2019

企业级数据仓库实战

分层的边界

ODS层

数据接入层，也叫ODS层，是最接近数据源中数据的一层，数据源中的数据，经过抽取、洗净、传输，也就说传说中的 ETL 之后，装入本层。本层的数据，总体上大多是按照源头业务系统的分类方式而分类的。

一般来讲，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，至于数据的去噪、去重、异常值处理等过程可以放在后面的DWD层来做。

DWD层

该层一般保持和ODS层一样的数据粒度，并且提供一定的数据质量保证。同时，为了提高数据明细层的易用性，该层会采用一些维度退化手法，将维度退化至事实表中，减少事实表和维表的关联。

DWS层

该层会在DWD层的数据基础上，对数据做轻度的聚合操作，生成一系列的中间表，提升公共指标的复用性，减少重复加工。

直观来讲，就是对通用的核心维度进行聚合操作，算出相应的统计指标。

DWM层

又称数据集市或宽表。按照业务划分，如流量、订单、用户等，生成字段比较多的宽表，用于提供后续的业务查询，OLAP分析，数据分发等。

一般来讲，该层的数据表会相对比较少，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。

APP层

在这里，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、PostgreSql、Redis等系统中供线上系统使用，也可能会存在 Hive 或者 Druid 中供数据分析和数据挖掘使用。比如我们经常说的报表数据，一般就放在这里。

DIM层

在表层主要包含两部分数据：

1. 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。
2. 低基数维度数据：一般是配置表，比如枚举值对应的中文含义，或者日期维表。数据量可能是个位数或者几千几万。

分层的边界

APP

- 数据应用层(Application Model)
- 个性化指标加工：不公用性；复杂性（指数型、比值型、排名型指标）
- 基于应用的数据组装，与应用强耦合

DW

- 数据仓库层(Data Warehouse)
- Dim（dimension）：公共维表层，一致性维度建设
- Dwm(data warehouse market)：数据集市层，宽表集市，跨多业务场景、行为数据组装
- Dws(data warehouse summary)：数据汇总层，单业务场景，行为数据组装，提升公共指标的复用性，减少重复的加工
- Dwd(data warehouse detail)：数据明细层，存储经过标准化后的数据

ODS

- 数据接入层(Operational Data Store)
- 同步：结构化，非结构化数据增量或全量同步到HDFS;
- 字段名与业务平台表保持一致,根据数据业务需求保存历史数据;

THANK YOU FOR YOUR GUIDANCE.

谢谢