第 2 届超大数据库会议（XLDB2008）大会报告（中文版）

REPORT FROM THE 2nd WORKSHOP ON EXTREMELY LARGE DATABASES

Jacek Becla*1 and Kian-Tat Lim2
SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
***1 Email: becla@slac.stanford.edu**
2 Email: ktl@slac.stanford.edu

温馨提示：本文由厦门大学计算机系林子雨老师翻译自 XLDB 会议网站的英文报告，转载请注明出处，仅用于学习交流，请勿用于商业用途。
[本文翻译的原始出处：厦门大学计算机系数据库实验室网站林子雨老师的超大数据库技术资料专区 http://dblab.xmu.edu.cn/XLDB]
翻译者林子雨个人主页：http://www.cs.xmu.edu.cn/linziyu

【摘要】在科学界和业界，大规模分析的复杂性已经在近些年有了很大的提升。分析人员正在努力尝试使用复杂的技术，比如时间序列分析和分类算法，因为他们平时所熟悉的工具，虽然功能强大，但是可扩展性较差，无法有效使用可扩展的数据库系统。第 2 届 XLDB 大会，主要目的在于了解这些存在的问题，剖析这些问题的背后原因，并寻找相应的解决方案。大会还讨论了建设一个新的开源科学数据库 SciDB，这个构想是在第 1 届 XLDB 大会（XLDB2007）上提出来的。本文是本次大会活动和讨论的总结报告。
【关键词】分析;数据库;千兆级;亿亿次级;大型数据库;超大数据库;

1 大会总结

第 2 届 XLDB 大会，主要关注大规模复杂分析。与会代表包括数据密集型的科学和产业应用领域、数据库研究群体和数据库厂商。

复杂分析。大会讨论了许多复杂分析任务的实例。产业应用通常集中在寻找和发现客户行为模式。这些产业分析所采用的工具，和科学家在执行发现模式和异常的分析时所使用的工具一样，比如时间序列分析和分类。

数据集的大小正在急剧增加，增长率也在不断提高。当前一个最大的项目，每年都会新增加数万 PB 的数据。类似 R、MATLAB 和 EXCEL 等工具无法跟上数据增长的步伐，这使得分析人员不得不生成可以放入内存的样本数据，而不是使用所有数据。这些超大数据集的结构和针对这些数据集的应用，已经变得越来越复杂，因此，XLDB 系统必须在数据表示、处理和硬件方面，都能够保持灵活性。有一种可能的方法，虽然需要某种程度的文化改变，但是可以最大化灵活性并同时降低代价，这种方法是，以服务的方式提供分析工具，即使用一个中央 XLDB 来支持分散在各地的、具有不同分析需求的群体。必须避免管理成本随着数据量的快速增加而增加，因此，需要一个在发生硬件故障时仍然能够正常运行的自适应系统。

SQL 的面向集合的特性和行级别的 ODBC/JDBC 接口，给使用数据库的人员造成了一定的障碍。基于数组的数据模型，在直观上和科学界以及业界遇到的数据类型比较匹配。和

分析工具的集成，以及和编程语言（比如 C++和 DLL）的集成，也会有一定的帮助。发明一种能够直接表达分析人员意图的语言，也是可能的，尽管这个语言被接纳的过程会有一些坎坷。面向过程的 MapReduce 阵营和面向声明语句的数据库阵营，已经开始逐渐意识到对方的优点，开始走向融合。

随着分析变得越来越复杂，涉及的数据量越来越多，分析工作流及其结果的可再现性，就变得尤为重要。虽然起源和再现性通常和科学界相关，业界现在也开始认识到这些特性的重要型，这些特性很容易在数据库中进行处理。但是，同时我们也要注意到，完美的再现性需要高昂的代价，甚至是不可能做到的，因此，有选择性地放松一致性保证也是很重要的。**SciDB**。最初的 XLDB 活动达成了一个共识，那就是建设一个开放的开源科学数据库 SciDB。到目前为止，SciDB 创建者已经确定了最初的合作伙伴，组建了一个数据库研究智囊团，收集了详细的用户用例，完成了最初的设计，募集了经费，建立了一个非盈利的机构，并且开始招募工作人员。SciDB 的设计，采用了层次结构的、多维数组数据模型，具备相关的数组操作，这些操作和关系数据库中的操作完全对应。查询是采用解析树的形式进行表达，预期可以绑定到 MATLAB、C++、Python、IDL 和其他工具和编程语言上。SciDB 将运行在增量可扩展的、由商业服务器构成的簇或云上面。它提供了一种可选的方式，允许对数据进行就地分析，而不需要什么数据加载过程。

SciDB 是由一个非盈利机构进行管理的。由 Mike Stonebraker 和 David DeWitt 领导的智囊团，负责 SciDB 的设计工作。科学研究人员和高端商业用户的大量参与，使得该智囊团队可以获得详细的应用需求和用例，从而使得初始设计更加有效。设计工作还得到了大的工业界合作伙伴的支持，比如 EBay 和 Microsoft。SciDB 系统的第一个版本在 2010 年底可以投入使用。

下一步工作。与会者达成一个共识，SciDB 应该独立于 XLDB。未来将会创建一个科学难题，用来对不同的 XLDB 系统（包括 SciDB）进行基于统一标准的性能测试。下一届 XLDB 大会计划在 CERN（欧洲核子研究会）举行，从而可以充分利用 CERN 和 VLDB 大会举办地（法国，里昂，2009 年 8 月）比较接近的优点。XLDB2009 大会的主要目标，就是加强和非美国的 XLDB 群体的联系，并且涉足更多的科学领域。

2 关于大会

第 2 届 XLDB 大会继续提供一系列论坛，来讨论超大数据库问题。大会于 2008 年 9 月 29 日到 30 日在美国加州 Menlo Park 的 SLAC(斯坦福直线加速中心)举行。大会的主要目的在于：

- 继续了解和超大数据库相关的主要困难和挑战，这里重点关注的是复杂分析；
- 继续在 XLDB 不同群体之间构建沟通的桥梁和纽带，这些群体包括科学界、工业界、数据库研究人员和数据库厂商；
- 构建开放的开源 SciDB 社区；

本次大会的网址是 http://www-conf.slac.stanford.edu/xldb08.

附录中给出了大会日程表。

大会组织委员会包括 Jacek Becla / SLAC (chair), Kian-Tat Lim / SLAC, Celeste Matarazzo / LLNL, Mayank Bawa / Aster Data, Oliver Ratzesberger / eBay 和 Aparajeeta Das / LG CNS.

2.1 参会情况

大会采用邀请参会的方式，这样可以使得参与人员数量尽量少，同时可以保证在没有麦

克风的情况下可以充分交流，并且要保证来自不同群体的参会人员数量保持相对平衡。来自工业界（数据库用户和数据库厂商）、科学界（数据库用户）和学术界（数据库研究人员）的 64 人，参加了本次大会。和第一届大会相比，第二届大会参会者中包含了更多的来自学术界的人员。所有参会者的名字和单位信息可以在网站上面得到。

2.2 结构

大会的大部分时间，都用来进行交互式讨论，讨论主要关注超大数据库建设时所需要具备的复杂分析和管理特性。大会的很大一部分讨论都放在了 SciDB，这是一个新的数据库项目，它是由第一届 XLDB 大会发起的，是第一届大会的后续工作。大会讨论中，也有一少部分内容是由 CERN/LHC, Pan-STARRS 和 eBay 等讲述的他们和超大数据集“战斗”的故事，他们讲述了和超大数据集打交道的一些经验教训。

2.3 关于本报告

本报告的结构，和大会日程没有严格对应，因为，我们想涵盖所有的主题和大会的讨论主线。

第 3 节阐述复杂分析，重点给出一些实例，第 4 节讨论对数据表示方法的影响，第 5 节的内容阐述对数据处理的影响，第 6 节讨论 SciDB，第 7 节讨论了大会就下一步工作和未来 XLDB 大会所达成的共识。

我们特意弱化了特定项目和与会者的名字，从而可以抽象出科学界和工业界群体之间的共性和差异。

3 复杂分析—介绍

本次大会重点关注了基于超大数据库的复杂分析。大会指出，当应用于千兆（peta-scale）级别的数据集时，即使是相对简单的计算也会变得很复杂。大会的目标是，透过这些普通的统计数据和聚合数据，发现什么才是高级科学家和商业分析人员所需要的，以及这些需求是如何影响 XLDB 的结构和用途的。

很多关于业界的复杂分析实例，都是和发现和理解客户行为模式相关。这些模式，或者在某些情况下属于这些模式的异常，可以用于很多商业应用。比如，可以确定广告销售目标群体，预测波动，探测垃圾，寻找欺骗和分析社会网络等。和可控的实验相结合，复杂分析可以用来改进产品，因为，可以事先预测产品的变化给客户行为和销售收入带来的变化。在科学界，类似的任务包括分析天体光谱和位置，压碎高吞吐量的基因组和蛋白质数据，对来自成千上万个传感器的气象数据进行拼装，对海洋、地震和燃烧的计算仿真进行比较，对有冲突和迷惑性的实验数据进行筛选等等。并不是所有这些科学领域都使用数据库来存储原始和衍生数据，但是，他们都有大量的数据集，数据规模达到 TB 级别，甚至达到 PB 级别。有大量的技术可以用于处理这些数据，包括对原始数据的转换，以及可以应用到衍生属性的高级机器学习算法。

科学届和工业界的分析方法具有很大的重叠性。二者都使用统计技术、分类算法和时间序列分析。二者都很喜欢找出不符合模式的异常值，同时使用大量数据来找出模式。

下面两节内容，将重点阐述在基于大规模数据的复杂分析方面，科学界和工业界所面临的共同问题。第 4 节讨论了以分析人员的视角看数据库时的相关问题，包括数据表示方法和查询接口。第 5 节继续讨论如何在数据库内部处理数据。

4 复杂分析—数据表示

4.1 规模和代价

数据量不仅在急剧增长，而且增长的速度也在不断增加。工业界的数据量已经可以达到每年几十个 PB 的级别，每天都有几十 TB 的原始数据生成。科学界的数据集的大小，也达到了这个规模，CERN 计划在类似的日常数据生成速度下，每年存储 15PB 的数据。当然，这些都是最大的数据集，实际上，大部分与会者在工作场合使用的数据集的规模都在 0.1 到 10PB 的范围。

对于来自不同数据仓库厂商的数据仓库产品而言，已经有一些可扩展性的解决方案。但是，具备可扩展性的分析工具，目前还没有同步跟上。许多令人满意的分析工作，都是借助于类似 R 或 SAS 或者 EXCEL 等工具，如果要在大规模数据集上采用这些工具进行分析，就必须对数据进行采样，把数据缩小到可以放入内存的规模。目前的算法变得越来越复杂，分析工具厂商又无法提供可扩展性的解决方案，这就使得在执行复杂分析时，单位数据的成本不断上升。

在这种超大的数据规模下，维护复杂的、规范化的关系模式，是很难的，代价也太高。许多项目都采用了折中的方案，即把数据存储在文件或非结构化字符串中，在传统的 RDBMS 中，只存储元数据或者一些衍生数据。

大会指出，即使是免费软件也都不是真正免费的。所有的软件都需要维护和操作代价，这其中的人员开销就非常大。

4.2 复杂性

就像规模不断在扩展一样，分析的复杂性也在不断增加。

首先，数据的结构变得更加复杂，这就使得处理过程也变得更加复杂。目前观察到得一些现象是，用一些属性来捕捉一些无法重构的条件。为 Internet 查询存储中间临时结果，以及为宇宙探测存储一些图像，就是关于这个方面的实例。时间序列变得越来越重要，在时间序列中，时间之间的顺序和间隔是很重要的。许多科学数据和不断增加的工业界数据，都有空间属性，因为在空间和时间上的多点关联分析是很有必要的。不确定性数据，为查询增加了不确定范围和中间计算，这进一步增加复杂性。

其次，分析人员自身也正变得越来越复杂。不同系统之间的转换也是必须的。在许多情形下，大家熟知的衍生数据产品，比如关键特性指标，是由高度优化的过程生成的，可以支持采用简单的工具进行一些基本分析。但是，对于今天的分析人员而言，这是远远不够的，他们需要在更大的细节层面对数据进行分析，必须对原始数据和衍生数据以即席的方式进行集成，然后发现一些新的模式或者新的性能指标。

4.3 灵活性

对于任何复杂分析而言，事先往往都无法准确知道全部需求。科学家通常不知道他们到底需要什么。工业界的需求都处于快速变化之中。分析系统在构建时，必须具备足够的灵活性，从而可以处理未知的需求。XLDB 的设计必须要满足这些新的、前所未有的需求。

随着存储的数据变得越来越复杂，越来越结构化，数据的变化性也在不断增加。相应地，模式灵活性，尤其是容易地增加新的属性，是 XLDB 操作的一个非常重要的方面。这些新的属性可能是由原始数据中生成的，也可能是从衍生的新指标中得到的，或者也可能是从终

端用户的注解中得到的。

随着尖端技术的不断进步，采用这些尖端技术的系统的能力也在不断增强。计算机领域的一些新的进展，比如多核计算、RAM 的存储容量、闪存、快速网络、极大增加的磁盘 I/O 带宽等等，可以支持许多分析技术。在不同时期，系统必须能够处理这些新方法的指令。

4.4 数据模型

许多种数据模型都可以用来表示 XLDB 中的信息。可能的数据模型包括：关系表、对象、数据流、数组、图、网格、字符串（比如 DNA 或氨基酸序列）、非结构化文本和 XML。对于任何给定的数据集而言，上述这些数据模型没有一个是完美的，而且一个系统也不可能很好地支持所有类型的数据模型。关系模型已经流行几十年了，它是综合考虑下面两个方面以后的折中方案：（1）灵活性和表达能力方面；（2）限制操作的数量从而让优化器更好的工作。对象数据库和 MapReduce 系统，则侧重于其中一个方面，它们都支持极大的表达能力和灵活的操作，但是，没有进行优化。

许多与会者认为，以数组的形式表示数据，可能是在关系模型基础上往前迈出的一大步。优秀的数组所表现出的性能，要比在关系表上模拟出数组的性能好得多。对于许多科学领域而言，数组是一个很自然的、直观的数据模型，这些科学领域包括天文学、物理学、遥感、海洋学和地震学。通常需要很少量的维度，一般不超过 4 个维度。有一点是非常重要的，那就是应该支持不规则数组，也就是在每个维度上具有不同的数组元素个数。

数组对于自身的数组元素具有内在的排序性。有些行业对这个特性很感兴趣，因为，他们可以使用数组元素的排序和间隔来表示事件序列。

一些科学领域需要更加专业的数据模型。生物学主要处理序列，化学一般处理图和网络结构，所有这些结构都可以采用表或数组进行模拟，但是，需要一定的代价。

4.5 接口

数据表示的最后一个方面就是接口，通过这些接口，分析人员就可以访问数据。就像上面提到的那样，终端用户的统计工具（比如 R、MATLAB 和 EXCEL），现在只能应用于一些小规模的数据集上面，因此，就需要把这些工具集成到数据库当中，这样就可以充分利用数据库的能力，同时也可以充分利用 XLDB 后端的可扩展性。

除了工具接口以外，科学界和工业界也会使用编程语言来实现一些高级的分析，比如采用 C++、IDL 甚至 FORTRAN。这些编程语言的接口要比更低的级别的 JDBC 或 ODBC 更加自然，可以在很大程度上提高分析人员的工作效率。Microsoft 的 LINQ 被作为一个例子提了出来。一种可能的替代方法是，设计一种新的编程语言，比如 Sawzall 或 Pig Latin，但是问题在于，有多少人愿意采用这些新的编程语言。

与会者认为，必须制定一些基本的操作语句对数组进行操作，而且这些操作语句将成为一个新的数据模型的基础。与会者预测，一些任务原语（比如执行时间序列分析）的集合，可以只需要几行代码就可以完成分析任务，而不像 SQL 代码那样，需要几页的代码。针对 AT&T 的 Daytona 系统的专业图处理操作就是一个例子。

此外，一个接口必须在数据库内部进行定义，从而支持定义新的操作。这个接口可以采用现有的编程语言实现，而且必须通过把计算放在数据存储位置来实现充分利用 XLDB 系统的并行性。一些技术在这个领域可能会比较有用，比如把 Daytona 的 Cymbal 编程语言转换成 C 代码进行编译和执行的技术。

5 复杂分析—处理

5.1 体系架构

对于最大规模的数据集而言，应该把计算放置在数据存储的地方，而不是把数据转移到计算发生的地方，这一点是毋庸置疑的。在微观层面，这需要一个非共享体系架构，这在现有的 XLDB 系统中很常见，比如那些采用 Teradata 或者 Pan-STARRS GrayWult 实现的系统。在宏观层面，难度很大，因为，当一个项目开始后，数据必须转移到项目投资的地方。无论在科学界还是在工业界，大家都已经发展出一种文化，那就是，一定要自己拥有数据和控制数据，有时候，这也是出于竞争的考虑。但是，拥有一个集中的统一的分析平台，将会带来很大的收益。它可以避免产生太多的数据集市，要知道，每个数据集市都是需要消耗管理资源的。在平台上以服务的形式提供分析功能，看起来是一种很有潜力的方式。这样就可以根据外部负载的变化，而动态调整平台的用途，实际上，一些外部因素，比如大会或季度报告，都会造成平台负载的变化。这样就可以实现宝贵资源的最大化利用，而且可以让这些宝贵的资源分配给最重要的问题，从而减少探索的时间。更深度的集中，并不排斥对数据具有严格的访问控制权。

在“蛮力阵营”和“数据库阵营”这二者之间，仍然存在一些分歧，蛮力阵营强调 MapReduce 框架和完全的表扫描。数据库阵营则强调优化能力。MapReduce 在容错方面具有优势，渐进方式给出的结果，可以让我们快速发现错误，也简化了编程模型，但是，它仍然需要进行编程，因此，更适合进行批处理，而不适合进行交互式查询，在资源利用率方面也不高。但是，两个阵营都认为，二者之间应该有更多的融合，彼此都能够采用对方模型。

5.2 再现性

随着分析变得越来越复杂，对一个分析过程及其结果能够再现，变得尤为重要。这个工作不仅仅只需要跟踪已经发生了什么，或者维护一些版本元数据。数据自身的版本演化以及使用旧版本数据的能力，是非常关键的，因为，在更新后的数据上面运行一个旧版本的过程，会生成完全不同的结果。这种能力也可以让我们对发生错误的数据进行源头追溯，或者从错误的源头数据开始，找出所有在这个错误数据基础上衍生出的其他数据。虽然，起源和再现性是和科学领域紧密相关的问题，但是，工业界也开始意识到这个特性的重要性，有时候是出于新的法律要求。

在数据库内部发生的操作，进行起源追溯是很容易做到的。但是，并非所有感兴趣的操作都在数据库中发生。在许多情形下，外部系统会被用来处理原始数据，或者作为一个“黑盒子”计算包。在这些情形下，必须对数据库内部和外部的数据起源信息进行合并。把外部的数据起源信息加载到数据库内部，是一种可行的方案，可以提供统一的查询能力。另外一种方法是，把数据库内部的数据起源信息通过一个外部工具导出成一种标准的格式。

在这个层面上维护数据起源信息，是需要代价的。与会者指出，XLDB 系统通常有大量的磁盘，从而可以支持足够的 I/O 带宽。在典型的应用中，只有 10%的磁盘处于使用状态，因此，剩余的磁盘就可以用来存储数据起源和版本信息。

与此同时，也有与会者认为，完美的再现性就是一个海市蜃楼，并非所有时间都需要再现性，尤其是在探索性的分析中。首先，在数据库外部的计算环境，比如硬件、操作系统、编译函数库等等，都会影响结果。虽然某个环境可以被记录，但是，重现一个给定的配置，也许需要很高的代价，或者根本就不可能实现。第二，如果放松准确性要求，那么，就可以降低很多代价。由于这些大规模系统中都会不可避免地发生硬件故障，具备在不完整的数据

上执行分析的能力，是非常关键的。类似地，通过放松对关系数据库中所采用的标准的 ACID 四性的要求，就可以提高响应速度和性能。在这些情形下，会使用不完整和不一致的数据，这个时候，系统应该明确给出提示，让用户知道结果有一定的不确定性。

5.3 工作流

有几种情形下，在复杂分析的过程中，必须需要工作流管理。最初，数据加载到系统中，数据处理程序会把原始数据转换成适合长期存储的数据格式。另一个数据处理过程，会把原始数据转换成更加适合分析的衍生数据。最后，分析工作自身还包括多个步骤。

对这些工作流进行跟踪，对于起源和再现性而言，是非常重要的，这一点已经在前面阐述过。以一种合适的方式为它们分配资源，也是很关键的。工业系统通常具有严格的要求，从而可以满足服务水平协议，但是，即使是科学系统，也必须要保证资源的公平分配。从维护起源和支持潜在优化（比如把计算放到数据存储的位置进行）的角度而言，在数据库内部管理工作流，是最强大的解决方案。但是，在许多情形下，把所有的处理都放入数据库内部是不可能的。因此，数据库必须能够和外部的工作流管理系统进行集成。

5.4 响应

尽管 XLDB 的超大规模，但是，许多终端用户仍然希望得到快速的响应。商业人员需要对事件进行实时响应，科学家也需要对一些在短暂观察窗口内存在的现象进行响应。在许多情形下，当原始数据被收集或加载到 XLDB 中进行处理时，就会触发这些响应，但是，最好能够把数据库的处理能力用来处理新鲜数据。

响应的另外一种方式是，系统具备显示渐进结果的能力，或者具备逐渐提高结果准确性的能力。这种能力在传统的 SQL RDBMS 中很少存在，具备这种能力以后，就可以快速发现错误的查询，或者在发现结果准确性已经达到某种程度以后，就可以快速停止查询。

5.5 管理

在超大规模情况下，如果管理的代价随着数据量的增加而同步增加，那么，这种系统是不可持续发展的。系统必须具备自我维护、自我恢复、自我负载均衡和自调整的能力，从而避免需要大量的数据库管理人员。在与会者的经验认识中，对大规模系统在每个层面进行细节监控，是非常重要的。把数据库的强大能力应用到对自身日志的分析中，也是很有用的。数据库性能指标和数据库配置之间的反馈环路，可以实现指标管理的自动化，减少数据扭曲，但是，这个反馈环路不应该完全排除人工操作，因为，短暂的加载异常都可能会带来长期的影响。

必须对数据库使用的资源进行精细的管理。准确的代价评估通常是很难进行的，因为存在许多相关的因素和数据依赖性。一个好的系统，应该可以快速捕获那些正在执行中的性能低下的查询，并且终止那些已经超出资源配额的查询。这就很自然要求实现类似操作系统中所采用的优先级机制。

当一个系统中包含了大量的计算机的时候，容错变得非常关键。一个需要关注的情形是，一个复杂的、长时间运行的用户自定义函数，运行失败了。虽然，数据库可以处理自己的内部操作发生失败的情形，但是，为用户代码提供失败处理能力（也许可以采用检查点机制），可能更加困难。

6 SciDB

第一届 XLDB 大会明确指出，缺乏一个可供大规模数据库用户使用的共享基础平台。每个数据密集型的群体，都直接在操作系统上面运行他们的软件，很少有例外情况。这就导致系统构建好以后，根本就无法应用到其他项目中。虽然数据密集型的工业界用户，一般都有足够的财力，可以购买这些定制的解决方案，但是，在科学界，这种方式是不可持续的。为了解决这个问题，第一届 XLDB 大会提出了下列建议：

（a） 加强科学界和数据库研究界的合作；

（b） 定义一个公共的数据库需求列表，可以被不同科学领域所共享；

后来，在 2008 年 3 月，在 Asilomar 举行的科学数据库小型研讨会上，决定构建一个新的开放的开源数据库，称为 SciDB。从那个时候开始，SciDB 的创始人已经确定的最初的合作伙伴，组建了一个数据库研究智囊团，收集了详细的用户用例，完成了最初的设计，募集了资金，成立了一个非盈利的组织，并且开始招募工作人员。本节内容重点介绍 SciDB 的初步设计和一些初期开展的活动。

6.1 科学需求

第一届 XLDB 大会，科学数据库小型研讨会，以及由初期的合作伙伴提供的用户用例等等，所有这些都透露出，科学数据库用户对目前的关系数据库存在普遍的不满意情绪。一些原因如下：

- **错误的数据模型**：科学数据和关系数据模型，很少能够实现自然的匹配。科学领域的理想数据模型，可能多少有些差别，但是，无论如何，都不可能仅仅是表的形式，而且在表上面模仿其他数据模型，效率非常低。
- **错误的操作**：最频繁执行的操作，比如回归分析和傅里叶变换，在关系数据库中是几乎无法实现的。类似的，被频繁执行的复杂分析，比如时间序列分析，也不可能用 SQL 语句进行表达。
- **没有起源**：所有的科学家都希望一个 DBMS 支持数据起源和再现性。
- **没有时间跨度**：科学领域的用户必须能够重现发表过的结果，因此，在当前的 DBMS 里面，会把原来的数据给覆盖掉，这是不行的。
- **缺乏可扩展性**：一些科学项目的数据量已经达到 PB 级别，其他项目也很快会达到这个级别。没有一个现有的 DBMS 可以支持几个 PB 级别的可扩展性。当需要考虑软件使用许可证的开销和硬件开销时，情况就变得更加糟糕。

6.2 设计

构建一个能够支持科学领域需要的所有特性的系统，需要大量的开发工作，即使是设计最低级别的系统，比如存储管理器。这就使得我们很难直接在现有的开源系统（DBMS 或 MapReduce）上面构建。虽然我们希望有些现有的授权软件可以得到重用，但是，SciDB 的全部设计必须从头开始。这个部分着重强调已经做出的关键决定、需要克服的技术难题和 SciDB 未来设计计划。

6.3 数据模型

很显然，没有任何一种数据模型可以满足所有科学领域的需求。在对一些因素做了评估以后，我们选择了数组模型，评估的因素包括：哪些数据模型可以满足大部分用户的需求，哪些数据模型是最容易实现的，哪些模型是可以在其他模型基础上实现的。数组模型对于大

部分科学用户而言是很合适的，包括天文学和许多地理科学的分支（海洋学、遥感、大气科学）。

SciDB 支持嵌套的多维数组。可以支持两种类型的数组：

（a） **基本数组**：属于 MATLAB 类型，具有整型维度，从 0 开始，每个维度可以是有边界或无边界的。

（b） **改进的数组**：用户自定义函数定义了数组的样子。改进的数组在不同维度上可以是不规则的，即不同维度所包含的数组元素的个数可以不相同。

数组的每个元素都会包含一系列属性，这个和关系数据库中的行类似。每个属性都可以是一个嵌套数组。我们希望基于数组的数据模型具有较好的表达能力。还采用了一些特殊的技术，来实现低损失的压缩。

6.4 查询语言和操作

SciDB 中，使用解析树的形式来定义查询语言，并将会构建与其他常用工具的绑定，比如 MATLAB、C++、Python、IDL 等。C++绑定，将会第一个得到实现，因为它对于内部开发很有用。希望社区能够完成其他绑定工作。

SciDB 将支持标准的关系操作，比如过滤或连接，再加上其他常用的数组操作。目前还没有确定这些操作的完整列表，还需要像不同的科学群体进行咨询。被频繁提及的实例包括回归分析和傅里叶转换。除了一些系统定义好的基本操作以外，用户也可以自己定义 PostgreSQL 类型的操作。

这个领域的一些研究主题包括，如何采用类似 C++或 Python 等语言表示数组操作。一方面，如果语法看起来和所有编程语言都很类似，那么，这个语言就会很有用；另一方面，一些工具语言已经很好地定义了数组操作。

6.5 基础架构

SciDB 将会运行在由工业标准硬件构成的、增量可扩展的簇（集群）或云上，可以是一个单个的笔记本电脑，或者是由私人项目或实验室管理的小型簇，或者是非常大的商业云。SciDB 建设完成后，应该具有很好的可用性、失效备援能力和灾难恢复能力。数据应该被分区到可用的硬件上，从而最大化吞吐量。

就地处理

SciDB 可以进行就地数据处理，也就是说，不需要把外部数据加载到 SciDB 中。这种数据无法使用某些 SciDB 服务，比如复制和灾难恢复。SciDB 会提供一种机制来描述这种类型数据的内容。此外，针对流行的数组格式（比如 HDF-5 和 NetCDF）的适配器也会被开发出来。

不确定性

根据来自不同科学家的输入情况来看，几乎所有科学领域都需要利用某种程度的不确定性。对不确定的需求，差别很大。因此，SciDB 在初期阶段只支持不确定性的定义，并且把它应用到简单的计算（比如比较）中。即使实现这个层次的功能，也不是一个简单的任务。更加复杂的错误模型，会在以后逐渐得到实现。不同科学领域的共性，会再次进行确认。

起源

为了捕捉到数据的起源（provenance），SciDB 必须记录下每个数据发生的每一次更新。此外，也可以通过特殊的接口把外部的起源导入，同时还会支持起源的导出。和数据的版本特性相结合，就有可能重新生成任何操作的条件，跟踪它的输入和效果。预期将会开发出专

用的工具来查询数据的起源。

6.6 项目组织和当前状态

SciDB 组织是以下面三个组织之间的合作为基础的：

- 科学领域和高端商业用户。他们的任务包括提供用户用例和对设计进行评估。
- 数据库研究智囊团。这个小组的任务包括设计系统、开展必要的研究工作、监督软件的开发。
- 一个非盈利的基金会。这个基金会的任务是管理这个开源项目，并且为开发出来的系统提供长期支持。

在初期，来自科学领域的合作伙伴包括 Large Synoptic Survey Telescope/Stanford Linear Accelerator Center (SLAC National Accelerator Laboratory) (LSST/SLAC), Pacific Northwest National Laboratory (PNNL), Lawrence Livermore National Laboratory (LLNL)和 University of California at Santa Barbara (UCSB)。初期的探索性用户是 LSST 和 eBay。其中的一些团队早已经提供了用户用例。已经成立了一个科学顾问组，来协调来自不同科学领域的输入，并把科学领域的术语翻译成数据库领域的术语，并对提出的要求进行优先级分类。

工业界的合作伙伴包括 eBay、Vertica 和 Microsoft。

智囊团的成员包括：Mike Stonebraker (MIT), David DeWitt (University of Wisconsin → Microsoft), Jignesh Patel (University of Wisconsin), Jennifer Widom (Stanford), Dave Maier (Portland State University), Stan Zdonik (Brown Institute), Sam Madden (MIT), Ugur Cetintemel (Brown University), Magda Balazinska (University of Washington) 和 Mike Carey (UC Irvine)。智囊团已经提出了最初的设计，目前正在重新定义和完善。

6.7 时间线

一个可供演示的 SciDB 系统将在 2009 年年底可以完成开发，到 2010 年年底，就可以完成可以投入使用的第一个系统。

6.8 总结

SciDB 项目激发了许多 XLDB 大会参会者的兴趣。大家普遍认为，它是一个很有雄心壮志的项目，因为，为了获得成功，在初期阶段必须把焦点放在实现定义良好的核心特性上面。这个系统是从零开始搭建的，因此，用户不要期望在两年内就可以获得 TB 级别的服务，最终获得好的服务性能，可能还需要一个很长的时间。

7 下一步工作

大会的最后一段时间用来讨论未来的工作。讨论了两个主题：（1）SciDB 和 XLDB 二者在未来的关系；（2）XLDB 大会的未来。

7.1 SciDB 和 XLDB

SciDB 是第一届 XLDB 大会的产物，很显然，在 XLDB 和 SciDB 之间具有很大的重合性。但是，也很显然，SciDB 的设计和开发不适合在 XLDB 这种大的环境中进行。因此，需要把 SciDB 独立出来。与会者认为，本次大会的大部分时间都被用来讨论 SciDB，是很

有价值的，可以让 XLDB 群体了解 SciDB 的最新进展。同时，与会者也认为，应该让 XLDB 群体定期了解 SciDB 项目的进展，但是，这个事情不需要占用未来 XLDB 大会的时间。

与会者建议，应该把 SciDB 收集到的用例发布给其他厂商来研究。与会者还建议，SciDB 应该涵盖更多科学领域。

和未来 SciDB 紧密相关的一个热门话题是，如何衡量这个项目是否成功。结论是，应该尽早制定一些衡量成功的标准，SciDB 的设计应该周期性地和这些标准进行比照。

7.2 科学难题

与会者认为，应该定义一个面向科学领域的数据库难题。这个难题和之前的一些测试基准（比如 TPC-X 系列）应该有所不同，因此，用难题(challenge)这个词比用测试基准(benchmark)这个词更好一些。一些与会者认为，一个基于 Sloan Digital Sky Survey 数据集和查询的难题，可以作为一个综合性测量指标，但是，最终大家认为，这个数据集和查询实际上已经是针对关系数据库引擎做了过多的优化。Mike Stonebraker 和 David DeWitt 同意定义一个科学难题。

7.3 维基百科

大会简单讨论了 XLDB 维基百科。与会者认为，XLDB 维基百科应该具有更好的可见性，得到更好的发布。可能需要具备一定杂志经验的主持人，来维护一个吸引人的网站。我们将会招聘一个主持人。

7.4 下一届大会

与会者普遍认为，我们应该继续举行 XLDB 大会。这个大会具有很好定义的、独特的主题，同时也为数据库用户和数据库社区成员（研究人员和厂商）提供了很好的交流平台。大会还指出，XLDB 和其他大会之间不存在重叠性，虽然其他群体（比如 SSDBM）已经尝试创建类似的论坛，但是，从来没有获得成功。

大多数与会者表示，XLDB 大会应该每年召开一次。会议日程安排为 2 天较为合适。我们将继续采用邀请参会的形式，来推动坦诚的沟通和交流。下一届 XLDB 大会仍然采用交互式讨论的方式，虽然也有人建议加入一些演示和论文来推动讨论。

有些与会者认为，或许下一届的 XLDB 大会，可以考虑和其他数据库会议（比如 SIGMOD 或 VLDB）一起举办。但是，最后大会讨论决定还是采用独立办会的方式。每个与会者都很清楚，VLDB 和 SIGMOD 会议上的较大的数据库群体，如果能够聆听真实的科学领域需求，将会是很有裨益的。因此，我们将会尝试在下一次 VLDB 大会上组织一个指导性质的讨论，但是，它不会取代 XLDB 大会。

与会者建议，我们应该把 XLDB 群体扩充到美国以外。大规模数据库的相关活动正在其他地区开展，比如欧洲（MonetDB）以及亚洲（尤其是中国和日本）。鉴于这个原因，把下一届大会放到欧洲和亚洲也是可以考虑的。许多与会者认为，最好的会议举办地是 CERN（欧洲核子研究委员会），日期可以定在与 VLDB 大会临近的前面或后面几天。2009 年 8 月，VLDB 大会将会在法国里昂举行。一旦 CERN 官方给予确认可以承办会议，那么，下一届 XLDB 大会的会址会很快确定。如果这个计划成行，那么，Europeans Maria Girone/CERN 和 Martin Kersten/CWI 将会负责大会组委会的工作，组委会成员还应该包括至少两个来自前两届 XLDB 大会的组织者。

与会者指出，在某几个主要科学领域，我们没有足够数量的参会代表。尤其是生物领域，没有代表参会，本次 XLDB 大会中的生物领域参会代表，其实都是和生物学家一起工作过的数据库研究人员。

下一届 XLDB 大会的两个重要目标是：

- 和非美国的 XLDB 群体联系；
- 和更多的科学领域和群体建立联系。

下一届 XLDB 大会应该包括一个关于 SciDB 项目的简短报告。另一个可能的讨论主题是“嵌入式传感器和 RFID”。大会的可能日程安排是，第一天讨论现有的引擎和解决方案，主要关注欧洲已经开发的系统，然后，在第二天开始讨论如何把这些解决方案推向前进。

未来的 XLDB 大会欢迎更多的 DMBS 厂商的参会代表。让数据库厂商始终了解大规模数据问题和需求，可以促进相关研究，最终可以导致开发出更多支持 XLDB 特性的 DBMS 产品。与会者也建议，让大家提前知道下一届 XLDB 大会的日程安排，或许对大家有帮助，比如，可以在 2008 年 12 月左右发布大会日常安排。

8 致谢

本次大会感谢以下赞助商：

- eBay
- Greenplum
- Facebook
- LSST Corporation.

9 术语

CERN – The European Organization for Nuclear Research
DBMS – Database Management Systems
HEP – High Energy Physics
LHC – Large Hadron Collider
LLNL - Lawrence Livermore National Laboratory
LSST – Large Synoptic Survey Telescope
MIT – Massachusetts Institute of Technology
Pan-STARRS – Panoramic Survey Telescope & Rapid Response System
RDBMS – Relational Database Management System
SDSS – Sloan Digital Sky Survey
SLAC – SLAC National Accelerator Laboratory, previously known as Stanford Linear Accelerator Center
SSDBM – Scientific and Statistical Database Management Conference
UCSB – University of California in Santa Barbara
VLDB – Very Large Databases
XLDB – Extremely Large Databases

10 附录：大会日程表（略）