## 《大数据导论》研读讨论作业

## Main Ideas and Major Contributions of Readings Discussion 1 (in Chinese)

备注：小组讨论翻译并理解给出的研读文献，无需全文直译，但需要撰写文献的核心思想、主要过程和主要贡献，依据撰写质量评价给分

Name: \_\_\_\_黄昊\_\_\_\_ Student ID: \_\_\_\_20204205\_\_\_\_

Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275 (2014) 314–347

这篇文章主要介绍了大数据的一些基础知识，包括大数据的简介、大数据的机遇和挑战，以及现今采用的处理大数据问题的最先进的技术和工艺。还讨论了处理大量数据的几种基本方法，粒计算、云计算、仿生计算和量子计算。

该文首先介绍了什么是大数据，即大数据是高容量、高速度和高多样性的数据。一般说来，如果在目前的技术条件下，对一个数据集进行捕捉、整理、分析和可视化是非常困难的，这样的数据就可以称之为大数据。数据密集型科学正在成为前三种科学范式（即经验科学、理论科学和计算科学）中的第四种科学范式。

然后介绍大数据问题在商业和企业、社会管理和科学研究领域的几个应用。

在商业中的大数据体现在大公司尤其是跨国公司中，这些公司每天都有很大的信息产生，各大公司为了增加商业竞争力都非常重视这些数据的处理，未来基于大数据的商战或许会日益增多。

在公共管理领域的大数据体现在社会的每个人在每个公共部分产生大量数据，在人口多的国家所产生的数据量更加巨大 ，各国政府把这些数据当作潜在资源，充分发挥大数据的作用，为公共部门提供了提高生产力和更高水平的效率和效益的机会。

在科学研究中的大数据体现在许多科学领域随着计算机的发展产生了大量的各种类型的数据，如何利用这些数据获取知识还仍是未知的。

在第三部分本文介绍了大数据的机遇和挑战。机遇即为大数据能够给多方带来很大的好处，成为企业的基本竞争，利用大数据可以让企业在商业领域获得许多优势，也能为社会带来更多就业岗位和精通数据的人才，给国家的卫生保健，公共管理等带来许多潜在价值和利益。挑战即为面对大数据有很多困难，困难在于数据采集、存储、搜索、共享、分析和可视化。

在数据采集和存储方面，数据集的规模不断扩大。在许多领域，一些有价值的数据经常因为没有足够的空间来存储这些数据而被删除。

在数据传输方面，带宽容量是云和分布式系统的瓶颈，尤其是在通信量很大的情况下。

在数据管理方面，现有的数据库管理工具无法处理如此庞大和复杂的大数据。当前，数据仓库和数据集市是基于标准查询语言 (SQL) 的数据库系统。NoSQL也是目前大型和分布式数据管理和数据库设计的一种方法，两者各有优势。

在数据分析方面，当要处理的数据量非常大时，如何才能保证响应的及时性，这仍然是一个很大的挑战。数据安全问题也是很大的挑战，因为大数据的规模非常大，使保护方法受到影响以及工作量的增加。分布式方式存储的大数据也会有来自网络的威胁。

在数据可视化方面，由于大数据的大尺寸和高维度，进行数据可视化非常困难。如何思考大数据的可视化方式非常重要，这也需要更多的大数据方面的人才，而现在大数据方面的人才短缺，给可视化也带来了更大的困难。

在第四部分主要介绍了处理大数据的最新技术和工艺，主要介绍了大数据相关的技术以及一些处理工具。

大数据涉及多个学科的不同的技术，包括优化方法，统计，数据挖掘，机器学习，人工神经网络（ANN），社会网络分析（SNA），信号处理，模式识别，可视化方法等。高维数据也是大数据的一个问题，而通过降维是现今最先进的处理高维度数据的技术。

该文介绍的批处理工具主要内容总结如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 主要内容 | 指定用途 | 优势 |
| Apache Hadoop | / | 基础设施和平台 | 高扩展性、可靠性、完整性 |
| Dryad | / | 基础设施和平台 | 分布式执行引擎，良好的可编程性 |
| Apache Mahout | 核心算法包括聚类、分类、模式挖掘、回归、降维、进化算法和基于批次的协作排序 | 商业中的机器学习算法 | 良好的成熟度 |
| Jaspersoft BI Suite | 具有快速数据可视化的能力，可以快速探索大数据，而无需提取、转换（ETL）和加载 | 商业智能软件 | 具有成本效益的、大规模的自助式BI |
| Pentaho Business Analytics | 可以从结构化和非结构化的大量数据中生成报告 | 商业分析平台 | 知识发现的稳健性、可扩展性和灵活性 |
| Skytree Server | / | 机器学习和高级分析 | 高速准确地处理海量数据集 |
| Tableau | Tableau Server提供基于浏览器的分析，Tableau Public用于创建交互式可视化 | 数据可视化，商业分析 | 更快、智能、精细、漂亮和易于使用的仪表板 |
| Karmasphere Studio and Analyst | / | 大数据工作空间 | 基于协作和标准的无约束分析和自我服务 |
| Talend Open Studio | / | 数据管理与应用一体化 | 易于使用的、基于eclipse的图形环境 |

批处理工具中最著名和最强大的是Apache Hadoop，实现了名为Map/Reduce的计算范式。Map/Reduce的工作原理即为归并算法的原理，分而治之然后结合各个子问题的解决方案作为原问题的解决方案。主要是Map步骤和Reduce步骤。Hadoop集群中分主节点和工作节点。主节点接受输入，将其划分为更小的子问题，并在Map步骤中将其分配给工作节点。然后主节点收集所有子问题的答案，并以某种方式将其合并，形成Reduce步骤中的输出。在Hadoop中主节负责监控所有的数据节点。如果有一个数据节点未能执行相关任务，主节点将要求该数据节点或另一个数据节点重新执行失败的任务。

流处理工具主要内容如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 简介 | 指定用途 | 优势 |
| Storm | / | 实时计算系统 | 可扩展、可容错，且易于设置和操作 |
| S4 | / | 处理连续的无限制的数据流 | 成熟的、分布式的、可扩展的、容错的、可插拔的平台 |
| SQLstream s-Server | 使用内存中处理(“NoDatabase”)技术 | 传感器、M2M和远程信息处理应用 | 基于SQL的实时流式大数据平台 |
| Splunk | 特点包括索引结构化或非结构化的机器生成的数据、实时搜索、报告分析结果和仪表盘 | 收集和利用机器数据 | 快速和易于使用，动态环境，可从笔记本电脑扩展到数据中心 |
| Apache Kafka | 通过内存分析技术管理流媒体和操作数据，以获得实时决策 | 分布式发布‑订阅消息系统 | 高吞吐量的不可变活动数据流 |
| SAP Hana | / | 实时业务平台 | 快速内存计算和实时分析 |

在Storm集群上，用户为不同的Storm任务运行不同的拓扑结构，与Map/Reduce不同，Map/Reduce作业最终会结束，而拓扑结构则一直在处理信息，或者直到用户终止它。一个Storm集群由两种工作节点组成，一个主节点和几个工作节点。主节点和工作节点实现了两种守护进程Nimbus和Supervisor。Nimbus负责在Storm集群中分发代码，调度工作，将任务分配给工作节点，监控整个系统。如果集群中出现故障，Nimbus将检测到它并重新执行相应的任务。Supervisor遵从Nimbus分配的任务，并根据Nimbus的指示在必要时启动或停止工作进程。整个计算拓扑结构被分割并分配给若干个工作进程，每个工作进程实现拓扑结构的一部分。守护进程Zookeeper用于协调系统，它在本地磁盘上记录Nimbus和Supervisor的所有状态。

交互式分析工具将数据呈现在交互式环境中，允许用户进行自己的信息分析。数据可以以表格或图形格式或两者同时进行审查、比较和分析。本文主要介绍了两款交互式分析工具。

谷歌的Dremel。通过结合多级执行树和列式数据布局，可以在几秒钟内对万亿行的表格运行聚合查询。该系统可扩展到数千个CPU和数千字节的数据。

Apache Drill，有更多的灵活性来支持各种不同的查询语言、数据格式和数据源。目标是在10,000台服务器上进行扩展，并在几秒钟内达到处理PB级数据和数万亿记录的能力。

Drill和Dremel都是专门为有效利用嵌套数据而设计的，使用HDFS进行存储，使用Map/Reduce进行批量分析。

第五部分主要讲的是设计大数据系统的原则，原则如下：

1. 良好的架构和框架是必要的，而且是最优先的。
2. 支持各种分析方法。
3. 没有万能的模式。
4. 将分析带入数据。
5. 处理过程必须是可分配的，以实现内存中的计算。
6. 数据存储必须是可分配的，用于内存存储。
7. 处理单位和数据单位之间需要协调。

第六部分主要讲的是基础技术和未来研究，包括粒计算、云计算、仿生计算和量子计算。

粒计算（GrC，是各种算法的合称）可以将数据大小降低到不同的粒度级别。GrC的理论基础非常健全，涉及集合论、模糊集、粗糙集和随机集。不同级别的颗粒所代表的信息显示出不同的知识、特征和模式，其中不相关的特征被隐藏，有价值的特征被突出。缺点是如果数据规模缩小到很小，隐藏在大数据中的信息可能会部分丢失。

虚拟计算机的使用被称为云计算，它通过Internet提供应用程序和服务，还扩展到基础设施即服务。云环境带来的好处是云存储，为存储大数据提供了可能的工具。云存储在存储信息方面具有良好的可扩展性和可扩展性，还解决了与数据传输和共享有关的挑战之一，因为云中的数据集和分析结果可以与他人共享。缺点是在云环境中上传和下载大量数据所需的时间和成本很大。在公众可访问的服务器上托管数据集也存在隐私问题。

仿生计算，有高效的机制来组织、访问和处理数据。计算智能的灵感来自于自然，与自然界相类似，仿生的硬件系统可以被分为三个轴心，即系统发育、本体发育和表观发育。生物计算机的灵感来自于生物分子，以进行涉及存储、检索和处理数据的计算计算。它整合了生物衍生材料来执行计算功能并获得智能和高效的性能。包括生物化学计算机、生物机械计算机和生物电子计算机。

量子计算，量子计算机的内存非常大，并且可以同时操作一组指数级的输入。它还可以在希尔伯特空间的黄昏地带进行计算。量子计算机可以解决目前计算机上异常困难的问题，包括大数据问题。量子计算机使用量子比特或量子位。这是因为量子算法比传统算法更有效、更快速。