大数据 文献总结（2）

MIT-Plato 团队

中文文献：

【摘要】 大数据分析相比于传统的数据仓库应用,具有数据量大、查询分析复杂等特点.为了设计适合大数据分析的数据仓库架构,文中列举了大数据分析平台需要具备的几个重要特性,对当前的主流实现平台——并行数据库、MapReduce 及基于两者的混合架构进行了分析归纳,指出了各自的优势及不足,同时也对各个方向的研究现状及作者在大数据分析方面的努力进行了介绍,对未来研究做了展望.

[1]王珊 王会举 覃雄派 周烜**.**架构大数据**:**挑战、现状与展望**[J].** 计算机学报**.2011** 年 **10** 期

【摘要】大数据泛指巨量的数据集，因可从中挖掘出有价值的信息而受到重视。《华尔街日报》将大数据时代、智能化生产和无线网络革命称为引领未来繁荣的三大技术变革。麦肯锡公司的报告指出数据是一种生产资料，大数据是下一个创新、竞争、生产力提高的前沿。世界经济论坛的报告认定大数据为新财富，价值堪比石油。因此，发达国家纷纷将开发利用大数据作为夺取新一轮竞争制高点的重要抓手。

[2] 邬贺铨**.**大数据的机遇与挑战**[J].** 科技与教育**.2013** 年 **4** 期

【摘要】随着全球数据量爆炸式的增长，大数据时代已经到来。文章从大数据时代的基本特征入手，引出了大数据时代面临的挑战以及大数据带来的价值，提出了以创新为核心的务实应对策略。

[3] 陈如明**.** 大数据时代的挑战、价值与应对策略**[J].** 中国国际信息通信展专刊**.2012** 年

【摘要】服务有四大特性,其中不可储存性导致了服务总是存在供求上的不平衡,这种供求矛盾有两种情况:供过于求和供不应求。服务的需求出现了波峰和波谷。怎样解决这种矛盾一直是营销界重要的研究问题之一,并因此产生了需求管理的概念。目前很多服务企业花了很多精力来研究怎么将需求的波峰填补需求的波谷,但是,创造一种与消费者消费规律曲线相平行的灵活的服务生产动态变化机制,在今天以顾客为导向的营销界,似乎显得比改变消费者需求更有效和更理性。为此,本文创造性地提出服务参与层级理论,并将眼光放在当今世界比较热门的大数据概念上,利用层级理论和大数据概念提出一种需求管理的新思路。

[4] 沈晓雨**.** 大数据时代下的服务需求管理新思路**[J].** 商场现代化**.2013** 年 **20** 期

【摘要】 随着互联网应用的飞速发展和信息的社会化，数据呈爆发式的增长，传统的关系数据库在处理分析如此海量的数据时出现性能和可扩展性的瓶颈，所以必须研究新的有效的大数据分析平台。大数据技术目前还没成熟，也没形成统一标准，但工业界已经广泛使用 Hadoop 作为其大数据处理平台，这也带动了国内学术界对 Hadoop 相关技术研究。除了 Hadoop 外，NoSQL 相关技术也得到较快发展，涌现了一批优秀的开源项目，如 HBase 和 Cassandra 等都被工业界广泛应用。本文基于国家核高基科技重大专项——非结构化数据管理系统 LaUDMS 来研究和实现对大数据的处理分析相关技术。非结构化数据管理系统 LaUDMS 重点就是深入研究大数据的存储和分析技术，并结合理论和实践来解决对大规模非结构化数据的管理难题。本文首先对大数据处理分析平台的研究现状进行了综述；其次在综合比较分析现有平台优缺点的基础上介绍了非结构化数据管理系统 LaUDMS 的内核清华知云 Kloud 的平台架构；再次是清华知云 Kloud 中的大数据分析平台的技术研究和实现。技术研究包括深入分析了分布式数据仓库 Hive 的设计和组件，并将其融合到基于 P2P 架构的 Cassandra 内部实现中；为实现 Hive 组件完全融合到 Cassandra 中，定义了基于 Cassandra 自由表的面向对象数据模型来存取 Hive 的元数据信息；为提高自由表访问效率，描述了基于 Cassandra 自由表的辅助索引设计和实现，并且将其融合到 Hive 的分布式索引插件框架中，实现 Hive 分析的性能优化。该大数据分析平台实现后对某网站用户访问日志进行了实验分

析，性能和可用性得到相应的提升，取得良好效果。
[5] 卓安**.** 基于 **P2P** 可伸缩架构的大数据分析平台研究与实现**[M].** 清华大学**.2012**
【摘要】 云计算、物联网、社交网络等新兴服务促使人类社会的数据种类和规模正以前所未有的速度增长,大数据时代正式到来.数据从简单的处理对象开始转变为一种基础性资源,如何更好地管理和利用大数据已经成为普遍关注的话题.大数据的规模效应给数据存储、管理以及数据分析带来了极大的挑战,数据管理方式上的变革正在酝酿和发生.对大数据的基本概念进行剖析,并对大数据的主要应用作简单对比.在此基础上,阐述大数据处理的基本框架,并就云计算技术对于大数据时代数据管理所产生的作用进行分析.最后归纳总结大数据时代所面临的新挑战.
[5] 孟小峰 慈祥**.** 大数据管理**:**概念、技术与挑战**[J].** 计算机研究与发展**.2013** 年 **01** 期
【摘要】 在科学研究、计算机仿真、互联网应用、电子商务等诸多应用领域,数据量正在以极快的速度增长,为了分析和利用这些庞大的数据资源,必须依赖有效的数据分析技术.传统的关系数据管理技术(并行数据库)经过了将近 40 年的发展,在扩展性方面遇到了巨大的障碍,无法胜任大数据分析的任务;而以 MapReduce 为代表的非关系数据管理和分析技术异军突起,以其良好的扩展性、容错性和大规模并行处理的优势,从互联网信息搜索领域开始,进而在数据分析的诸多领域和关系数据管理技术展开了竞争.关系数据管理技术阵营在丧失搜索这个阵地之后,开始考虑自身的局限性,不断借鉴 MapReduce 的优秀思想改造自身,而以 MapReduce 为代表的非关系数据管理技术阵营,从关系数据管理技术所积累的宝贵财富中挖掘可以借鉴的技术和方法,不断解决其性能问题.面向大数据的深度分析需求,新的架构模式正在涌现.关系数据管理技术和非关系数据管理技术在不断的竞争中互相取长补短,在新的大数据分析生态系统内找到自己的位置.
[6] 覃雄派 王会举 杜小勇 王珊**.**大数据分析**——RDBMS** 与 **MapReduce** 的竞争与共生**[J].** 软件学报**. 2012** 年 **01** 期
【摘要】 自从计算机出现后,计算机对数据的管理经历了程序管理、文件管理和数据库管理三个阶段。数据库是数据管理的高级阶段,是数据管理最有效的手段,是现代计算机信息系统和计算机应用系统的基础和核心。本文介绍了数据库的定义、发展历史及各代数据库所采用的数据模型、各代数据库的优缺点。结合当今应用需求和新技术对数据库发展趋势、应用前景作了展望。
[7]曹文平**,**闫金梅**.** 数据库综述**[J].** 科技管理研究**. 2006(09)**
【摘要】 重点介绍了数据库技术的发展与主流技术。
[8] 宋淑玲**,**丁蕊**.** 数据库技术的发展与主流技术**[J].** 黑龙江科技信息**. 2009(01)**
【摘要】 随着计算机应用领域的扩张,以及 Internet 技术的广泛运用,对做为信息管理的重要技术支柱——数据库的专门研究也出现了新技术。本文从数据挖掘和数据仓库、XML 技术、数据流管理和网格数据管理等几个方面讨论目前数据库研究领域中最热门技术的发展现状和研究方向。
[9] 魏萌**.** 浅谈数据库技术的研究与发展**[J].** 科协论坛**(**下半月**). 2007(04)**
【摘要】 数据挖掘可视化是数据挖掘中的一个重要组成部分和必然的发展趋势。本文对数据挖掘可视化技术的研究现状及发展趋势进行了分析与概括,包括数据挖掘可视化的含义、内容、应用现状以及发展趋势等。
[10] 王华金**,**蔡虔**.** 数据挖掘可视化技术综述**[J].** 科技广场**. 2009(01)**
【摘要】 数据挖掘作为当前国际学术界的一个研究热点,本文对它的研究意义、定义、分类等概念做了深入的阐述,详细介绍了数据挖掘的全过程,为深入研究数据挖掘的应用奠定了基础。
[11] 潘春花**.** 数据挖掘理论及挖掘过程浅析**[J].** 科技信息**. 2009(04)**

【摘要】 数据库的应用已十分广泛，深入到了各个领域，但同时也带来了数据的安全隐患。本文从应用角度出发，介绍 SQL Server 2008 的安全及加密设置，并提出 SQL Server 数据库应用时的安全措施。
[12] 孟宪颖,毛应爽,赵慧玲． **SQL Server** 数据库安全性研究**[J].** 计算机光盘软件与应用**.**
【摘要】 近来,大数据引起了产业界、科技界和政府部门的高度关注。本文简要阐述了大数据的研究现状与重大意义,探讨了大数据的科学问题,介绍了大数据应用与研究所面临的问题与挑战。最后,对大数据发展战略提出了几点建议。
[13] 李国杰 程学旗**.** 大数据研究**:**未来科技及经济社会发展的重大战略领域**——**大数据的研究现状与科学思考**[J].** 中国科学院院刊**. 2012** 年 **06** 期
【摘要】 <正>大数据:正在到来的数据革命,以及它如何改变政府、商业与我们的生活,涂子沛,广西师范大学出版社,2012 当说到大数据,你想到什么?媒体已不止一次报道过,而且都算不上什么"新闻"了。譬如,你爱看书,经常上网淘,一来二去,基于购买行为,网站就了解到了你的偏好,除了在你浏览时..
[14] 杨吉． 涂子沛的“大数据”**:**不是你想的那样**[J].** 中国图书评论**. 2012(10)**
【摘要】 <正>IDC 认为,中国在大数据领域具有巨大的市场潜力,将成为全球最重要的大数据市场之一随着云计算概念日渐深入人心,大数据也越来越受到关注。IDC 在长期对云计算市场进行跟踪研究的同时,也对大数据市场保持着密切关注。IDC 将大数据技术定义为:"为了更为经济的从高频率获取的、大容量的、不同结构和类型的数据中获取价值,而设计的新一代架构和技术。"IDC 发现,目前大..
[15] 周震刚． 中国大数据市场 **10** 大预测**[J].** 通讯世界**. 2012(10)**
【摘要】 <正>关于大数据的发展背景、重大意义、最新动向、未来趋势以及中国的机遇与挑战等相关问题,中国工程院院士、中科院计算所首席科学家李国杰接受了《新经济导刊》专访。李国杰表示,大数据对经济社会发展和科学研究具有革命性的意义,其兴起有着内在的需求和利益驱动,因为数据里蕴藏着巨大的价值。未来将形成数据服务、数据探矿、数据化学、数据材料、数..
[16] 牛禄青． 构建大数据产业环境 专访中国工程院院士、中科院计算所首席科学家李国杰**[J].** 新经济导刊**. 2012(12)**
【摘要】 <正>尽管大数据时代的风暴早已席卷全球,EMC、微软、甲骨文、IBM 等巨头早已在风暴中手擎闪电,争当大数据时代下新世界的宙斯主宰,但是大数据和云计算的双驾马车却以神速疾驰,在"诸神之战"中为大数据未来留下一片澄澈的蓝海。而大数据和云计算之所以同车同轨、交蔓相生近乎唇齿,正因为它们对彼此的需求正盛,而数不尽的分析报告竞相表明,大数据和云计算将在长到难以预知的一段时期内保持这种紧密的关系。
[17] 何鹏**.**移动互联网时代的企业信息聚合**——**从大数据的实践到云计算的应用**[J].** 互联网天地**. 2012(12)**
【摘要】 数据将成为运营商开展移动互联网业务的核心优势资产,本文分析了运营商的数据获取原则和数据获取策略,以及开展大数据应用的关键保障。
[18] 顾芳,刘旭峰,左超． 大数据背景下运营商移动互联网发展策略研究**[J].** 邮电设计技术**. 2012(08)**
【摘要】 在大数据时代,政府、贸易、金融保险、信息技术等诸多行业都将在大数据技术中获得极大的价值提升。本文通过对大数据的跟踪和研究,解释快速发展变化的新技术,洞察信息产业发展规律,发现在其影响下金融市场的变化,以期做出更好的投资决策和判断。
[19] 韦雪琼,杨晔,史超． 大数据发展下的金融市场新生态**[J].** 时代金融**. 2012(21)**
【摘要】 伴随 Internet 和 Web 技术的飞速发展,语音、视频、网络日志、互联网搜索索引、互联网文本文件等技术的广泛使用带来了数据量的急剧增长,这预示着大数据时代的到来。

大数据时代的数据具有数据量剧增、数据结构更复杂化的特点,导致数据存储和处理的难度加大。而Hadoop的出现大大简化了大数据时代数据的存储和处理的问题,所以本文对Hadoop技术的研究和优化具有重要的现实意义。本文研究的主要内容是：首先对 Hadoop 的核心技术 HDFS 和 MapReduce 的原理进行了研究和分析。分别从名字节点、数据节点、接口、类、调用关系等方面进行详细的研究,并分析了 HDFS 和 MapReduce 的工作机制。同时,针对 Hadoop 目前存在的两个性能问题,在深入研究源码的基础上,提出初步改进方案。其次,对第一个性能问题 Hadoop 推测执行算法在异构环境中性能较差的问题进行研究和分析,提出改进的算法,该算法根据系统负载的情况自动的调节后备任务的执行,以实现系统负载的均衡；采用 Zaharia 提出的历史平均剩余完成时间估算剩余时间,并将剩余时间的值大于 0.2 的方法判断掉队者,进而得到更精确的掉队者队列。新算法在一定程度上提高了异构环境中推测执行的性能。最后,对第二个性能问题 DBInputFormat 操作关系数据库中的海量数据时所出现的性能缺陷问题进行深入的分析和研究,并对 DBInputFormat 接口进行改进,提出新的分片策略,构建新接口。该接口在一定程度上提高了 Hadoop 操作关系数据库的效率,改善了 Hadoop 读取关系数据库的性能。搭建实验平台,分别对新提出的算法和改进的接口进行实验,经过验证,说明它们在一定程度上都提高了 Hadoop 性能。

[20] 曹英**.**大数据环境下 **Hadoop** 性能优化的研究**[D].** 大连海事大学 **2013**

英文文献：

[1] Das S, Sismanis Y, Beyer K S, Gemulla R, Haas P J,McPherson J. Ricardo: Integrating R and Hadoop//Proceed-ings of the ACM SIGMOD International Conference on Management of Data(SIGMOD‘ 10). Athens, Greece, Ind-ianapolis, Indiana, USA, 2010

[2] Karayannidis N, Tsois A, Sellis T K, Pieringer R, MarklV, RamsakF, FenkR, Elhardt K, Bayer R. Processing starqueries onhierarchically-clusteredfact tables//Proceedingsofthe 28th International Conference on Very Large Data Bases(VLDB‘ 02). Hong Kong, China, 2002

[3] Bajda-PawlikowskiKamil, Abadi Daniel J, SilberschatzAvi,Paulson Erik. Efficient processing of data warehousing que-ries in a split execution environment//Proceedings of theACM SIGMOD International Conference on Management ofData (SIGMOD‘11). Athens, Greece, 2011: 985-996.2011

[4] Wang Huijui, Qin Xiongpai, Zhang Yansong, Wang Shan,Wang Zhanwei. LinearDB: A relational approach to makedata warehouse scale like MapReduce//Proceedings of theDatabase Systems for Advanced Applications-16th Interna-tional Conference (DASFAA‘11). Hong Kong, China,2011

[5] Okcan A,RiedewaldM. Processing theta-joins using MapReduce //Proceedings of the ACM SIGMOD International Confer-ence on Management of Data (SIGMOD‘ 11). Athens,Greece, 2011

[6] Jiang D, Tung A K H, Chen G. Map-join-reduce: Towardsscalable and efficient data analysis on large clusters. TKDE,2010

[7] Lin Y, Agrawal D, Chen C, OoiBC, Wu S. Llama: Lever-aging columnar storage for scalable join processing in theMapReduceframework//Proceedings of the ACM SIGMOD In-ternationalConferenceon Management ofData(SIGMOD‘ 11).Athens, Greece, 2011

[8] AfratiF N, Ullman J D. Optimizing joins in a map-reduceenvironment//Proceedings of the 13th International Confer-ence on Extending Database Technology. Lausanne, Switzer-land, 2010

[9] Yang H-C, Dasdan A, Hsiao R-L, ParkerD S. Map-reduce-merge: Simplified relational data processing on large clus-ters//Proceedings of the ACM SIGMOD International Con-ference on Management of Data (SIGMOD’ 07). Beijing,China, 2007

[10] Blanas S, Patel Jignesh, Ercegovac V, Rao J, Shekita E J,Tian Y. A comparison of join algorithmsforlog processing inMaPreduce//Proceedings of theACM SIGMODInternationalConference on Management of Data(SIGMOD’ 10). Indian-apolis, Indiana, USA, 2010

[11] LiBoduo, MazurEdward, Diao Yanlei, McGregor Andrew,Shenoy Prashant J. A platformforscalable one-passanalyticsusing MapReduce//Proceedings of the ACM SIGMOD Inter-nationalConference on Management of Data(SIGMOD’ 11).Athens, Greece, 2011

[12] Nykiel T, Potamias M, Mishra C, Kollios G, Koudas N.MRShare: Sharing across multiple queries in MapReduce.PVLDB, 2010

[13] Condie T, Conway N, Alvaro P, HellersteinJM, Elmeleegy K, Sears R. MapReduce online//Proceedings of the 7thUSENIX Symposium on NetworkedSystems Design and Im-plementation(NSDI‘ 10). San Jose, California, 2010

[14] Jiang D, OoiBC, ShiL, WuS. The performanceof MapRe-duce: An in-depthstudy. PVLDB, 2010

[13] Stonebraker M, AbadiD J, DeWitt D J, MaddenS, PaulsonE, Pavlo A, Rasin A. MapReduce and parallel DBMSs:Friendsor foes? Communications of the ACM, 2010

[15] DeanJ, Ghemawat S. MapReduce: A flexible data process-ing tool. Communications of the ACM, 2010

[16] Olston C, Reed B, Srivastava U, Kumar R, Tomkins An-drew. Pig latin: A not-so-foreign

language for data process-ing//Proceedings of theACM SIGMODInternationalConfer-ence on Management of Data (SIGMOD‘08). Vancouver,BC, Canada, 2008
[17] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel J Abadi,Alexander Rasin, Avi Silberschatz. HadoopDB: An architec-tural hybrid of MapReduce and DBMS technologies for ana-lyticalworkloads//Proceedingsof the35th International Con-ference on Very Large Data Bases (VLDB‘09). Lyon,France, 2009
[18] http: //www. vertica. com/the-analyticb-iet-land-hadoop-mapreduce-integration/
[19] Upadhyaya P, Kwon Y C, Balazinska M. A latency andfault-tolerance optimizerforonline parallel query plans//Pro-ceedings of the ACM SIGMOD International Conference onManagement of Data(SIGMOD‘ 11). Athens, Greece, 2011
[20] Yang C, Yen C, Tan C, Madden S. Osprey: ImplementingMapReduce-style fault tolerance ina shared-nothing distribu-ted database//Proceedings of the 24th International Confer-ence on Data Engineering(ICDE‘10).Long Beach, California,USA, 2010
[21] Floratou A, Patel JM, Shekita E J, Tata Sandeep. Column-oriented storage techniques for MapReduce. PVLDB, 2011
[22] Jens Dittrich, Jorge-Arnulfo Quian-Ruiz, Alekh Jindal,YagizKargin, Vinay Setty, Jrg Schad. Hadoop+ + : Mak-ing a yellow elephant runlike acheetah (without it even not-icing). PVLDB, 2010
[23] Pavlo A, PaulsonE, RasinA, AbadiD J, DeWitt D J, Mad-den S, StonebrakerM. A comparisonof approachesto large- scale data analysis//Proceedings of the ACM SIGMODInter-national Conference on Management of Data(SIGMOD’09).Providence, Rhode Island, USA, 2009
[24] Brewer E A. Towards robust distributed systems//Proceed-ings of the 19th Annual ACM Symposium on Principles of Distributed Computing (PODC‘00). Portland, Oregon,USA, 2000
[25] Fushimi S, Kitsuregawa M, Tanaka H. An overview of thesystem software of a parallel relational database machine//Proceedings of the 12th International Conference on VeryLarge Data Bases(VLDB‘86). Kyoto, Japan, 1986
[26] DeWitt D J, Gerber R H, Graefe G, Heytens M L, KumarK B, Muralikrishna M. GAMMA A high performance dat-aflow database machine//Proceedings of the 12th International Conference on Very Large Data Bases(VLDB‘ 86).Kyoto, Japan, 1986
[27] Dean J, Ghemawat S. Map Reduce: Simplified data process-ing on large clusters//Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI‘ 04).San Francisco, California, USA, 2004
[28] TDWI Checklist Report: Big Data Analytics. http://tdwi.org/research/2010/08/Big-Data-Analytics.aspx
[29] Brewer E A. Towards robust distributed systems. Proceedings of the 19th Annual ACM Symposium on Principles of Distributed Computing (PODC’’00) .2000
[30] Chaudhuri S, DayalU. Anoverview ofdata warehousing and OLAP technology. SIGMOD Rec, 1997