

# 《数据架构》

## 图书基本信息

书名：《数据架构》

13位ISBN编号：9787115438439

出版时间：2016-11

作者：[美] W.H. Inmon,[美] Daniel Linstedt

页数：277

译者：唐富年

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu111.com](http://www.tushu111.com)

# 《数据架构》

## 内容概要

本书是数据仓库之父Inmon的新作，探讨数据的架构和如何在现有系统中最有效地利用数据。本书的主题涵盖企业数据、大数据、数据仓库、Data Vault、业务系统和架构。主要内容包括：在分析和大数据之间建立关联，如何利用现有信息系统，如何导出重复型数据和非重复型数据，大数据以及使用大数据的商业价值，等等。

本书的读者对象包括数据工程技术人员、管理人员以及从事数据分析和研究的科研人员。

# 《数据架构》

## 作者简介

W.H. Inmon

数据仓库之父，最早的数据仓库概念提出者，在数据库技术管理与数据库设计方面拥有30多年的经验。2007年，Inmon被ComputerWorld杂志评为计算机行业史上最具影响力的十大名人之一。

Daniel Linstedt

世界知名数据仓库专家、商业智能分析家，Empowered Holdings公司创始人兼主席，有20余年的IT行业打拼经验。Linstedt还是下一代数据仓库模型Data Vault的发明者。

## 书籍目录

第1章 企业数据	1
1.1 企业数据	1
1.1.1 企业的全体数据	1
1.1.2 非结构化数据的划分	2
1.1.3 业务相关性	3
1.1.4 大数据	3
1.1.5 分界线	4
1.1.6 大陆分水岭	5
1.1.7 企业数据全貌	6
1.2 数据基础设施	6
1.2.1 重复型数据的两种类型	7
1.2.2 重复型结构化数据	7
1.2.3 重复型大数据	8
1.2.4 两种基础设施	9
1.2.5 优化了什么	10
1.2.6 对比两种基础设施	11
1.3 分界线	12
1.3.1 企业数据分类	12
1.3.2 分界线	12
1.3.3 重复型非结构化数据	13
1.3.4 非重复型非结构化数据	15
1.3.5 不同的领域	17
1.4 企业数据统计图	17
1.5 企业数据分析	22
1.6 数据的生命周期——随时间推移理解数据	27
1.7 数据简史	31
1.7.1 纸带和穿孔卡片	31
1.7.2 磁带	32
1.7.3 磁盘存储器	32
1.7.4 数据库管理系统	32
1.7.5 耦合处理器	33
1.7.6 在线事务处理	33
1.7.7 数据仓库	34
1.7.8 并行数据管理	34
1.7.9 Data Vault	35
1.7.10 大数据	35
1.7.11 分界线	35
第2章 大数据	37
2.1 大数据简史	37
2.1.1 打个比方——占领制高点	37
2.1.2 占领制高点	38
2.1.3 IBM360带来的标准化	38
2.1.4 在线事务处理	39
2.1.5 Teradata的出现和大规模并行处理	39
2.1.6 随后到来的Hadoop和大数据	39
2.1.7 IBM和Hadoop	39
2.1.8 控制制高点	40

2.2	大数据是什么	40
2.2.1	另一种定义	40
2.2.2	大数据量	40
2.2.3	廉价存储器	41
2.2.4	罗马人口统计方法	41
2.2.5	非结构化数据	42
2.2.6	大数据中的数据	42
2.2.7	重复型数据中的语境	43
2.2.8	非重复型数据	44
2.2.9	非重复型数据中的语境	44
2.3	并行处理	45
2.4	非结构化数据	50
2.4.1	随处可见的文本信息	50
2.4.2	基于结构化数据的决策	51
2.4.3	业务价值定位	51
2.4.4	重复型和非重复型的非结构化信息	52
2.4.5	易于分析	53
2.4.6	语境化	54
2.4.7	一些语境化方法	55
2.4.8	MapReduce	56
2.4.9	手工分析	56
2.5	重复型非结构化数据的语境化	57
2.5.1	解析重复型非结构化数据	57
2.5.2	重组输出数据	58
2.6	文本消歧	58
2.6.1	从叙事到分析数据库	58
2.6.2	文本消歧的输入	59
2.6.3	映射	60
2.6.4	输入/输出	61
2.6.5	文档分片/指定值处理	61
2.6.6	文档预处理	62
2.6.7	电子邮件——一个特例	62
2.6.8	电子表格	63
2.6.9	报表反编译	63
2.7	分类法	65
2.7.1	数据模型和分类法	65
2.7.2	分类法的适用性	66
2.7.3	分类法是什么	66
2.7.4	多语言分类法	68
2.7.5	分类法与文本消歧的动态	68
2.7.6	分类法和文本消歧——不同的技术	69
2.7.7	分类法不同类型	70
2.7.8	分类法——随时间推移不断维护	70
第3章	数据仓库	71
3.1	数据仓库简史	71
3.1.1	早期的应用程序	71
3.1.2	在线应用程序	71
3.1.3	抽取程序	72
3.1.4	4GL技术	73

3.1.5	个人电脑	73
3.1.6	电子表格	74
3.1.7	数据完整性	75
3.1.8	蛛网系统	76
3.1.9	维护积压	77
3.1.10	数据仓库	78
3.1.11	走向架构式环境	78
3.1.12	走向企业信息工厂	78
3.1.13	DW 2.0	79
3.2	集成的企业数据	81
3.2.1	数量众多的应用程序	81
3.2.2	放眼企业	82
3.2.3	多个分析师	83
3.2.4	ETL技术	84
3.2.5	集成的挑战	86
3.2.6	数据仓库的效益	86
3.2.7	粒度的视角	87
3.3	历史数据	89
3.4	数据集市	92
3.4.1	颗粒化的数据	92
3.4.2	关系数据库设计	93
3.4.3	数据集市	93
3.4.4	关键性能指标	94
3.4.5	维度模型	94
3.4.6	数据仓库和数据集市的整合	95
3.5	作业数据存储	96
3.5.1	集成数据的在线事务处理	96
3.5.2	作业数据存储	97
3.5.3	ODS和数据仓库	98
3.5.4	ODS分类	99
3.5.5	将外部数据更新到ODS	99
3.5.6	ODS/数据仓库接口	100
3.6	对数据仓库的误解	101
3.6.1	一种简单的数据仓库架构	101
3.6.2	在数据仓库中进行在线高性能事务处理	101
3.6.3	数据完整性	102
3.6.4	数据仓库工作负载	102
3.6.5	来自数据仓库的统计处理	103
3.6.6	统计处理的频率	104
3.6.7	探查仓库	104
第4章	Data Vault	106
4.1	Data Vault简介	106
4.1.1	Data Vault 2.0建模	107
4.1.2	Data Vault 2.0方法论定义	107
4.1.3	Data Vault 2.0架构	107
4.1.4	Data Vault 2.0实施	108
4.1.5	Data Vault 2.0商业效益	108
4.1.6	Data Vault 1.0	109
4.2	Data Vault建模介绍	110

4.2.1	Data Vault模型概念	110
4.2.2	Data Vault模型定义	110
4.2.3	Data Vault模型组件	111
4.2.4	Data Vault和数据仓库	112
4.2.5	转换到Data Vault建模	112
4.2.6	数据重构	113
4.2.7	Data Vault建模的基本规则	114
4.2.8	为什么需要多对多链接结构	114
4.2.9	散列键代替顺序号	115
4.3	Data Vault架构介绍	116
4.3.1	Data Vault 2.0架构	116
4.3.2	如何将NoSQL适用于本架构	117
4.3.3	Data Vault 2.0架构的目标	117
4.3.4	Data Vault 2.0建模的目标	118
4.3.5	软硬业务规则	118
4.3.6	托管式SSBI与DV2架构	119
4.4	Data Vault方法论介绍	120
4.4.1	Data Vault 2.0方法论概述	120
4.4.2	CMMI和Data Vault 2.0方法论	120
4.4.3	CMMI与敏捷性的对比	122
4.4.4	项目管理实践和SDLC与CMMI和敏捷的对比	123
4.4.5	六西格玛和Data Vault 2.0方法论	123
4.4.6	全质量管理	124
4.5	Data Vault实施介绍	125
4.5.1	实施概述	125
4.5.2	模式的重要性	126
4.5.3	再造工程和大数据	127
4.5.4	虚拟化我们的数据集市	128
4.5.5	托管式自助服务BI	128
第5章	作业环境	130
5.1	作业环境——简史	130
5.1.1	计算机的商业应用	130
5.1.2	最初的应用程序	131
5.1.3	Ed Yourdon和结构化革命	132
5.1.4	系统开发生命周期	132
5.1.5	磁盘技术	132
5.1.6	进入数据库管理系统时代	133
5.1.7	响应时间和可用性	133
5.1.8	现代企业计算	136
5.2	标准工作单元	136
5.2.1	响应时间要素	136
5.2.2	沙漏的比喻	137
5.2.3	车道的比喻	138
5.2.4	你的车跑得跟前面的车一样快	139
5.2.5	标准工作单元	139
5.2.6	服务等级协议	139
5.3	面向结构化环境的数据建模	140
5.3.1	路线图的作用	140
5.3.2	只要粒度化的数据	140

5.3.3	实体关系图	141
5.3.4	数据项集	142
5.3.5	物理数据库设计	143
5.3.6	关联数据模型的不同层次	143
5.3.7	数据联动的示例	144
5.3.8	通用数据模型	146
5.3.9	作业数据模型和数据仓库数据模型	146
5.4	元数据	146
5.4.1	典型元数据	146
5.4.2	存储库	147
5.4.3	使用元数据	148
5.4.4	元数据用于分析	149
5.4.5	查看多个系统	150
5.4.6	数据谱系	150
5.4.7	比较已有系统和待建系统	150
5.5	结构化数据的数据治理	151
5.5.1	企业活动	151
5.5.2	数据治理的动机	152
5.5.3	修复数据	152
5.5.4	粒度化的详细数据	153
5.5.5	编制文档	153
5.5.6	数据主管岗位	154
第6章	数据架构	156
6.1	数据架构简史	156
6.2	大数据/已有系统的接口	166
6.2.1	大数据/已有系统的接口	166
6.2.2	重复型原始大数据/已有系统接口	167
6.2.3	基于异常的数据	168
6.2.4	非重复型原始大数据/已有系统接口	169
6.2.5	进入已有系统环境	170
6.2.6	“语境丰富”的大数据环境	171
6.2.7	将结构化数据/非结构化数据放在一起分析	172
6.3	数据仓库/作业环境接口	172
6.3.1	作业环境/数据仓库接口	172
6.3.2	经典的ETL接口	173
6.3.3	作业数据存储/ETL接口	173
6.3.4	集结区	174
6.3.5	变化数据的捕获	175
6.3.6	内联转换	175
6.3.7	ELT处理	176
6.4	数据架构——一种高层视角	177
6.4.1	一种高层视角	177
6.4.2	冗余	177
6.4.3	记录系统	178
6.4.4	不同的群体	180
第7章	重复型分析	181
7.1	重复型分析——必备基础	181
7.1.1	不同种类的分析	181
7.1.2	寻找模式	182



7.1.3	启发式处理	183
7.1.4	沙箱	186
7.1.5	标准概况	187
7.1.6	提炼、筛选	188
7.1.7	建立数据子集	188
7.1.8	筛选数据	190
7.1.9	重复型数据和语境	192
7.1.10	链接重复型记录	193
7.1.11	日志磁带记录	193
7.1.12	分析数据点	194
7.1.13	按时间的推移研究数据	195
7.2	分析重复型数据	196
7.2.1	日志数据	198
7.2.2	数据的主动/被动式索引	199
7.2.3	汇总/详细数据	200
7.2.4	大数据中的元数据	202
7.2.5	相互关联的数据	203
7.3	重复型分析	204
7.3.1	内部、外部数据	204
7.3.2	通用标识符	205
7.3.3	安全性	205
7.3.4	筛选、提炼	207
7.3.5	归档结果	208
7.3.6	指标	210
第8章	非重复型分析	211
8.1	非重复型数据	211
8.1.1	内联语境化	213
8.1.2	分类法/本体处理	214
8.1.3	自定义变量	215
8.1.4	同形异义消解	216
8.1.5	缩略语消解	217
8.1.6	否定分析	218
8.1.7	数字标注	219
8.1.8	日期标注	220
8.1.9	日期标准化	220
8.1.10	列表的处理	220
8.1.11	联想式词处理	221
8.1.12	停用词处理	222
8.1.13	提取单词词根	222
8.1.14	文档元数据	223
8.1.15	文档分类	223
8.1.16	相近度分析	224
8.1.17	文本ETL中功能的先后顺序	225
8.1.18	内部参照完整性	225
8.1.19	预处理、后处理	226
8.2	映射	227
8.3	分析非重复型数据	229
8.3.1	呼叫中心信息	229
8.3.2	医疗记录	237

# 《数据架构》

第9章	作业分析1	242	
第10章	作业分析2	249	
第11章	个人分析	259	
第12章	复合式的数据架构	264	
词汇表	268		

# 《数据架构》

## 精彩短评

- 1、质量较差。没有太多营养。感觉全书在东凑凑，西凑凑，凑字数
- 2、太偏向于传统资讯行业的方法论，实操性很差，基本没有参考意义。对于高速公路的比喻还是比较恰当，是加星的唯一理由。

## 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:[www.tushu111.com](http://www.tushu111.com)