“数据仓库之父”Bill Inmon最新力作

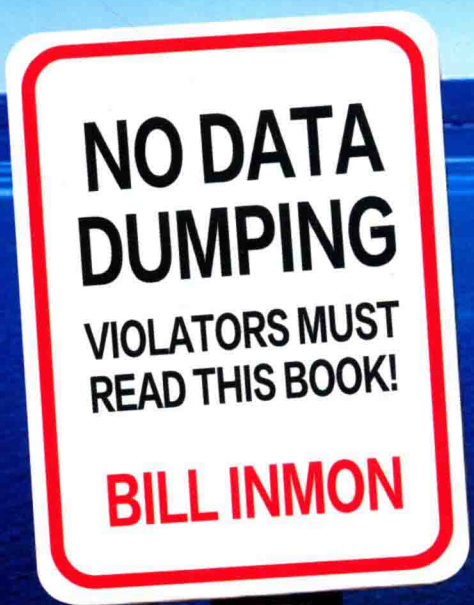
设计数据湖以避免垃圾存储

异步图书
www.epubit.com.cn

数据湖架构

DATA LAKE ARCHITECTURE

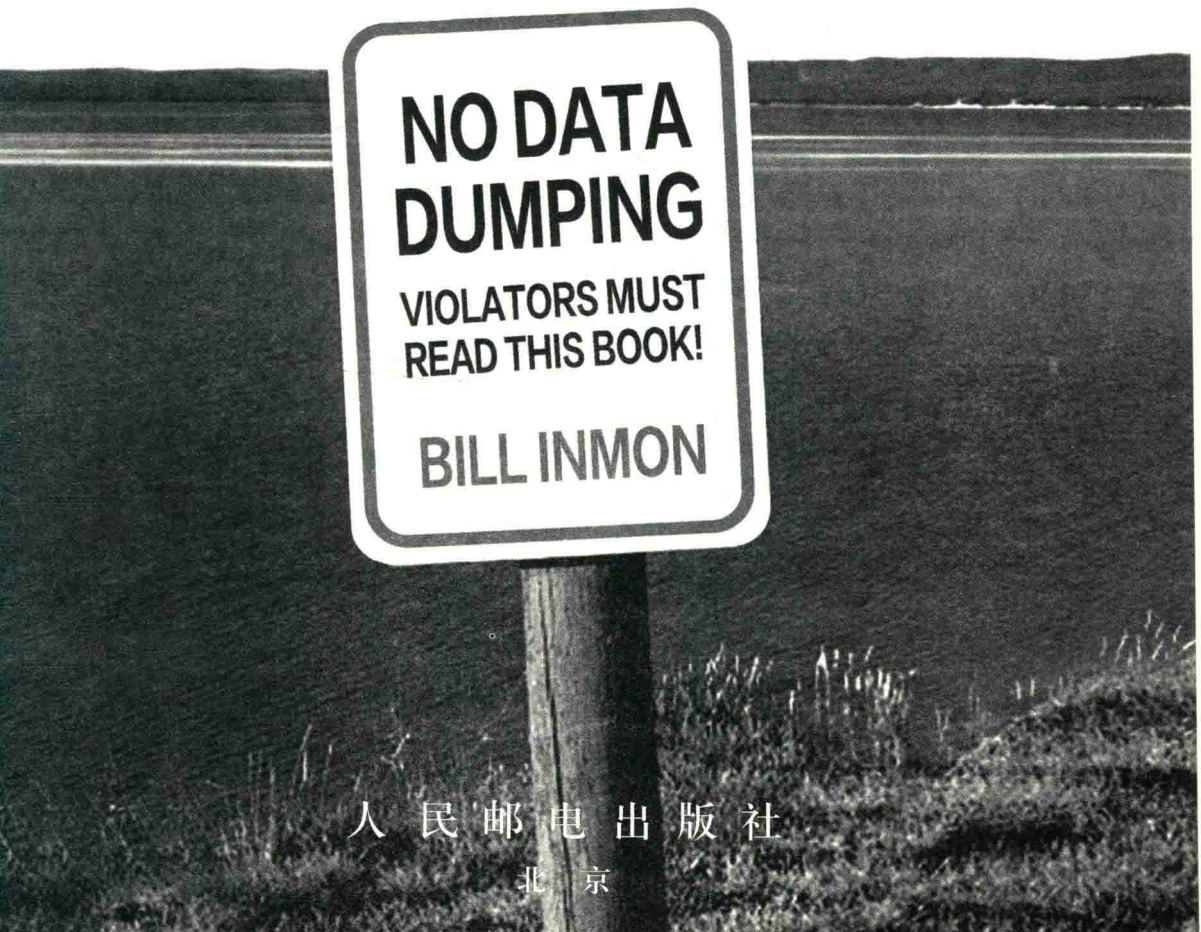
[美] Bill Inmon 著　吴文磊 译



中国工信出版集团　人民邮电出版社
POSTS & TELECOM PRESS

数据湖架构
DATA LAKE ARCHITECTURE

[美] Bill Inmon 著　吴文磊 译



人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

数据湖架构 / (美) 恩门 (Bill Inmon) 著 ; 吴文磊译. -- 北京 : 人民邮电出版社, 2017.5
ISBN 978-7-115-45173-6

Ⅰ. ①数… Ⅱ. ①恩… ②吴… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第067778号

版权声明

◆ 著 [美] Bill Inmon
译 吴文磊
责任编辑 陈冀康
责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：720×960 1/16
印张：10
字数：123 千字 2017 年 5 月第 1 版
印数：1 – 2 000 册 2017 年 5 月河北第 1 次印刷

著作权合同登记号 图字：01-2016-3963 号

定价：49.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

随着大数据的蓬勃发展，不少机构开始将源源不断的数据流导入到一个叫“数据湖”的设备中去。

本书是“数据仓库”之父撰写的最新著作，是帮助读者认识数据湖架构，并把数据湖打造成公司资产的指导手册。全书共15章，分别涉及数据湖简介、数据池据湖内部结构、数据池及其结构、各种类型的数据池等技术话题，目的在于讲解如何构建有用的数据湖，以便数据科学家和数据分析师能够解决商业挑战并找出新的商业机会。

本书适合数据管理者、学生、系统开发人员、架构师、程序员以及最终用户阅读。

前言

在错误的方向上，我们耗费了数年时间，花费了上百万美元，但是，我们是不是可以省出一点儿时间和金钱用到正确的方向上来呢？

如今，众多公司正在疯狂地建设数据湖泊——一种大数据狂热的副产品。有朝一日，这些公司幡然醒悟，发现他们根本不能从数据湖中攫取出任何有用的东西。即便真的从数据湖中找到了一丁点儿有用的信息，起码也要经历呕心沥血的努力。

他们花费了巨额的资金和大量人年（man years）的努力，却只换回了昂贵的累赘。

终有一天，这些企业会惊觉于他们所建造的不过是一个“单

向”的数据湖。数据被引入数据湖，却产生不了任何东西。在这种情况下，数据湖不会比垃圾场好到哪儿去。

这本书就是写给那些想要建造数据湖，并期望能够从中获得价值的机构。数据湖中当然有业务价值，但前提是建造得法。如果你正打算建造一个数据湖，那么你最好把它建造成公司的一项重要资产，而不是累赘。

本书探究了为什么众多公司在从他们的数据湖中获取数据时会面临如此艰难的困境。关于这个重要问题有数种答案。其中一个原因是，数据被不加区别地一股脑地打包丢入数据湖中。第二个原因是数据没有被整合起来。第三个原因是数据是以文本化的形式保存的，而你没办法轻易地分析文本数据。

本书建议要以高层（high level）的视角来组织数据，整合数据，“调校”数据，其目的就是使调整后的数据能够成为用于分析和处理的基石。数据湖当然可以成为公司的良性资产，但前提是在构建数据湖时要足够谨慎，并深谋远虑。

数据湖需要被划分成几个被称为数据池（data pond）部分，它们是：

- 初始数据池（Raw data pond）；
- 模拟信号数据池（Analog data pond）；
- 应用程序数据池（Application data pond）；
- 文本数据池（Textual data pond）；
- 归档数据池（Archival data pond）。

在创建之后，数据池需要经历调整过程，使数据容易访问，以便进一步加以利用。举例来说，模拟信号数据池需要对数据进行缩减（reduction）和压缩。应用程序数据池需要让数据经历经典的 ETL 整合。文本数据池则需要对文本进行消歧，以便使文本可以规整成一致的数据库结构，这样，文本所在的语境就可以被识别出来。

一旦数据池中的数据经历过算法的调整，那么该数据池就可以作为基础，为分析和处理流程提供服务。一旦数据湖中的数据被区划成不同的数据池，并且数据在池中经历了调整，那么这些数据池就会成为公司的资产，而不是负累。此外，当数据走完了它在数据池中的生命周期，它就会被移入归档数据池。

这本书是写给管理者、学生、系统开发人员、架构师、程序员以及最终用户的，并希望能成为那些想把数据湖打造成公司资产而非负担的机构的指导手册。

目录

第 1 章
数据的湖泊

打孔卡（punch card）被发明出来之后，磁带（tape）被发明出来，然后是磁盘存储（disk storage）和数据库管理系统（DBMS），紧跟着的是第 4 代编程语言（4GL）、“元数据”、软盘和移动计算。技术前进得如此之快，以至于我们甚至来不及记清楚它们的名字。很快，个人电脑和电子表单就会像西装和领带一样随处可见。

在这高速发展的几十年里，公司经历了从没有自动化到高度自动化的转变。但在转变过程中，存储却始终是一项制约因素。长久以来，在面对大量数据的时候，存储不是容量不够就是价格太高。这个瓶颈制约了既有系统的性能，并且对系统未来的可选方案产生了深刻的影响。

1.1 大数据来了

随后，大数据技术改变了世界。Hadoop 分布式文件系统（HDFS）是大数据技术最好的代表。这个开源软件框架的设计初衷就是解决在分布计算集群中的存储和处理大量数据集的难题。大数据技术有效地解放了存储包括在价格和技术能力上的限制。更为重要的是，在大数据技术的帮助下，一个全新的世界正向我们敞开大门。

简单来说，大数据刷新了我们对数据的认识。激增的数据可以被大数据系统保存并分析，这不仅是一项工业界的革命，更是一次世界性的革命。MB、GB、TB……旧有的数据量单位在这个存储容量被解放了的新世界中已不再适用。图 1.1 描绘了大数据降临的场景。

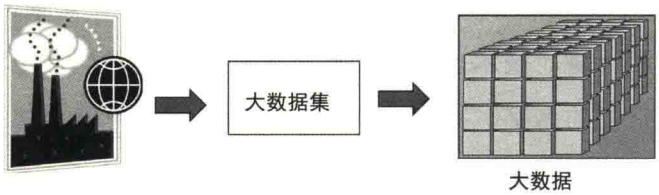


图 1.1 利用大数据创造无限机遇

1.2 数据湖来了

随着大数据的蓬勃发展，不少机构开始将源源不断的数据流

导入到一个叫做“数据湖”的设备中去。

把数据放进去是小菜一碟儿，然而，想从这浩瀚的知识海洋中拽出点什么有用的东西却极具挑战。一些机构开始向数据科学家们寻求帮助。于是，大量的经费被投入研发，然而，如同这些机构一样，大数据对于数据科学家们而言也是一个全新的领域。尽管投入高昂，但分析上难有突破，而误报和其他错误倒是时有发生。图 1.2 展示的是大数据催生了用广袤的数据湖泊来筛查数据。

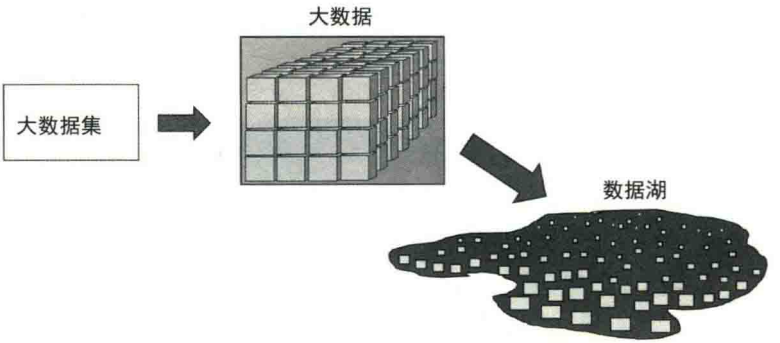


图 1.2 在数据湖中置入大数据

图 1.3 所展现的是在数据湖中，过去商业社会所崇尚的规模产生价值在数据湖中的失效。对于数据湖来说，数据确实在持续增长，却很难用财富堆积出其中的价值。

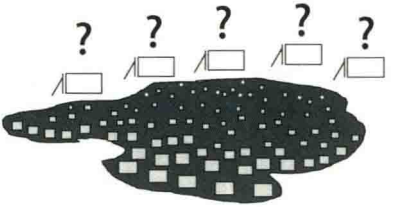


图 1.3 醒来吧，我们没能在数据湖中找到任何东西

1.3 “单向”的数据湖

业务用户会对数据湖中池化（pooling）的信息感到一筹莫展的原因有很多。核心的问题在于，湖中的数据增长得越多，其分析难度也越大。任何规模可观的数据湖都常常会被人戏谑为“单向湖”，因为数据被不断地推进湖里，但分析报告却始终难产，或者数据被推入湖中之后仅被访问一次。图 1.4 描绘了“单向”数据湖。

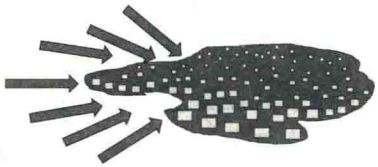


图 1.4 数据被推入“单向”数据湖，但什么也产出不了

这是一项昂贵且令人沮丧的悖论（Catch-22）。数据湖成长得越大，就越具有潜在的洞察能力，但对于机构而言，却越无用（useless）。如果没人去使用数据湖中的数据，那么数据湖对机构就毫无意义。然而，为了从数据湖中榨取出有用的信息，机构却在存储和雇佣专业人员上投入了大量资金。

那么问题来了，为什么数据湖会变成“单向”湖，对此，我们又能做些什么呢？大数据和数据湖中确实蕴含着巨大的潜力，但似乎没有人能从他们的投资中获得与其相当的回报。数据湖变成“单向”数据湖有很多原因。但追根溯源，这些问题都指向同