浅析数仓数据管理

caolei

Exported on 01/10/2020

Table of Contents

1 数据管理定义

数据管理，顾名思义就是管理数据。下面是来自wikipedia[1]的定义。

> Data Management comprises all disciplines related to managing data as a valuable resource.
> 数据管理包括与将数据作为有价值的资源进行管理相关的所有学科。

1 https://en.wikipedia.org/wiki/Data_management

2 数据管理范畴

数据管理涉及非常多的学科，所以管理范畴也会比较宽泛，常见的范畴主题包括数据治理、数据架构、数据建模、数据存储、数据安全、数据集成与互操作、数据仓库与商业智能、元数据、数据质量等。

这篇文章主要讨论大数据数据仓库领域内的数据管理，主要有元数据管理、数据存储管理、数据质量管理、数据建模管理四个主题。

3 元数据管理

元数据是关于数据、操纵数据的进程和应用程序的结构和意义的描述信息，其主要目标是提供数据资源的全面指南。元数据不仅定义了数据仓库中数据的模式、来源以及抽取和转换规则等，而且整个数据仓库系统的运行都是基于元数据的，是元数据把数据仓库系统中的各个松散的组件联系起来，组成了一个有机的整体。

元数据打通数据源、数据仓库、数据应用，记录了数据从产生到消费的完整链路。它包含静态的表、列、分区信息（也就是MetaStore）；动态的任务、表依赖映射关系；数据仓库的模型定义、数据生命周期；以及ETL任务调度信息、输入输出等。

元数据是数据管理、数据内容、数据应用的基础。例如可以利用元数据构建任务、表、列、用户之间的数据图谱；构建任务DAG依赖关系，编排任务执行序列；构建任务画像，进行任务质量治理；数据分析时，使用数据图谱进行字典检索；根据表名查看表详情，以及每张表的来源、去向，每个字段的加工逻辑；提供个人或BU的资产管理、计算资源消耗概览等。

3.1 基础信息管理

3.2 血缘关系管理

4 元数据管理组成

数据仓库元数据的管理主要分为以下几个方面

基本信息管理：包括库、表、字段、注释等信息的管理

数据质量管理：数据质量可被理解成具有一系列特征的数据满足要求的程度。这些特征可以是：完整性，有效性，准确性，一致性，可用性和及时性。要求被定义为所陈述的需求或期望，通常是暗示的或强制性的。针对质量的管理需要从定义出发。

生命周期管理：数据不能只是一味的产出，还应该有消亡的过程，生命周期管理会让数据经历产生、使用、迁移、清理、销毁等过程

血缘关系管理：在数仓建设中，三种实体会存在血缘关系的概念：任务、表、字段；例如任务A需要等待任务B和任务C运行成功之后才能运行，或者需要等待任务B和任务C至少一个运行成功之后才能运行，我们就可以定义任务A和任务B、任务C存在血缘关系，即任务A的上游任务是任务B和任务C。表和字段的血缘关系也是一样的，后面会详细介绍。

数据热度管理：数据热度可以理解为哪些数据被经常是使用，哪些数据在一定时间内从来没有被使用过，这块的核心是如何定义冷热，满足什么样条件的数据算是冷数据，反之满足什么样条件的数据算是热数据。

5 基本信息管理

TODO

6 数据质量管理

TODO

7 生命周期管理

TODO

8 血缘关系管理

TODO

9 数据热度管理

TODO

10 总计与展望

TODO