435a7ad09KDT5e#title-abq-o71-jub)

本文为您介绍CDM明细层的表、数据存储与生命周期管理和各种事实表的设计规范。

**表命名规范**

命名规则：**{project\_name}.dwd{业务缩写/pub}{数据域缩写}{业务过程缩写}[{自定义表命名标签缩写}]{刷新周期标识}{单分区增量全量标识}**。

命名说明：

* pub表示数据包括多个业务的数据。
* 单分区增量全量标识：i表示增量，f表示全量。

**数据存储及生命周期管理规范**

CDM明细层的表的类型为事实表，存储方式为按天分区。

事务型事实表一般永久保存。周期快照型事实表根据业务需求设置生命周期管理。您可依据3个月内的最大需要访问的跨度设置保留策略，具体计算方式如下：

* 当3个月内的最大访问跨度小于或等于4天时，建议将保留天数设为7天。
* 当3个月内的最大访问跨度小于或等于12天时，建议将保留天数设为15天。
* 当3个月内的最大访问跨度小于或等于30天时， 建议将保留天数设为33天。
* 当3个月内的最大访问跨度小于或等于90天时，建议将保留天数设为93天。
* 当3个月内的最大访问跨度小于或等于180天时， 建议将保留天数设为183天。
* 当3个月内的最大访问跨度小于或等于365天时，建议将保留天数设为368天。

**事务型事实表设计准则**

事务型事实表主要用于分析行为与追踪事件。事务事实表获取业务过程中的事件或者行为细节，然后通过事实与维度之间关联，可以非常方便地统计各种事件相关的度量，例如浏览UV，搜索次数等等。

* 基于数据应用需求的分析设计事务型事实表，如果下游存在较大的针对某个业务过程事件的分析指标需求，可以考虑基于某一个事件过程构建事务型事实表。
* 事务型事实表一般选用事件发生日期或时间作为分区字段，这种分区方式可以方便下游的作业数据扫描执行分区裁剪。
* 明细层事实表的冗余子集的原则能有利于降低上层数据访问的IO开销。
* 明细层事实表维度退化到事实表原则能有利于减少上层数据访问的JOIN成本。

**周期快照型事实表**

周期快照型事实表主要用于分析状态型或者存量型事实。快照是指以预定的时间间隔来采样状态度量。

**累计快照事实表**

累计快照事实表是基于多个业务过程联合分析从而构建的事实表，如采购单的流转环节等。

累计快照事实表主要用于分析事件之间的时间间隔与周期。例如，用交易的支付与发货之间的间隔，来分析发货速度，或在支付和退款环节分析支付退款率等等。

累计快照事实表同时也可以用于帮助分析一些少量的、且对刷新时间不是非常敏感的指标统计。例如，在当前事务型事实表不支持，且只有少量的统计指标时，需要分析交易的关闭和发货，就可以基于累计快照事实表进行计算。

CDM汇总层设计规范

更新时间：2020-06-18 14:10:42

[**本页目录**](javascript:void(0))

* [命名规范](https://help.aliyun.com/document_detail/117438.html?spm=a2c4g.11186623.6.919.495346abcTz7wL#title-m2l-o8e-f9y)
* [数据存储及生命周期管理规范](https://help.aliyun.com/document_detail/117438.html?spm=a2c4g.11186623.6.919.495346abcTz7wL#title-xl6-4wc-vsq)

本文为您介绍CDM汇总层设计规范。

**命名规范**

命名规则：**{project\_name}.dws{业务缩写/pub}{数据域缩写}{数据粒度缩写}[{自定义表命名标签缩写}]{统计时间周期范围缩写}{刷新周期标识}{单分区增量全量标识}**。

命名说明：

* 在默认情况下，离线计算应该包括最近一天（1d）、最近N天（nd）和历史截至当天（td）三个表。

如果nd表的字段过多，需要拆分时，只允许以一个统计周期单元作为原子拆分，即一个统计周期拆分一个表。例如，最近7天（1w）拆分一个表，不允许拆分出来的一个表存储多个统计周期。

* 对于{刷新周期标识}和{单分区增量全量标识}在汇总层不做强制要求。单分区增量全量标识：i表示增量，f表示全量。
* 对于小时表不管是按天刷新还是按小时刷新，都用\_hh来表示。
* 对于分钟表不管是按天刷新还是按小时刷新，都用\_mm来表示。

**数据存储及生命周期管理规范**

CDM汇总层的表的类型为事实表，存储方式为按天分区。

事务型事实表一般会永久保存。周期快照型事实表根据业务需求设置生命周期管理。您可依据3个月内的最大需要访问的跨度设置保留策略，具体计算方式如下：

* 当3个月内的最大访问跨度小于或等于4天时，建议将保留天数设为7天。
* 当3个月内的最大访问跨度小于或等于12天时，建议将保留天数设为15天。
* 当3个月内的最大访问跨度小于或等于30天时，建议将保留天数设为33天。
* 当3个月内的最大访问跨度小于或等于90天时，建议将保留天数设为93天。
* 当3个月内的最大访问跨度小于或等于180天时，建议将保留天数设为183天。
* 当3个月内的最大访问跨度小于或等于365天时，建议将保留天数设为368天。

CDM接口数据层设计规范

更新时间：2019-12-16 17:27:24

本文为您介绍CDM接口数据层设计规范。

接口数据层将不同数据域的汇总数据预关联在一个物理表，开放给应用使用，以减少应用层多次重复JOIN的成本开销，CDM接口数据层更适用于实时计算。

命名规则：**{project\_name}.dwi{业务 BU 缩写/pub}{数据域/hbd}{数据粒度缩写}[{自定义表命名标签缩写}]\_{统计时间周期范围缩写}**。统计时间周期范围和数据域说明如下：

* 关于统计时间周期范围缩写，默认情况下，离线计算包括最近一天（1d）、 最近N天（nd）和历史截至当天（td）三个表。

**说明** 如果出现nd 的表字段过多，需要拆分时，只允许以一个统计周期单元作为原子拆分，即一个统计周期拆分一个表，例如最近7天（1week）拆分一个表，不允许拆分出一个表存储多个统计周期。

* 如果一个汇总表出现混合多个数据域时，表名称中需要使用hbd（hybird 缩写）进行标识，这种情况当前只用于准实时情况，离线计算不建议跨数据域存储数据。