## 数据仓库层级

### 为什么要分层

我们对数据进行分层的一个主要原因就是方便管理，能对数据有一个更加清晰的掌控，详细来讲，主要有下面几个原因：

* 清晰数据结构：每一个数据分层都有它的作用，这样我们在使用表的时候能更方便地定位和理解。
* 数据血缘追踪：在层级的规范下，如果有一个数据出问题了，我们就能够快速准确地定位到问题，并清楚它的危害范围。
* 减少重复开发：规范数据分层，开发一些通用的中间层数据，能够减少极大的重复计算。提高计算效率。
* 把复杂问题简单化。讲一个复杂的任务分解成多个步骤来完成，每一层只处理单一的步骤，比较简单和容易理解。而且便于维护数据的准确性，当数据出现问题之后，可以不用修复所有的数据，只需要从有问题的步骤开始修复。
* 屏蔽原始数据的异常。处理一些数据错乱或者质量上的问题。

### 层级结构

数据仓库的层级分为，ODS层、DW层、DM层、APP层。层级不要太多，太多会造成数据一致性上的问题，同时也会增加ETL的难度。

#### ODS层

对源系统数据的copy，加上时间戳，通常只保留最近（1-7天）抽取的数据。基本不使用任何的数据转换，原因：

1. 提高ETL效率。

2. 便于与源系统进行数据核对。

3. 减轻源系统的压力。通常数仓抽取的数据量是巨大的，如果在抽取的过程采用复杂的转换，势必对源系统造成很大的IO压力。

#### DW层

数据仓库层，Inmon的3NF建模层。根据数据仓库主体域进行建模，包含历史数据，粒度通常与业务系统一致。通常满足3NF，也向外部提供一些明细查询的接口。在优秀数据仓库的项目中，对源系统历史数据的查询，可以迁移到DW层。

在ODS->DW的过程中要进行数据的标准化，清洗等操作。

3

常用术语

|  |  |
| --- | --- |
| ETL(Extract Transform Load) | 数据的清洗，转换，加载。常用的ETL工具有，Kettle，Datastage、Informatic  PowerCenter |
| *DM*（*Data Mart*）数据集市 | 数据集市也可叫做“小数据仓库”。那么数据集市就是企业级数据仓库的一个子集,主要面向部门级业务。数据集市可以在一定程度上解决访问数据仓库的瓶颈，通常按照Kimball的维度建模理论进行设计。查询效率会非常高。 |
| *Cube* | 数据立方体 |
| 主数据 | 企业级的权威数据，例如客户信息、卡、帐户等，通过打通各个业务系统的孤岛，整合而成的权威数据。 |
| 元数据*Metadata* | 数据的数据，描述数据的技术属性，业务属性，相关接口，版本变更情况等信息。 |
|  |  |