

---

## 《大数据下的数据管理》

-段方

某世界 100 强企业大数据总设计师

教授 北京大学博士后

---

---

### 1 概述

---

#### 1.1 数据要素化、资产化

##### 1.1.1 大数据成为新的生产“要素”

##### 1.1.2 大数据的基本框架

##### 1.1.3 数据资产化基础

##### 1.1.4 数据治理的重要性

#### 1.2 大数据下的数据管理

##### 1.2.1 行业数据管理的特点

##### 1.2.2 大数据为什么强调数据管理？

##### 1.2.3 大数据管理与传统数据管理的异同

##### 1.2.4 如何让萝卜卖出“奢侈品”的价格？

##### 1.2.5 数据管理的业界案例

###### [1.2.5.1 电信行业案例](#)

### [1.2.5.2 互联网行业案例](#)

## 1.3 数据管理的范畴

### 1.3.1 有关的标准——DAMA 概述

### 1.3.2 数据标准

### 1.3.3 数据资产

### 1.3.4 数据质量

### 1.3.5 数据安全

## 1.4 数据管理基础

### 1.4.1 数据的数据——元数据

### 1.4.2 元数据

#### [1.4.2.1 数据仓库的元数据](#)

#### [1.4.2.2 HADOOP 的元数据](#)

#### [1.4.2.3 元数据的价值和应用](#)

### 1.4.3 CWM 标准

#### [1.4.3.1 数据仓库的元模型](#)

#### [1.4.3.2 HADOOP 的元模型](#)

## 1.5 数据管理的意义

### 1.5.1 提供大数据管理的“抓手”

### 1.5.2 证明大数据的内部管理水平

## 1.6 数据管理包含数据治理

### 1.6.1 数据治理的概念

### 1.6.2 数据管理包含数据治理

## 1.7 【例】附件-数据治理(DAMA)框架介绍

---

## 2 数据管理基础

---

### 2.1 数据内容

#### 2.1.1 客户数据

#### 2.1.2 产品数据

#### 2.1.3 营销数据

#### 2.1.4 员工数据

#### 2.1.5 服务数据

#### 2.1.5

### 2.2 数据标准的颗粒度

#### 2.2.1 详细数据

#### 2.2.2 汇总数据

#### 2.2.3 衍生数据

### 2.3 数据如何收集

#### 2.3.1 从内部各个 IT 系统获取

#### 2.3.2 外部数据获取方法

#### 2.3.3 批量获取/实时获取

### 2.4 数据清洗过程

#### 2.4.1 数据抽取

#### 2.4.2 数据转换

#### 2.4.3 数据加载

## 2.5 跨平台数据接口标准

### 2.5.1 数据接口的定义

### 2.5.2 数据接口的管理职责

### 2.5.3 接口设计的原则

## 2.6 逻辑模型设计及实例

### 2.6.1 原因

### 2.6.2 逻辑模型设计的原则

### 2.6.3 设计实例

## 2.7 物理模型设计及实例

### 2.7.1 原因

### 2.7.2 物理模型设计原则

### 2.7.3 设计实例

## 2.8 数据字典及实例

### 2.8.1 概念

### 2.8.2 范围和內容

### 2.8.3 设计实例

## 2.9 【案例】某企业的數據资产治理案例

---

## 3 数据建模基础

---

### 3.1 什么是数据模型？

### 3.2 数据仓库的数据模型

- 3.3 大数据系统不用数据模型吗？
- 3.4 数据模型是数据治理的基础
- 3.5 数据接口的标准是数据模型的基础
- 3.6 数据字典概念基础
- 3.7 【例】附件——某企业数据字典举例

---

## 4 建模方法论

---

- 4.1 数据模型基本方法
- 4.2 数据模型与行业应用的关系
- 4.3 第三范式
- 4.4 实体-关系模型
- 4.5 建模的三阶段
  - 4.5.1 概念模型
  - 4.5.2 逻辑模型
  - 4.5.3 物理模型
- 4.6 建立模型的关键问题
  - 4.6.1 逻辑模型与业务的关系
  - 4.6.2 逻辑模型与需求的关系
  - 4.6.3 逻辑模型与业务系统的关系
- 4.7 【例】附件——某金融企业数据建模举例

---

## 5 数据仓库逻辑模型（示例）

---

### 5.1 数据仓库模型概述

### 5.2 某企业数据模型概述及举例

### 5.3 参与人主题域

### 5.4 服务主题域

### 5.5 资源主题域

### 5.6 营销主题域

### 5.7 事件主题域

### 5.8 账务主题域

### 5.9 财务主题域

### 5.10 【例】附件——某企业的数据仓库建模举例

---

## 6 汇总层数据模型

---

### 6.1 什么是汇总层数据模型？

### 6.2 某企业汇总层数据模型及案例

### 6.3 详单汇总

### 6.4 账务汇总

### 6.5 个人用户汇总

### 6.6 集团客户汇总

### 6.7 产品信息汇总等

## 6.8 【例】附件——某企业的数数据质量管理介绍

---

## 7 数据接口设计

---

### 7.1 数据接口设计要求

### 7.2 数据接口分类

### 7.3 数据接口实现机制

### 7.4 接口双方职责

### 7.5 数据转换规则

### 7.6 消息级接口规则

### 7.7 【例】附件——某企业的数数据接口规范举例

---

## 8 建模工具

---

### 8.1 Erwin 建模工具介绍

### 8.2 主要特点

### 8.3 主要功能

### 8.4 Erwin 支持的各种接口

### 8.5 【例】附件——Erwin 工具介绍

---

## 9 数据治理基本内容

---

## 9.1 数据架构管理

### 9.1.1 企业数据模型

### 9.1.2 价值链分析

### 9.1.3 相关数据架构

## 9.2 数据开发

### 9.2.1 分析

### 9.2.2 数据建模

### 9.2.3 数据库设计

### 9.2.4 实施

## 9.3 数据操作管理

### 9.3.1 获取

### 9.3.2 恢复

### 9.3.3 调优

### 9.3.4 保留

### 9.3.4 清除

### 9.3.5

## 9.4 数据安全和管理

### 9.4.1 标准

### 9.4.2 分级

### 9.4.3 管理

### 9.4.4 授权

### 9.4.4 审计

### 9.4.5

## 9.5 参考数据和主数据管理

9.5 外部规范  
.1

9.5 内部规范  
.2

9.5 客户数据  
.3

9.5 产品数据  
.4

9.5 维度管理  
.5

## 9.6 文档和内容管理

9.6 获取和存储  
.1

9.6 备份和恢复  
.2

9.6 内容管理  
.3

9.6 检索  
.4

9.6 保留  
.5

## 9.7 元数据管理

9.7.1 架构

9.7.2 整合

9.7.3 控制

9.7.4 交付

## 9.8 数据质量管理

9.8.1 规范

9.8.2 分析

9.8.3 度量

9.8.4 改进

9.9 其它

## 9.10 【例】附件-数据治理的案例分享

---

## 10 元数据管理

---

### 10.1 从数据仓库说起

#### 10.1.1 数据仓库的技术特征——元数据

#### 10.1.2 数据仓库的元数据应用举例

### 10.2 大数据更需要元数据

#### 10.2.1 大数据的业务特征

#### 10.2.2 大数据更需要元数据

#### 10.2.3 大数据的元数据与数据仓库元数据的关系

#### 10.2.4 HADOOP 内部元数据内容

#### 10.2.5 大数据的元数据管理意义（更清晰的数据脉络）

### 10.3 元数据的标准-CWM

#### 10.3.1 为什么是 CWM?

#### 10.3.2 元模型的概念及举例

#### 10.3.3 CWM 的演进

#### 10.3.4 CWM 的问题

### 10.4 元数据的对外开放

#### 10.4.1 元数据的 API

#### 10.4.2 元数据的开放应用意义

### 10.5 元数据的应用

10.5.1 血缘分析

10.5.2 影响分析

10.5.3 元数据的易用性问题

10.6 元数据运维

10.6.1 元数据的建设和运维

10.6.2 运维更加重要

10.6.3 运维的管理制度保障

10.7 元数据的“事前”管理

10.7.1 事前元数据的概念

10.7.2 降低元数据的同步难度

10.7.3 从事前开始

10.8 元数据应用案例分享

10.9 【例】附件-某企业元数据应用案例展示

---

11 数据资产管理

---

11.1 数据成为资产

11.1.1 数据如何能成为资产？

11.1.2 数据资产如何评估？

11.2 数据资产范围

11.2.1 数据资产范围

11.2.2 数据资产的梳理

11.2.3 数据资产与固定资产间的关系

11.3 数据资产管理的内容

11.3.1 资产的注册

11.3.2 资产的变更

11.3.3 资产的统计

11.3.4 资产的审计

11.3.5 资产的权限管理

11.4 数据资产审计

11.4.1 数据资产审计内容

11.4.2 数据资产审计的意义

11.4.3 数据资产审计的延伸

11.5 数据资产使用分析

11.5.1 哪些数据的访问量更多？

11.5.2 哪些数据资产的存储成本过高？

11.5.3 哪些数据资产的价值更大？

11.5.4 资产使用中有哪些问题？

11.6 【例】附件-数据资产管理的具体案例分享

---

12 数据质量管控

---

12.1 数据质量管理的源起

12.1.1 从“挑战者”航天飞机说起

- 12.1.2 60%的精力在数据质量上
- 12.1.3 数据质量的表现形式
- 12.2 数据质量管控的内容
  - 12.2.1 数据处理全程的全监控
  - 12.2.2 数据质量的“度量 ”
  - 12.2.3 数据质量监控的方法
  - 12.2.4 数据质量中的“数据回滚 ”
  - 12.2.5 数据质量管控的“实时性 ”要求
- 12.3 数据质量的知识库
  - 12.3.1 数据质量管理的知识库
  - 12.3.2 “众人拾柴火焰高 ”——群策群力
  - 12.3.3 知识库的表征方式
- 12.4 数据质量管理的业务应用
  - 12.4.1 数据质量评估应用
  - 12.4.2 数据质量分级告警
- 12.5 数据质量管理的制度建设
  - 12.5.1 “3 分技术，7 分管理 ”
  - 12.5.2 如何细化管理？
- 12.6 数据质量管理的“窍门 ”
- 12.7 【例】附件-某企业数据质量管理案例

---

## 13 主数据的概念

---

## 13.1 概念和特点

### 13.1.1 共享数据

### 13.1.2 维度或者层次

## 13.2 主数据管理

### 13.2.1 主数据内容

### 13.2.2 数据保持一致性

### 13.2.3 数据管理流程控制

### 13.2.4 主数据与数据仓库

## 13.3 主数据管理意义

### 13.3.1 关键数据的统一

### 13.3.2 业务的一致性

## 13.4 【案例】附件——某企业主数据管理案例

---

## 14 大数据安全管理

---

## 14.1 从数据仓库的安全说起

### 14.1.1 安全颠覆使用

### 14.1.2 数据仓库安全介绍

## 14.2 系统的安全和网络的安全

### 14.2.1 安全的层级

### 14.2.2 系统安全

### 14.2.3 网络安全

## 14.3 HADOOP 的安全特点

### 14.3.1 开源系统如何安全？

### 14.3.2 HADOOP 各个组件的安全防护

### 14.3.3 安全管控的方案选择

## 14.4 数据的脱敏

### 14.4.1 数据安全的最后防线

### 14.4.2 数据脱敏的算法

### 14.4.3 不可逆算法的优势

## 14.5 “逐级设防”的安全管控策略

### 14.5.1 如何层层防御？

### 14.5.2 “城防图”的设计

## 14.6 安全管理的制度建设

### 14.6.1 从安全管理开始

### 14.6.2 安全管理的制度设计举例

## 14.7 大数据安全的反思

### 14.7.1 有绝对的数据安全吗？

### 14.7.2 如何有限的成本用到刀刃上？

## 14.8 【案例】附件——某企业大数据安全管理案例

---

## 15 数据管理中的互联网思维落地

---

## 15.1 互联网思维的概述

### 15.1.1 互联网思维的概念

### 15.1.2 互联网思维的特点

### 15.1.3 互联网思维的应用举例

## 15.2 大数据与互联网思维的关系

### 15.2.1 大数据为什么需要互联网思维？

### 15.2.2 大数据互联网思维与 IT 思维的对比

### 15.2.3 大数据互联网思维的意义

## 15.3 数据管理中的互联网思维

### 15.3.1 数据管理的使用者分析

### 15.3.2 数据管理中的产品分析

### 15.3.3 数据管理中极致思维的落地

## 15.4 元数据的互联网思维

### 15.4.1 用户是谁？

### 15.4.2 产品如何？

### 15.4.3 如何降低使用门槛？

## 15.5 数据质量的互联网思维

### 15.5.1 互联网思维在数据质量管理中的落地

### 15.5.2 如何提升数据质量监控的可视化程度？

## 15.6 数据资产管理的互联网思维

### 15.6.1 如何让普通用户也能够理解数据资产管理？

### 15.6.2 数据资产管理中的跨界思维

## 15.7 数据安全中的互联网思维

### 15.7.1 数据安全的产品化

### 15.7.2 数据安全的“可视化”

## 15.8 【例】附件-数据治理中互联网思维的落地案例

---

## 16 总结

---