The Tips of Data Warehouse

> 版本号：0.1.0
>
> 发布地址：https://github.com/dantezhao/woodlab
>
> 作者：木东居士， dantezhao91@outlook.com

概述

大数据时代，作为数据的掌握者，我们不仅要更好地使用数据，也要更好地管理数据。而数据仓库正是这样一套管理和组织数据的解决方案。

本文试图从一种经验的角度来描述在数据仓库建设中的会遇到的各种坑和需要注意的关键点，希望以此帮助踏上数据仓库之路的小伙伴们。

注意：本文不会详细地解释数据仓库的各个概念，亦不会给出各种示例代码来阐述数据仓库的建设细节。

正文

0x01 请理解数据仓库和数据平台的区别

当你开始建设数据仓库之前，需要明白数据仓库和数据平台是两个不同的概念，不要把搭建一套 Hadoop + Hive 的平台叫数据仓库，这是数据平台的范畴。

我们常说的数据仓库不仅仅是指数据接入、数据存储和数据计算，它也要包括数据治理、数据建模和数据挖掘。比如元数据管理、维度建模和 OLAP 分析，这些都是我们在建设数据仓库时候要考虑的内容。

0x02 提前规划你的数据仓库

数据仓库是公司数据体系的核心模块，数据仓库可以做的不好，但是不能不做。

因此，在数据体系设计的前期最好要有一定的规划，即使最简单的表和字段命名的规范也能带来很大的收益。

另外，从数据开发的角度出发，在做各种临时数据处理需求的时候也要有数据仓库的思维，多尝试抽象出来数据中间层，这样对公司和对自己的成长都是有帮助的。

0x03 实现轻量级的数据仓库

如果业务的快速发展不能留给你太多的时间来实现一个完善的数据仓库，那么可以考虑在前期实现一个轻量级的数据仓库，以尽可能小的成本带来最大收益。关于这个轻量级的数据仓库，建议优先考虑如下几个点：

1. 明确数据分层
2. 确定可执行的表和字段命名规范
3. 定期抽象出常用的中间表
4. 建设元数据管理系统，或者建设文档库，提供中间表的文档说明

0x04 不要脱离业务场景

做数据一定要记得贴近业务，虽说会有很多临时和重复需求，但却能切实地创造价值。

切记不要以为可以完全脱离业务去做一套数据仓库，我们可以在数据仓库的某个层次不以业务需求为导向来设计，但是最终面向业务的数据一定会是和业务理解有关。

0x05 文档！文档！

数据仓库建设的初期，要逐步沉淀出各种文档，比如模型设计文档、字段命名规范文档、SQL 开发规范文档。文档是数据仓库沉淀的最直观的一种体现，这也是技术积累的一部分。

最重要的是，如果元数据系统没有成型，那就要把数据仓库中间表的内容沉淀到文档中，尽量做到一表一文档。这样不管是从节约沟通成本的角度，亦或是增加团队积累，更或是完成 KPI 的角度考虑，都是有很大益处的。

0x06 尽早布局数据质量管理

请尽早布局数据质量管理的内容，不要等到发生严重的数据事故后才注意到数据质量问题。关于数据质量监控，如果没有足够的时间和精力做一套完整的系统，可以先从以下几个点入手，这样至少能对自己有一层基本的保护：

1. 核心数据每日数据量级监控和告警
2. 重要业务指标监控和告警
3. 主要业务流程各阶段数据的监控和告警

0x07 多使用视图表

多使用视图表对外提供数据服务，它可以有效地屏蔽业务方对最底层表结构变更的感知，同时加强权限管理。

如下场景可以多考虑使用视图表：

1. 该表经常会有加字段的需求
2. 该表的计算口径会出现变化，需要并行跑多份数据，某个时间点进行表切换
3. 该表可能会对不同人或部门提供服务，希望不同人或部门可读的字段不同

视图表主要是来晚上表结构变更、口径修改和权限管理的场景，不要滥用而增加维护成本。

0x08 考虑你的职业发展

不要一直埋着头搞 ETL，可以搞半年或一年来了解大致的业务和技能，但不能长期这样发展。现在开源平台相对成熟，长时间搞 ETL，会弱化自己的技术深度，如果再没有数据挖掘相关的项目经验，很容易在以后得面试中被淘汰。

因此，建议各位数据开发的小伙伴，如果你近一年的工作主要都是在用 SQL 做 ETL，那就要有一点危机意识，经常反思一下自己是否有成长，核心竞争力是否有所提现。

如果有些心虚，可以考虑在数据仓库、数据挖掘或者核心平台开发上下一些功夫。