<<第四范式>>

图书基本信息

书名：<<第四范式>>

13位ISBN编号：9787030347251

10位ISBN编号：7030347250

出版时间：2012-6

出版时间：科学出版社

作者：潘教峰、张晓林

页数：247

字数：264500

<<第四范式>>

内容概要

《第四范式：数据密集型科学发现》系统介绍了地球与环境科学、生命与健康科学、数字信息基础设施和数字化学术信息交流等方面基于海量数据的科研活动、过程、方法和基础设施，生动揭示了在海量数据和无处不在网络上发展起来的与实验科学、理论推演、计算机仿真这三种科研范式相辅相成的科学研究第四范式——数据密集型科学发现，进一步探讨了这种新范式的内涵和内容，包括利用多样化工具不间断采集科研数据、建立系统化工具和设施来管理整个数据生命周期、开发基于科学研究问题的数据分析及可视化工具与方法等，并深入探讨了这种新范式对科学研究、科学教育、学术信息交流及科学家群体的长远影响。

《第四范式：数据密集型科学发现》将帮助从事科学研究、科技研究规划、科技政策等领域的科研人员和管理者理解和把握科研环境与科研方法的革命性变化，也将为学术出版、文献情报、科学数据及其他从事信息与知识管理的人士提供未来的战略视角，同时也有助于有志于科学研究和学术信息交流管理的高层次学生了解未来的挑战和需求。

作者简介

无

<<第四范式>>

书籍目录

<<第四范式>>

章节摘录

 插图： 大多数的科学数据分析以分级步骤进行。
在第一步中，对数据子集进行抽取，这一工作要通过过滤某些属性（如去除错误的数据）或抽取数据列的垂直子集完成。
在接下来的步骤中，通常以某种方式转换或聚合数据。
当然，在更复杂的数据集中，这些模式往往伴随着多个数据集的复杂连接，如外部校准或抽取和分析一个基因序列的不同部分[8]。
随着数据集的日益增大，进行大多数这些计算的最有效方法显然是尽可能地使分析功能与数据密切结合，这也使大多数的模式很容易通过集合型的表述语言来表达，这种语言的运用可以从基于成本的查询优化、自动并行化和索引中获得巨大收益。
格雷及其合作者展示了几个现有关系数据库技术成功应用于这方面的项目[9]。
有一些项目以无缝的方法来整合用程序语言编写的复杂类库，并将其作为底层数据库引擎的扩展[10,11]。
近年来，Map Reduce 2已经成为分布式数据分析和计算的普遍范式[12]。
这种范式的原理类似于分布式分组和聚合的能力，这些能力已经在并行关系数据库系统中存在了一段时间。
新一代的并行数据库系统，如Teradata、Aster Data和Vertica，已经将这些能力重塑为“数据库中的MapReduce”，并开发出可以比较每种方法优点的新基准[13]。
与科学家连接 设计科学数据库面临的最具挑战性的问题是在数据库建设者和对分析感兴趣的专门领域科学家（domainscientists）之间建立起有效的交流。
但大多数项目犯下了竭力追求“为所有人做所有事”（everything for everyone）的错误。
显然，有一些特征要比其他一些特征更重要。
因此，有必要对不同设计进行折中，当然，这也导致性能的折中。
吉姆·格雷提出了“20个询问”的启发式规则。
在他参与的每一个项目中，他都寻求研究人员想让数据系统回答的最重要的20个问题。
他认为，5个问题不足以识别广泛的模式，100个问题将导致重点不突出。
由于与人2译者注：Map Reduce是Google开发的分布式计算模型，在处理T级别以上巨量数据业务时有显著优势。
类选择有关的大多数决定都遵循“长尾理论”（或所谓的1/f分布），询问中的相关信息根据重要性排序显然是呈对数分布，大约在20（24.5）～100（26.5）范围内实现增益是适中的[14]。
“20个询问”规则是一种设计步骤的别称，这种步骤使专门领域科学家和数据库设计者可以对话，填补科学领域中使用的名词和动词之间，以及数据库中存储的实体和关系之间的语义鸿沟。
这些询问定义了专门领域科学家期望对数据库提出的有关实体和关系方面的精确问题集。
这种重复实践的结果是：专门领域科学家和数据库之间可以使用共同的语言。
这种方法非常成功地使设计过程聚焦于系统必须支持的最重要特征，同时帮助专门领域科学家理解数据库系统的折中，从而限制“特征的蠕变”。

<<第四范式>>

编辑推荐

《第四范式：数据密集型科学发现》以吉姆·格雷提出科学研究第四范式的著名演讲开篇，邀请国际著名科学家对数据密集型科学发现的理念、应用和影响进行了全面分析。
第一部分，Dan Fay等人介绍了地球、环境、海洋、空间等领域的大数据环境与科学应用；第二部分，Simon Mercer等人分析了医学、认知科学、生物系统、医疗服务等领域的数据密集型科学发现；第三部分，Daron Green等人提出了适应大数据时代的科学信息与科学计算基础设施面临的挑战；第四部分，Lee Dirks等人对数据密集型科学发现给学术信息交流带来的深刻变化做了描述。
全书视野开阔、思考深邃，既把握大势，又深入具体，为把握第四范式的要旨与含义提供了坚实的基础。

<<第四范式>>

版权说明