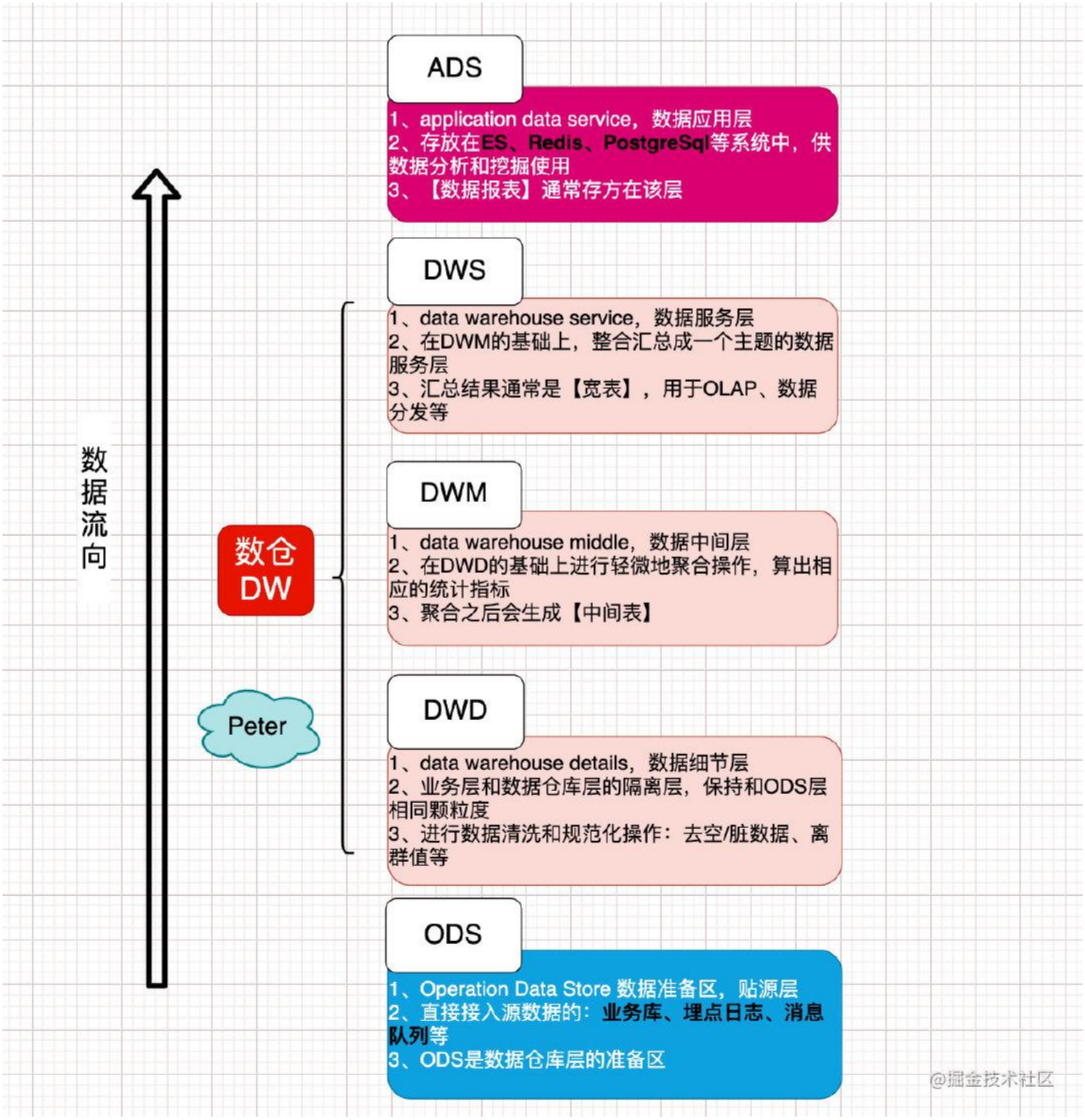
数据分层详解ODS、DWD、DWM、DWS、ADS

原创 铁憨憨b 于 2022-01-27 10:52:52 发布 18925

详解数仓中的数据分层：ODS、DWD、DWM、DWS、ADS



何为数仓DW

Data warehouse（可简写为DW或者DWH）数据仓库，是在数据库已经大量存在的情况下，它是**一整套包括了etl、调度、建模在内的完整的理论体系。**

数据仓库的方案建设的目的，是为前端查询和分析作为基础，主要应用于OLAP（on-line Analytical Processing），支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果。目前行业比较流行的有：AWS Redshift，Greenplum，Hive等。

数据仓库并不是数据的最终目的地，而是为数据最终的目的地做好准备，这些准备包含：清洗、转义、分类、重组、合并、拆分、统计等

为何要分层

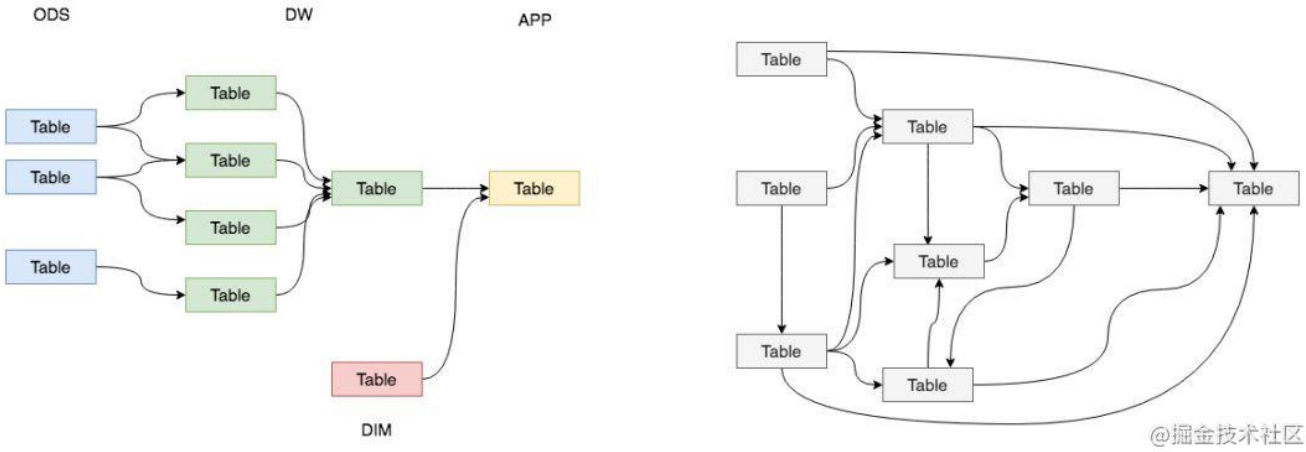
数据仓库中涉及到的问题：

1. 为什么要做数据仓库?

2. 为什么要做数据质量管理?
3. 为什么要做元数据管理?
4. 数仓分层中每个层的作用是什么?
5. ……

在实际的工作中，我们都希望自己的数据能够有顺序地流转，设计者和使用者能够清晰地知道数据的整个声明周期，比如下面左图。

但是，实际情况下，我们所面临的数据状况很有可能是复杂性高、且层级混乱的，我们可能会做出一套表依赖结构混乱，且出现循环依赖的数据体系，比如下面的右图。



为了解决我们可能面临的问题，需要一套行之有效的数据组织、管理和处理方法，来让我们的数据体系更加有序，这就是**数据分层。**数据分层的好处：

- 清晰数据结构：让每个数据层都有自己的作用和职责，在使用和维护的时候能够更方便和理解
- 复杂问题简化：将一个复杂的任务拆解成多个步骤来分步骤完成，每个层只解决特定的问题
- 统一数据口径：通过数据分层，提供统一的数据出口，统一输出口径
- 减少重复开发：规范数据分层，开发通用的中间层，可以极大地减少重复计算的工作

数据分层

每个公司的业务都可以根据自己的业务需求分层不同的层次；目前比较流行的数据分层：数据运营层、数据仓库层、数据服务层。

数据运营层ODS

数据运营层：Operation Data Store 数据准备区，也称为贴源层。数据源中的数据，经过抽取、洗净、传输，也就是ETL过程之后进入本层。该层的主要功能：

- ODS是后面数据仓库层的准备区
- 为DWD层提供原始数据
- 减少对业务系统的影响

为了考虑后续可能需要追溯数据问题，因此对于这一层就不建议做过多的数据清洗工作，原封不动地接入原始数据即可

这层的数据是后续数据仓库加工数据的来源。数据来源的方式：

1. 业务库：sqoop定时抽取数据；实时方面考虑使用canal监听mysql的binlog日志，实时接入即可
2. 埋点日志：日志一般是以文件的形式保存，可以选择使用flume来定时同步；可以使用spark streaming或者Flink、Kafka来实时接入
3. 消息队列：来自ActiveMQ、Kafka的数据等

数据仓库层

数据仓库层从上到下，又可以分为3个层：数据细节层DWD、数据中间层DWM、数据服务层DWS。

数据细节层DWD

数据细节层：data warehouse details，DWD

该层是业务层和数据仓库的隔离层，保持和ODS层一样的数据颗粒度；主要是对ODS数据层做一些数据的清洗和规范化的操作，比如去除空数据、脏数据、离群值等。

为了提高数据明细层的易用性，该层通常会才采用一些维度退化方法，将维度退化至事实表中，减少事实表和维表的关联。

数据中间层DWM

数据中间层：Data Warehouse Middle，DWM；

该层是在DWD层的数据基础上，对数据做一些轻微的聚合操作，生成一些列的中间结果表，提升公共指标的复用性，减少重复加工的工作。

> 简答来说，对通用的核心维度进行聚合操作，算出相应的统计指标

数据服务层DWS

数据服务层：Data Warehouse Service，DWS；

该层是基于DWM上的基础数据，整合汇总成分析某一个主题域的数据服务层，一般是宽表，用于提供后续的业务查询，OLAP分析，数据分发等。

一般来说，该层的数据表会相对较少；一张表会涵盖比较多的业务内容，由于其字段较多，因此一般也会称该层的表为宽表。

数据应用层ADS

数据应用层：Application Data Service，ADS；

该层主要是提供给数据产品和数据分析使用的数据，一般会存放在ES、Redis、PostgreSql等系统中供线上系统使用；也可能存放在hive或者Druid中，供数据分析和数据挖掘使用，比如常用的数据报表就是存在这里的。

事实表 Fact Table

事实表是指存储有事实记录的表，比如系统日志、销售记录等。事实表的记录在不断地增长，比如电商的商品订单表，就是类似的情况，所以事实表的体积通常是远大于其他表。

维表层Dimension

维度表（Dimension Table）或维表，有时也称查找表（Lookup Table），是与事实表相对应的一种表；它保存了维度的属性值，可以跟事实表做关联，相当于将事实表上经常重复出现的属性抽取、规范出来用一张表进行管理。维度表主要是包含两个部分：

- 高基数维度数据：一般是用户资料表、商品资料表类似的资料表，数据量可能是千万级或者上亿级别
- 低基数维度数据：一般是配置表，比如枚举字段对应的中文含义，或者日期维表等；数据量可能就是个位数或者几千几万。