**数据治理学习指南**

目录

[声明 3](#_Toc17500)

[一、 绪论 3](#_Toc15607)

[1.1. 数据治理背景 3](#_Toc21732)

[1.2. 数据治理范围 3](#_Toc18062)

[1.3. 数仓建设模型 4](#_Toc18736)

[1.3.1. 数仓分层 5](#_Toc28317)

[1.3.2. Mongo爬虫库 5](#_Toc6606)

[1.3.3. ODS 5](#_Toc7037)

[1.3.4. DW 6](#_Toc19353)

[1.3.4.1. DWD\_PRE 6](#_Toc7934)

[1.3.4.2. DWD 6](#_Toc12303)

[1.3.5. DIM 6](#_Toc30672)

[1.3.6. DSP 6](#_Toc6567)

[1.4. 术语 6](#_Toc19200)

[二、 原始数据（ODS） 6](#_Toc11207)

[1.1. 数据来源 6](#_Toc16047)

[三、 预处理数据（PRE） 7](#_Toc27441)

# 声明

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 更新人 | 更新内容 | 更新时间 |
| V0.1 | 丁子宸 | 创建 | 2021-9-3 |

# 绪论

## 数据治理背景

过去的十年是企业的IT系统经历了数据量高速膨胀的时期，这些海量的、分散在不同角落的数据导致了数据资源利用的复杂性和管理的高难度。同时，企业内部的业务区隔或行政分化也在不断地制造着企业数据交互的断层，而企业与外部业务交互所产生的“体外循环”数据与企业的核心数据体系并不能自然地融合。当这种数据的异构化所导致的应用冲突达到一定临界点时，数据治理便成为了规范企业数据的必要步骤。

数据治理专注于将数据作为企事业单位数据资产进行应用和管理的一套管理机制，能够消除数据的不一致性，建立规范的数据应用标准，提高数据质量，实现数据内外部共享，并能够将数据作为组织的宝贵资产应用于业务、管理、战略决策中，发挥数据资产价值。

## 数据治理范围

数据治理涉及的IT技术主题众多，包括元数据管理、主数据管理、数据质量、数据集成、监控与报告等。当然基于每家公司的数据治理目标不一样，每家公司的数据治理实施关键路径也会不一样。

## 数仓建设模型

提到数据仓库，必须提到恩门和金博尔两个人，他们共同开创的数仓建模的设计方法，对现代数仓的建设有着重大意义。

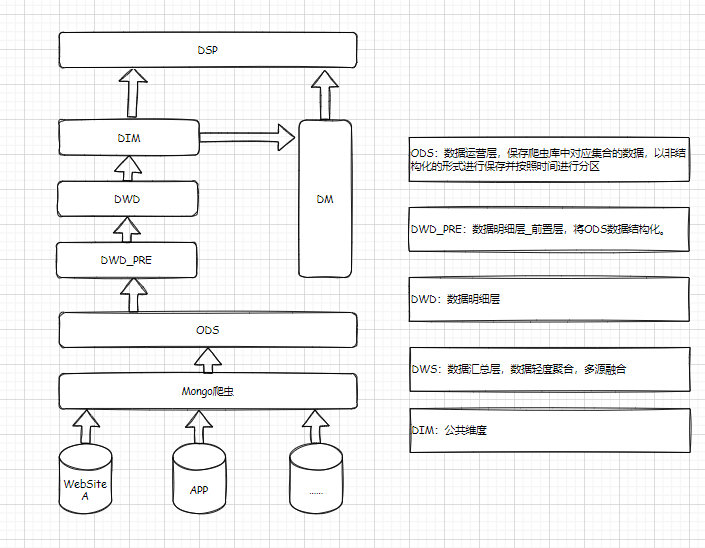
恩门提出的建模方法自顶向下（这里的顶是指数据的来源，在传统数据仓库中，就是哥哥业务数据库）。基于业务中各个实体以及实体之间的关系，构建数据仓库。

金博尔建模与恩门正好相反，是一种自底向上的模型建设，从数据分析的需求出发，拆分维度和事实。这两种方法各有优劣，恩门建模是从数据源开始构建，构建成本比较高，适用于应用场景比较固定的业务，比如金融领域，冗余数据少是它的优势。金博尔建模由于是从分析场景出发，适用于变化速度比较快的业务，比如互联网业务。

### 数仓分层

围绕恩门模型建设数据仓库，通过ETL流程对数据进行标准化、验证、清洗，并最终汇总到数据存储，提供上层业务查询、分析。数仓分层的目的是使用空间换时间，通过大量预处理，提升用户数据加工效率。如果不分层，源业务系统的业务规则发生变化将会影响整个数据清洗过程，工作量巨大。 通过数据分层管理可以简化数据清洗的过程，因为把原来一步的工作分到了多个步骤去完成，相当于把一个复杂的工作拆分成了多个简单的工作，把一个大的黑盒变成了一个白盒，每一层的处理逻辑都相对简单和容易理解，这样我们比较容易保证每一个步骤的正确性，当数据发生错误的时候，往往我们只需要局部调整某个步骤即可。

在冰鉴的数据体系中，数据主要来源于外采数据、爬虫数据，结合业务需求及数仓分层需要，将数仓分为ODS、DW、DIM三层，同时，延伸数仓的来源和使用，输入源边界统一为基于Mongo的爬虫库，输出源为业务同步的DSP。



### ODS（数据运营层）

数据运营层ODS，类似于数据湖概念，保存爬虫库中对应的Collection数据，以非结构化的形式进行保存并按照时间进行分区。通过同步工具，以T+1的形式将Mongo库中的数据同步到对应表中。

### DW

### DWD\_PRE

数据明细层—前置层DWD\_PRE，本层将ODS数据按照业务维度需要进行结构化。

### DWD

### DIM

### DSP

## 术语

ETL：

数据湖：

# 原始数据（ODS）

## 数据来源

# 预处理数据（PRE）