第 5 届超大数据库会议（XLDB2011）大会报告（中文版）

REPORT FROM THE 5th WORKSHOP ON EXTREMELY LARGE DATABASES

Jacek Becla1*, Daniel Liwei Wang2, Kian-Tat Lim3
SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
*1 Email: becla@slac.stanford.edu
2 Email: danielw@slac.stanford.edu
3 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

1 大会总结

第 5 届 XLDB 大会（XLDB2011），主要关注医疗和基因学领域所面临的挑战，基于电子表格的大规模分析，以及大规模应用统计信息和机器学习所面临的挑战。

XLDB2011 明确了在医疗和基因学领域的相关问题。一些问题是比较普遍的，比如一些软件、数据格式和使用模型，在概念上都是相同的，可是无法兼容。使用习惯并没有明显的趋同性，因为用户通常拒绝接受变化。在这个数据极大丰富的世界，一些分析者还是采用处理数据稀少情形所采用的思维，虽然，已经有部分人开始意识到这个问题。新机器和新技术（DNA 序列和医疗图像）所产生的数据，正在迅速增长，这让分析人员猝不及防，但是，同时，这也让我们发现了具备高可扩展能力的工具的缺失，并让我们意识到需要更加强大、扩展性更好的数据管理。

在 XLDB 会议中，电子表格被放在大数据的背景下进行讨论，这也正好延续了上一届会议的讨论兴趣。就单个电子表格而言，通常都很小，但是，它非常普及，数量非常庞大，无处不在，这就让它成为了一个需要关注的大问题。电子表格具有很直观的接口，因此，它很难被其他产品所取代，即使它存在着数据质量的问题。电子表格更像是原始数据，没有质量保证机制，比如模式、数据类型、一致性和真实性，因此，很难对电子表格进行检索和维护。没有严格约束，增加了电子表格的易用性，也减少了在记录新概念时的冲突。因此，处理电子表格问题的解决方案，主要关注把电子表格访问接口提供给其他技术，这些技术对大规模数据集具有很好的适应性和可扩展能力，比如 Hadoop 和并行 RDBMS。

基于大规模数据的统计，仍然是一个有待解决的问题，虽然现在已经有一些方案。统计软件包本身不具备可扩展性，可是，可以在构建可扩展的代码之前用来对算法进行原型实验。一些与会者注意到，设计一个可以综合考量可用性和可扩展性的软件，并不可行；而其他与会者则认为很多扩展性问题都是可以解决的。由于计算代价过高，一些普通算法很难实现扩展，因此，就需要新的更加聪明的算法，或者是一些近似算法。统计分析人员和技术人员之间缺少沟通，也是一个大问题，有时候就会出现一些问题，比如某个问题可能已经存在解决

方案，可是统计分析人员就是看不到，而技术人员则认为统计分析人员在描述问题和需求时，没有表现出积极合作的态度。对于二者在未来的合作，很多与会者都持乐观态度。技术人员也开始积极努力从统计分析人员那里收集需求和问题描述，这样做，既能够帮助加强统计分析人员之间的合作，也能够帮助技术人员找到解决方案。

经典的机器学习算法，会从数据仓库、归档数据或托管的数据存储中抽取数据，然后把这些数据输入给特定的算法。对于大规模机器学习而言，有三类主要的算法。第一类方法是，把逻辑能力注入到数据库中，从而充分利用数据库的优化能力和可扩展性。不幸的是，并非所有的逻辑能力都可以被注入到数据库中，这个过程很可能会导致不一致性、混乱和难以维护。但是，数据库应该成为解决方案需要考虑的一部分，数据库本身所具备的数据清洗和准备以及数据质量控制等方面的能力，是非常重要的。第二种方法，则强调根据经验得到的启发算法，而尽量避免复杂的机器学习模型。这种方法认为，现有的机器学习方法对于今天和未来的问题都已经足够。最后一种方法在统计分析群体中比较常用，即构建小规模的原型系统，然后为一些特定的大规模场景构建定制代码。

XLDB 大会是一个规模较小的非正式会议，就像一场没有准备的即席讨论会。对免费软件的兴趣，正在迅速增长，但是，较大的组织和机构会遇到障碍，如果缺少商业软件的支持。服务计算架构很有吸引力，但是，价格太高，尤其是在不需要高可靠性和高可扩展性时更是如此。在讨论沟通的鸿沟时，我们发现在 SQL 和非 SQL 支持者之间的鸿沟越来越大，这里包括文化的差异和方法的差异。

对于 XLDB 而言，下一步工作就是扩展到医疗领域，讨论数据集成问题，和高性能计算群体建立更多的联系和合作。XLDB2012 将会在旧金山举行，同时，也可能举办另一场分会（satellite conference）。下一届 XLDB 还没有考虑论文的问题。

2 关于 XLDB 大会

XLDB 大会提供了一个讨论场所，讨论的主题主要围绕大规模数据问题，数据规模达到 TB、PB 甚至 EB 级别。第 5 届 XLDB 大会（XLDB2011）于 2011 年 10 月 20 日在 SLAC Menlo Park, CA 举行。大会的主要目标是：

（1）扩展到医疗和基因学群体，这些人以往是不参加 XLDB 大会的；

（2）在大数据分析的背景下，对统计和机器学习算法进行评估；

（3）讨论基于电子表格的分析。

XLDB2011 大会包含了两天的会议日程，共有 280 人参加会议。本大会报告只包含了讨论会的内容。获取大会的演示报告等信息可以访问网站：(http://www-conf.slac.stanford.edu/xldb2011/)。

2.1 参会

和往届 XLDB 大会一样，XLDB2011 大会仍然采用邀请参会的方式，这样可以把参会群体控制在尽量小的范围，同时具有很好的代表性。XLDB2011 参会者中，包括了科学和工业数据库用户、学术数据库研究群体、数据库开发商。业界的用户代表的数量正在逐年增加。

2.2 结构

XLDB2011 大会延续了 XLDB 的传统，仍然采用互动的讨论形式。首先是医疗行业和基因学群体的专题讨论。然后是关注基于数据表的大规模分析，以及大规模统计分析和机器学习算法。最后，在总结性的讨论中规划了下一届 XLDB 大会。

3. 新的群体：医疗和基因学

XLDB2011 大会引入了两个新领域的用户群体参加会议，即医疗和基因学。其中，两名代表来自国家卫生研究院，一名代表来自 GNS 医疗机构。与会者讨论了这些领域的数据管理和分析，包括当前的实践、最严峻的问题、寻找解决方案所面临的障碍，以及他们和比其更大的 XLDB 群体如何能够取得进展。

数据的分片和小规模的方法

基因学和医疗群体是非常分散的，对于如何生成和管理数据，许多小团体之间都没有达成共识。这从实用主义的观点出发，两个群体都认为计算是必须的开销。但是，他们对于标准化和统一化没有什么积极性。他们的数据生成设备和数据分析方法，各不相同。在语言、定义和方法上很少具有共同的地方，这使得合作变得很困难。例如，排序机器都具有不同的解决方案、文件格式和接口，有时候，即使是同一个 机器的不同版本，这些内容都不相同。由此生成的混乱的数据，很难用于其它作业，由此也导致了群体之间的隔阂。所幸的是，人们已经开始认识到数据碎片问题的严重性。

一种解决方案是，尽量减少自己开发，而直接采用供应商的现成软件，这可能会增加可互操作性。基因学群体很愿意采用价格不高的商业软件和开源软件。但是，现实情况是，商业软件价格不菲，开源软件根本找不到，还需要时间使其走向成熟。因此，这些群体还是继续自己开发相应的解决方案。之所以采用自己开发应用，还有一个很重要的原因就是，有些需求事先无法准确知道，当这些需求被完全确定的时候，呈现在你面前的就是一个定制的、半生不熟的解决方案。

医疗行业会频繁购买商业软件，比如分析软件，这导致了极大的开销，而且还会存在一些浪费。一些公司同时扮演着用户和供应商的角色，比如 GNS 医疗机构，就专注于构建和销售定义化的解决方案。业界用户非常重视对开源的商业支持。

编程语言方面也存在少量“分裂”的问题。这两个群体都使用 Java、R 和不同的脚本语言。SQL 虽然不是很普及，却也是一种可以被接受的语言。R 是一个统计包，在基因学领域很普及，被用到很多项目中，比如 Bioconductor 就是一个对高吞吐量的基因数据进行分析和理解的框架。R 已经被广泛接受，并得到了认可，但是，大家都知道它的可扩展性很差。这个群体已经习惯于在 R 的各种限制下进行工作，也知道需要具有更高可扩展性的工具，不过，就是不知道有什么更好的解决方案。

技术进步带来的问题

在未来的 1 到 1.5 年时间里，基因学群体需要解决迫切的、令人绝望的数据爆炸问题。数据爆炸本身是由技术进步带来的。更好的解决方案和更高的设备性能（这些设备现在要比以前便宜几个数量级），使得数据的增长速度超出了摩尔定律。在本次大会召开的时候，美国国家卫生研究院，每年可以产生 1PB 的数据。

主要的问题还在于文化和人，而不是技术。生物学群体很慢才接受把计算作为研究的一个重要部分。生物学家还不习惯于把计算和分析开销也列入预算。在以前，基因组测序(sequencing)是很昂贵的，它的数据很稀少，这意味着存储和分析数据的开销几乎可以忽略不计。但是，现在的条件发生了极大的变化，美国国家人类基因租研究院报告显示，在 2007 年对一个人类基因租进行测序需要耗费 10M 美元，但是，到了 2011 年，就只需要花费 10K 美元。对于许多生物学家而言，硬件基础设施并没有及时更新换代，半数以上的人还要被迫使用无法满足需求的、扩展性差的、固定的硬件设施。

关于人的问题，还有一个方面，那就是这两个群体都缺少能够有效处理大数据难题的人。

有一个与会者认为，医疗领域应该需要 140 万能够处理大数据的数据科学家，但是，目前这个人数只有 20 万。缺少对现有工具和技术的使用技巧，也是一个问题。许多对于 XLDB 群体而言是一目了然的事情，对于生物学家而言却是一无所知。可能我们需要新的技术，缺乏对现有计算环境的认识，缺乏对现有计算环境的经验知识，是一个更加迫切的问题。许多群体都不具备编写定制代码和集成现有软件的能力。而在那些具备这方面能力的人，他们所掌握的知识可能都很粗浅，对于更深层次的系统架构和软件组成原理可能一无所知。有时候，只具备一些浅薄的知识是有害的。一个与会者嘲弄道：“一丁点知识，是一件危险的事情”。那些在计算方面投入了大量工作（比如设计数据库中的数据分布）的生物学家，通常被群体所蔑视，因为群体内的很多人会认为，结构化数据不是科学，编写代码不是生物学。程序开发者是二等公民。医院似乎对 IT 工作人员不那么心存感激，一个与会者讲述了一些很有天赋的 IT 工作人员离开了医院，因为，他们的工作没有得到认可和赞赏。一个商业软件销售商则认为，数据分析工具已经过分地民主化，更好的商业化的选择被忽略了。如何开药，主要还是一个口口相传的工作，而缺少计算方法。

过去的 XLDB 大会没有提到过的、一个新的有趣的方法是，从数据中确定因果关系。医疗领域用户尤其需要分析数据因果关系，也就是说，哪些数据会影响到其他数据，以及以什么样的方式影响其他数据。他们也指出，一些已知的因果关系在实践中是不成立的。

未来的方向

与会者都乐观地认为，文化方面的问题是可以得到解决的。软件工程师和生物学家之间，应该进一步加强合作。需要在科学家、计算科学、学术界和工业界之间建立起沟通合作的桥梁。医院和解决方案供应商之间需要更多的合作。与会者提出一种新想法，即把软件工程师融入到项目过程，从而作为改变文化的一种方式。虽然有一个与会者对于生物学家是否会接受这种做法持有怀疑态度，但是，另外一个与会者还是引述了一个荷兰科学基金的实例，这个基金需要计算机科学专业的学生直接在一个科学问题上开展工作，并作为获得学位的必要条件。更多的合作将会减少在错误事情上的投资。例如，IBM 为一个基因公司投入了大量资金，但是，过多关注商业购买者，而不是科学家，最终的结果是，构建了模型，演示了令人印象深刻的计算能力，但是，并没有取得科学上的进步。销售商代表建议，可以在制度的层面开展合作，从而不会让生物学家先入为主地觉得计算专家都是做系统和数据库管理方面的事情。

为了解决这个群体缺少专家知识的问题，可以采用服务的方式提供解决方案，而不是采用软件和硬件的方式（这需要客户进行更多的集成工作）。通过这种方式，这个群体就可以把它的计算需求外包给专家来做，这就减少了自己开发的必要性。但是，目前还不清楚，这种方式是否能够满足他们的数据分析需求。另外一种减少软件集成的方式是，发布一个针对科学问题的公共软件堆栈，就像网络公司采用 LAMP 堆栈进行标准化一样。LAMP 就是 Linux-Apache-Mysql-Perl/PHP/Python，是构建应用服务器的基础。在 Indiana 大学里，各个院系之间就是使用这种策略提高公共程度。

与会者认为，XLDB 大会可以催生针对上述问题的解决方案。大会得到了很多支助，用来推动群体间的沟通合作，以及支持针对大数据问题的解决方案。XLDB 已经构建了一个人际网络，它可以收集用户用例，建立一个精选的用例集合，在不同领域之间找到共同的需求，为现有的需求找到与之匹配的解决方案，生成建议，把它们放入堆栈，这个堆栈可以被很多领域的人使用。XLDB 已经对一些地方的技术使用产生了影响，并且具有很大的潜力可以继续影响其他方面的文化。

4 从电子表格到大规模分析

虽然单独的电子表格不是“大数据”，但是，汇总在一起，它们却包含了海量的数据。之前的 XLDB 与会者，一直在寻求帮助，从而使他们更好的管理这些数据。在存储关键的数据碎片方面，电子表格得到了大量的使用，这一点是无论如何不能被忽视的。这个小节的内容，我们来明确基于电子表格的大规模数据管理和分析问题。

虽然，还不知道电子表格的使用情况和数据量情况，一个与会者估计，90%以上的商业数据被存储在电子表格中。每个计算机用户几乎都会使用电子表格，它们的广泛性是毋庸置疑的。它们可能经常被用于涉及几十亿美元的商业决策，常见的应用包括：

- 领域数据的权威存储；
- 计算和呈现汇总数据的工具；
- 数据可视化的简单方式；
- 多个数据源的数据集成工作的执行场所；
- 作为数据条目的输入表单；
- 便签本；
- 原型分析技术的沙箱；

一个与会者认为，人类以表格中的行的方式进行思考。因此，电子表格接口模型永远不会被替换，尽管电子表格的使用方式和许多电子表格软件的操作方法方面，还存在不少问题。电子表格的表格式接口，在编辑、可视化和操作数据方面，功能是很强大的。强大的功能一般都是默认包含在电子表格软件中，更加专业化的定制功能，比如文本分析和使用 R 进行统计数据处理等，都可以很好地集成进来。注册过的 Microsoft Excel 用户估计有 500 万，他们中的许多人都不是数据专家，没有注册的用户也大概有这个数量。一些数据集，比如美国人口统计局发布的统计摘要，就是以电子表格的形式发布的。这些简单的行和列的结构，需要很少的限制，当用户输入、编辑、操作数据时，不会产生冲突。电子表格为探查数据、开发算法和构建模型等工作提供了坚实的基础。分析人员在电子表格中开发的算法，通常会被程序开发员完全重写，转换成真正的、可重复执行的流水线，从而支持在大规模数据集上执行。大数据集不会被存储在电子表格中，因为，电子表格统一的输入、编辑和可视化的表格式接口，在超过一定数量的行和列以后，就会变得非常笨拙。

不幸的是，当接口是松散灵活的时候，数据也是如此。数据类型、数据单位和其他语义信息，都没有直接存在在电子表格中，所以，都是根据用户对每个新的公式和图表的规范来理解数据。电子表格没有像传统的数据库那样定义和强行施加模式。电子表格数据会被频繁地拷贝，有时候还会进行转换和调整，有时候则是作为和其他软件进行集成的组件，有时候用于共享。随着拷贝的数量的增加，会很难确定哪个是权威的、经典的版本。数据库所具备的能力，比如数据起源、安全性和可再生性，对于电子表格而言，是很困难的。所有上述问题，都使得对存储在电子表格中的数据进行管理，是一件代价高昂的事情，甚至是不可能的事情。

与会者认为，解决这些问题，应该保留电子表格的接口，但是，计算和存储功能应该从桌面上移除。一种方法是，把来自现有的数据库、数据仓库和 Hadoop 簇中的数据，集成到一个电子表格接口，Datameer Analytics Solution 就是这种思路。另一钟方法是，实现可扩展的、基于云计算的电子表格系统，比如 Google Fusion Table 就是这种思路，它可以支持把电子表格数据用于搜索、可视化和合作。没有一种方法具有绝对优势，未来还会有更多的解决方案。这些方法都无法解决已经存储在电子表格中的数据所面临的问题，虽然还没有可用的解决方案，一个来自密歇根大学的与会者演示了一个电子表格搜索引擎，它可以从一个大的

电子表格集合中推断出语义信息，然后构建索引，回答文本查询。

5 大规模统计分析

与会者认为，执行大规模统计分析所面临的常见问题，仍然没有得到解决，尽管已经有相关报道，说 SAS 和其他产品供应商已经提供了针对特定行业的解决方案。就像其他的计算型应用在面对日益增长的数据量时已经显得力不从心一样，类似 MATLAB、R 和 SAS 等统计分析软件在面对大规模数据库时也是如此。现在提出的初步想法是，首先使用这些工具在一个样品数据集上开发相关的统计分析应用，然后把这些统计分析应用部署到具有可扩展性的平台上，比如 Hadoop。我们在开发一个软件时，通常也采用类似的做法，也就是说，我们首先使用一种速度较慢却简单灵活的编程语言来开发一个原型系统，然后，再用一种更快的、生产型的编程语言来开发产品。John Chambers，作为 S 统计编程语言的创立者，他很赞同这种方式，他认为，统计分析软件包的设计初衷，就是为了让统计分析人员把精力集中在问题上，而不是类似可扩展性和高效性这种细节技术问题上，因为，这些技术问题都会在重新设计产品的过程中得到解决。因此，统计分析软件包，比如电子表格软件包，应该被看成一种原型系统来使用，而与此相关的其他繁重工作应该在其他地方得到解决。这种观点正好也符合了前几届 XLDB 大会的观点，即没有一种解决方案是完美的。

不幸的是，大规模数据环境下的计算，是一个很棘手的问题，与会者也在努力寻求一些方式，希望能够把扩展性好的计算平台融入到统计分析软件中。SAS 的用户自定义函数这个功能，允许把函数功能委派给许多数据库后端，包括可扩展的并行数据库，比如 Teradata。类似地，Revolution Analytics 公司也提供了针对 R 的计算外包插件，这个公司专门为 R 提供商业和产品开发，这个公司也提供了 R 的并行实现版本。但是，与会者很快意识到，这些公司都不够成熟，需要做更多的工作。把数据处理工作推给外部的可扩展的后端去做，虽然在某种程度上时成功的，但是，也有与会者认为，通过统计分析软件获得越来越多的计算资源，即使成功了，也不能算作是一个完美的解决方案。

一些与会者认为，一个更大的问题是，算法的计算具有超线性的时间代价，也就是说，时间代价和数据量之间是平方或者立方（甚至更高次方）的关系。虽然采用更多的计算资源可以加快算法的执行速度，但是，总的执行代价仍然是无法接受的。需要更多的工作来开发更加高效的方法，或者采用全新的实现技术，比如采用 Strassen 算法进行矩阵乘法运算，或者采用随机方法将少需要计算的数据量。在谈到性能较好的近似算法时，一个与会者注意到，网页搜索问题在本质上就是一个特征值问题，Google 的 PageRank 为最大特征值问题提供了高效的近似算法。

通常来说，解决大规模统计分析问题所面临的最主要的障碍就是，对现有的方法和知识的普及不够。统计分析人员可能还不知道早已经存在某个解决方案，R 软件库是很大的，但是，对于很多用户而言，通常会晕头转向找不到北。软件开发商对于统计分析人员所面临的问题，也没有足够的了解，这就很难开发出有针对性的解决方案。一些计算机领域的研究人员总是抱怨，存在问题的统计分析人员通常不愿意公布问题的细节，这一般都是出于安全的考虑，尤其是在医疗行业，会涉及到病人的隐私问题，或者出于竞争的考虑，因为，盈利机构和研究机构之间都存在竞争关系。一个微软公司的代表注意到，科学群体没有明确指出，现有的工具还缺少什么特性。一些与会者建议建立一个典型问题及其解决方案的资料库，并提供一些包含了足够细节的研究案例，通过这种方式，统计分析人员就可以从资料库中找到和自己类似的问题，或者相关的解决方案，开发者就可以针对详实的问题开发出相应的解决方案。最终，一个关于统计所面临挑战的测试基准(benchmark)或者形式化描述，将会极大推动研究的发展，比如 PennySort benchmark，同时也会推动利用现有系统解决悬而未决的

问题，进而让大家知道，产品供应商的技术到底要多快才能解决大规模统计分析所面临的挑战和难题。

6 机器学习

XLDB2011 对大规模数据量情况下的机器学习算法和相关问题进行了探讨。

一个有趣的想法是，机器学习算法可以完全在数据库内部执行，而不需要把数据从数据库中导出到统计分析软件包中。如果机器学习原语可以在数据库层得到实现，那么，处理过程就可以充分利用公共的数据库优化功能，比如并行性、缓存和改进的 I/O 调度。与会者描述了如何在一个关系数据库中实现一个数据挖掘模型，包括把模型规范映射到“模型-表的创建”，学习查询，对一个输入表和模型表之间的连接操作进行预测等等。但是，实现方案的结果也表明，数据库很难配合处理机器学习的各个阶段，尽管数据库可以用来加快其中一些步骤的处理过程。与会者认为，机器学习应该充分地在数据库中利用数据，从而避免一些其他操作，比如数据清洗、规范化以及把数据导入到严格的模式中所涉及的准备工作。
另一个与会者认为，当前这个数据极大丰富的世界，很适合那些根据经验训练得到的模型，而不适合那些高级的预测算法，他还认为，经过调优的启发算法可以获得更好的性能。他还引用了 LinkedIn 的“People you may know”作为一个实例，证明启发算法可以在大规模数据环境下获得好的性能。人们一般会有一种感觉，认为机器学习领域的学术研究已经足够可以用来解决一般的现实世界问题，但是，真实的应用是比较特殊和分散的，虽然机器学习已经在某些领域得到了较为成功的应用。有一个与会者认为，启发算法没有坚实的理论基础，它们的结果不能被应用到各种情形，但是，其他一些与会者则认为，启发算法已经做得很好，在医疗行业已经得到了广泛的应用。

如何表示机器学习模型被认为是一个具有挑战性的任务。PMML（Predictive Model Markup Language：预测模型标记语言）就是其中的一个规范，但是，目前还没有被广泛接受的标准。质量控制被认为是另一个难题，一个与会者希望能够对机器学习的每个处理步骤进行质量控制，从而有助于对结果的理解。

与会者还讨论了从数据中寻找答案的其他需要考虑的因素。多重假设陷阱（即如果存在足够的数据就可以支持任何假设），在数据极大丰富的环境下，就是一个非常危险的事情。另一个问题是，算法在数据集上运行时，经常会假设这些数据代表了完整且封闭的世界。这个假设是很危险的，因为数据再多，也都只是现实世界的一部分抽样。

在某个软件环境或使用某种语言开发出原型系统，然后采用其他软件或语言去重新实现这个系统，这种做法在机器学习中不断被反复采用。重新实现，是一个改进算法性能的机会，但是，同时也是一个引入错误和误解的机会。算法的创建者和重新实现者，通常都是来自不同领域的不同群体，比如统计分析人员或者计算机科学家，由于角度不同，他们之间很难进行有效的沟通。消除“原型系统-重新设计”这种模式，而采用统一的开发过程，对于这种做法，大多数与会者都不看好，但是，与会者都强调克服不同群体之间的沟通障碍。与会者建议合作团体中应该包括来自不同领域的人员。与会者再次建议建立具有用例的数据集合，这样不仅可以帮助计算机科学家构建更好、更合适的工具，同时，也可以给数据科学家和其他领域的专家提供可供参考的最佳实践方法。

XLDB 群体可以为那些来自不同群体的人员之间建立联系，这些人员可以代表各自的群体，并且对自己的领域和计算知识都有所了解，这样做，就可以实现可扩展的机器学习。另一个能够起到帮助作用的做法是，确定一个原型用例，比如在 EXPO 肿瘤学数据集上进行双聚类（bi-clustering）分析，或者对 google.org 上的 200TB 测序数据进行分析。

7 其他主题

XLDB 大会在热烈讨论过程中，经常会涉及一些意想不到的主题，其中一些主题如下。

与会者在讨论中，把自由软件看成一种解决方案策略和一种发展模式。来自较大的企业的代表注意到，此前大家都比较喜欢商业解决方案，但是，目前大家也开始逐渐接受自由软件，自由软件也不再像以前那样被认为是不成熟的。对于不是从事计算机的企业而言，必须有商业企业为其提供技术支持和承担责任，自由软件才可能有生存的空间。

服务计算是一种比较有潜力的、可以满足存储和计算扩展性需求的方法。虽然，当前的价格对于盈利的企业而言还是比较合理的，但是，还需要能够提供给学术群体使用的替代产品，学术群体对产品的可靠性和性能要求低一些，而对价格比较敏感。据相关报道称，Amazon Web Services 正在调研类似的替代产品。

在讨论分析模型时，与会者意识到，数据库计算模型对于很大一部分分析处理而言是有效的。在这种模型中，一个问题是以声明语句的形式提交给查询引擎，并在数据存储的地方就近执行查询。但是，他们也指出，一些重复性查询技术，比如那些通过一些重复步骤得到的结果，数据库就不能很好地给予支持，而需要它们自己的定制实现方案。

为查询提供概率答案，也就是给出的答案是一个概率质量函数或者概率密度函数，比较受到统计分析人员的青睐，他们期望通过这种方式避免对结果的错误理解。但是，并没有现成的产品可以使用，一个与会者认为，实现这种引擎，要比人们预想的困难得多。XLDB 大会上的一个海报，给出了一个解决方案的原型系统实现，它可以通过为每个可能的结果复制数据库实例，从而计算出概率结果。对于绝大多数与会者而言，使用 SQL 语言的群体和不使用 SQL 语言的群体之间的鸿沟，似乎难以跨越。对于 SQL 的敌意，主要来自 SQL 数据库软件所附带的东西（比如事务和模式），死板和不灵活性，以及无法嵌入 C 或其他编程语言。与会者感觉到，虽然转换和协调似乎是不可能的，但是，不同群体之间还是可以从沟通交流中学到很多东西。

在高性能计算仿真中的数据的使用被划分成两大类。第一类是，对仿真产生的大量数据流进行监控，从而可以在早期探测到错误。第二种是，在这种仿真完成以后，执行离线分析。

8 下一步工作

就像过去一样，大会的一部分时间用来确定对未来工作的规划。

下一届 XLDB 大会应该扩展到医疗群体。在医疗领域，还有很多大数据问题有待探索和发现。而今年的大会，则已经很好地涵盖了基因学和网页规模问题。下一届大会需要涵盖更多的问题。业界群体建议，可以把移动通信、生产厂家、飞行数据（比如波音）和国家情报机构（比如美国国防部高级研究计划局）等都包含进来。大会也邀请了硬件制造商 Intel 公司，对于参会的大数据群体所反映出的趋势，Intel 可以提供长期的计划和观点。与会者还要求提供一个专题讨论会，从而让风险投资者阐述观点。

最迫切需要的讨论主题是数据集成。数据集成问题是一个仍然没有得到解决的大难题，有一些人已经理解该问题，并且有一些人已经开始研究解决方案。另一个需要讨论的主题就是云计算，在数据集中的情形下，需要考虑代价问题。其他的主题还包括对“基于每个查询的模式”（这是对 Hadoop 中无模式计算的回应）的数据库支持和阵列数据库。

到目前为止，对于 XLDB 群体而言，大部分人要求开展的未来的活动包括：

（1） 收集测试用例（数据和相应的应用软件）；

（2） 收集用户用例和挑战；

（3） 为文档建立一个单个的存储库，并发布收集的信息（比如上面的测试用例和用户用

例）；

（4） 确定一个参考的体系架构，针对数据集中的解决方案，给出详细的硬件和软件配置。

与会者也对与高性能计算（HPC）群体进行合作表达了浓厚的兴趣。许多人都认为，HPC 和 XLDB 群体之间都有许多可以互相学习的东西。最好的建议是，让 XLDB 代表每年都去参加一个 HPC 集会，比如其中一个超级计算会议。XLDB 本身太小，无法容纳一个大的 HPC 代表团，而与会者认为，一个小的 HPC 代表团又让人觉得有些不舒服和疏远。

在大会日程上，下一届大会应该和 XLDB2011 类似，在多样性、专注性和组织性之间找到一个好的平衡点。与会者提出的可能的改进的地方包括，设置一个额外的专题讨论会和演示（或许可以放在接待宴会之前）。XLDB 群体强烈感觉到，引入同行论文评审，并非一个好的建议，因为这样做会在很大程度上降低报告的质量，并且会压抑开放的、无审查的对话活动。卫星大会（satellite workshop），比如在爱丁堡举行的 XLDB-Europe，也是非常有益的，但是，组织一个在异地（比如亚洲）举行的讨论会，可能会存在一些难度，因为会存在距离问题、语言问题和签证问题（尤其是如果在中国举办的话）。由于 XLDB 核心团队的财力有限，人力资源也很有限，因为，必须谨慎确定这个任务，从而使得未来的 XLDB 大会和活动能够产生最大的收益。

(全文完，厦门大学计算机系教师林子雨 2012 年 5 月翻译)