计算机科学导论期末作业：计算机前沿综述

大数据与云计算

作者：XXX班级YYY

摘要

奥地利作家维克托·迈尔-舍恩伯格在他所出版的书籍《大数据时代》中说：“大数据带来的信息风暴正在变革我们的生活、事情和思维，大数据开启了一次重大的时代转型”，这才刚刚过去五年，大数据与云计算便如他所言，开始改变我们的生活。

关键词：大数据 云计算 物联网 云服务

**目录**

引言

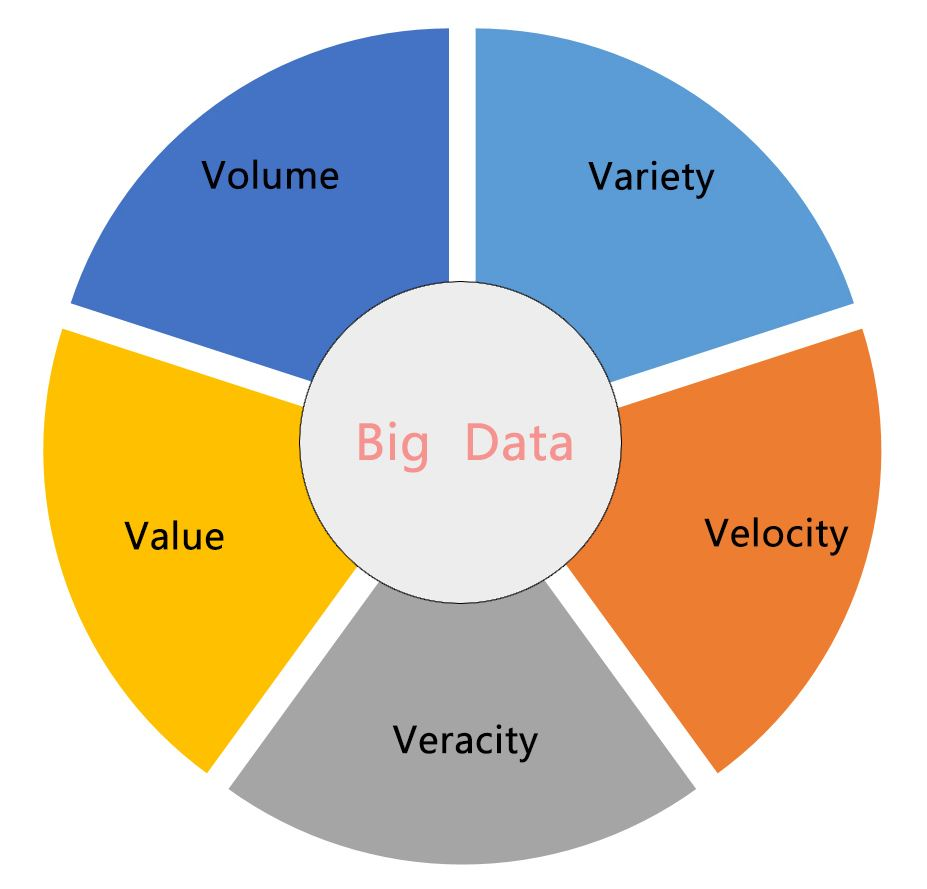
近年来，大数据和云计算已然成为互联网中的热门词汇，从阿里云到腾讯云，“云”被各个企业纷纷效仿，成为一种新技术的代名词，即使是一些与“云”关系不是十分密切的网盘也扣上了“云”的帽子。现如今，当我们在百度或者淘宝输入一组关键字进行搜索之后，不久就会发现页面上的广告、主页上的商品推荐仿佛懂得了我们的想法，向我们展示着我们想要获得的东西，类似的精准营销的确造成了一些用户隐私上的问题，然而不可否认的是能否掌握大数据已经成为企业的重要资产，成为企业在市场竞争的过程中的主要保障。云计算作为大数据得以实现的重要支撑，则是各个企业都在加紧研究的对象，它使得企业能够节约成本，同时提升性能和效率。

本文从大数据和云计算的概念开始介绍，以当前标准制定组织所发布的信息作为参考，深入浅出的将大数据的特点、应用、当前存在的问题以及云计算中涉及的概念进行阐释，并辅以生活中的联系，最后叙述了大数据与云计算之间的联系，揭示它们之间不可分割的关系，从而对大数据与云计算这个领域作了一个整体的概述。

一、大数据：

新兴的大数据浪潮来势汹涌，大数据时代的来临慢慢对我们生活的各个方面产生巨大的影响。到目前为止，大数据并没有一致的定义，国际数据公司（International Data Corporation，IDC）对大数据的定义是：为了更经济地从高频率获取的、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。正如其名字所言，大数据（Big Data），或者称作是巨量资料，通常是指在一定的时间区间内难以被常规的数据软件抓取并进行综合分析和管理的体量巨大的数据的集合，与此同时，大数据还泛指管理和利用大规模数据的商业模式和技术平台。

从IDC对大数据的定义中可以看出，大数据具有方面覆盖广阔、数据规模呈几何级数增长特性。IBM把大数据的特点总结为5个V，即：



1、数据体量巨大（Volume）：

一般认为大数据的单位以“拍字节（PB）”为起点，谷歌的Eric Schmidt则更是说道：从人类文明曙光到2003年数以万年计的时间长河里人类一共产生了5EB，而到2010年每两天人类就能产生5EB的数据。

2、数据种类丰富多样（Variety）：

数据可以根据是否可以使用关系型数据库来表示和存储分为结构化数据、半结构化数据以及非结构化数据，其中结构化数据通常具有因果关系，而大数据中出现更多的半结构化数据和非结构化数据（超过80%）主要是文本、网页，音视频等。大数据的多样性不仅表现在数据的类型上，还表现在它的来源以及用处上。

3、数据处理迅速（Velocity）：

大数据中数据具备动态的特征，但是动态数据也有着各异的形式：其中一些每秒钟可能产生PB量级的数据，如欧洲核子研究中心（CERN）的大型强子对撞机(Large Hadron Collider，LHC)，而另外一些却是产生缓慢而又持久的模式。要充分利用静止数据以及汹涌而来的动态数据，就要求我们对正使用中的数据的快速处理，这恰好与一对完全不同的处理范式相匹配：批处理和流处理。前者是数据不动逻辑动，通过发送代码计算数据得到其价值，其典型范式是Hadoop；而后者则是逻辑不动数据动，通过将动态数据引入并留下计算得到的价值、将原始数据扔掉或加入静止数据的方法处理数据，流处理范式丰富多样，在此不胜枚举。大数据处理起来要求快速，经常要求在一秒之内完成，能够从不同类型的数据中快速的挑取整合有用的内容，从而与通常意义上的数据挖掘技术有着根源上的差别。

4、数据准确、具有高质量（Veracity）：

大数据中要处理的数据必定要与现实世界中出现的事件有着密不可分的联系，而之所以分析大数据就是要从体量巨大的网络数据中得到可以说明并预测现实世界中会发生的事情的数据，因此提取出高质量的数据是分析大数据的必然要求。数据之所以重要性就是因为它能够支持决策，大数据分析中应当筛选并删除数据有偏向、作假、疑问的部分，避免这些数据破坏了资料系统的完整性和准确性，从而使最终的决策受到影响。

5、价值密度低、回报高（Value）：

大数据中的价值体现在从体量巨大的数据中对数据进行深度挖掘获得的信息，大数据中隐藏的价值非常巨大，但也存在价值密度低的特点，很多时候需要对数据进行高强度的挖掘才能得到获取到有实际价值的信息。

万国商业机器公司（International Business Machines Corporation，IBM）提出的大数据特征最为广泛接受，除此之外，由于大数据的数据来源渠道多样化、数据量体量巨大，一些其他人认为大数据具有可变性（Variability）、复杂性（Complexity）等其他特征。总之，要想全面的了解大数据，必须经过从理论、技术以及实践等方面系统的分析过程。

**大数据发展中存在的问题：**

大数据的发展过程中极易出现“孤岛效应”，大数据愈变愈大，有用信息掩盖的却愈来愈深，对大数据缺少有效的管理和分析办法，大数据的采集和分析面临诸多挑战，另外数据集中掌握在少数企业手中，而为了保护企业利益，这些数据也一般不会提供给其他人，信息的重复采集也造成了巨大浪费。另外，尽管在大数据方面中国已成为全世界最主要的市场之一，但中国的大数据产业才刚刚起步，企业愈发重视大数据，希望获得更多的更多个人的数据，然而却可能在不经意间侵犯了个人的隐私。近来互联网上关于隐私的事件频频发生，用户定应及时了解清楚各个平台的隐私协议，保护自己的权利不受侵害。

二、云计算：

云计算并不是如名字一样在云里计算，而是通过分布在不同地方的许多计算机构建而成的的资源池中实现计算的任务：一台计算机的计算能力有限，而如果能像并行计算那样把很多台计算机的计算能力叠加，计算功能就更加强大。云计算就是利用互联网上散乱分布的计算机将它们的计算能力充分利用来达到计算的目的。随着云计算技术的提升，伴随产生了商务云、政务云等各种云服务。一般客户只需要支付一定的费用向商家购买相关服务，商家要做的就是提供服务，包括资源的更新等，客户只需使用该服务（共享的硬件）而不必花钱来提高自己机器的配置，这种方式无疑更加廉价。对于用户而言，同时又可以将自己的闲置的资源（如计算机）出售或者共享，让自己的机器也成为云计算中的一部分，云计算使得有限的计算资源得到了更加合理的利用，而使用云服务就像使用一个工具一样简单。

按照美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）的说法，云计算（Cloud Computing） 是一个很方便的、可以实现无处不在的对一个共享的可配置资源池（例如，网络、服务器、存储设备、应用程序以及服务）进行访问的一个模型，这个资源池可以通过最小的管理工作投入或服务提供商（Service Provider，SP）交互来快速配置和发布。NIST认为云模型有五个本质特征：按需自助服务、广泛的网络连接、资源池、快速的可伸缩性以及可度量的服务，三个服务模型：软件即服务（Software as a Service，SaaS）、平台即服务（Platform as a Service，PaaS）以及基础建设即服务（Infrastructure as a Service，IaaS），四个发布模型：私有云、社区云、公共云、混合云，从中我们可以看出“云”的规模之大，它所带来的计算能力之强。

**使用云服务的优点：**

服务提供商所提供的云计算一般会使用虚拟化技术，一些比较有名气的云服务提供商如Amazon、阿里巴巴、Google云计算等提供的云服务，只需要注册一个账号，购买他们所提供的服务，就可以像使用一个应用程序一样使用它们，虚拟化技术已经让使用云服务变得“察觉”不到应用实体。

使用云服务的另一个好处是易于控制“云”的规模，当所需的硬件要求提升时可以很容易的做出改变，便于满足使用者的用户规模增长的需求。

当今社会，网络安全已然成为各类企业和我们个人所要直面的重大挑战，借助云服务提供商的安全团队能够有效应对恶意攻击，使得企业和个人的安全风险大幅度降低。也许你不曾注意到，云计算早已深入到我们的日常生活中，它已然是我们平时所用的很多网站和人家背后的强大服务支持。

三、大数据与云计算的关系

我们讨论大数据和云计算绝对不能将二者割裂开来，它们不是独立的，并且关系非同寻常。大数据技术的基础架构支持便是云计算平台和云存储技术，云计算自上世纪八十年代产生，历史显然比大数据更加长远。云计算与大数据情同姐妹，是相互之间不能缺少的，没有大数据，云计算难以发挥其优势，而没有了云计算，大数据更是难以普遍实现。

大数据的特色在于对海量数据的挖掘，然而它必须依赖于分布式数据挖掘、分布式处理和分布式数据库、云存储和虚拟化，其与云计算的关系可见下表：

|  |  |  |
| --- | --- | --- |
| 云计算服务模型 | 相关技术 | 相关软件 |
| SaaS | 分布式数据挖掘 | Mahout |
| PaaS | 分布式处理 | MapReduce，JobKeeper |
| 分布式数据库 | HBase，数据立方 |
| IaaS | 云存储 | HDFS，cStor |
| 虚拟化 | VMware，OpenStack |

随着大数据格式的愈来愈复杂化，数据源愈来愈多样化，传统的服务器已经难以处理如此庞大的数据，尤其是格式并不友好的半结构化数据和非结构化数据；而通过云计算，这些数据被分散到虚拟化的云服务器上，庞大的数据也被逐步分解，从而达到了高效处理数据的目的。目前，基于云计算的大数据分析蓬勃发展，这足以看出二者的密切关系。

而云计算与大数据的区别在于大数据的应用要比云计算的应用要更广泛，大数据的产生也要早于云计算；大数据存在的意义是想要挖掘体量巨大的数据中的信息，而云计算则是想要借助互联网这个平台更充分便捷地访问和管理各种资源，使资源得到充分地利用，云计算的发展使得传统的服务器面临挑战，传统服务器占用空间大并且难以维护，云计算帮我们解决了这些问题，并且减少了多余的服务器资源的浪费。二者概念上的不同决定了它们影响的方向，前者强调了计算的过程，而后者则反映了计算的对象。

结语

我们生活在一个互联网的时代，海量的数据无时无刻不围绕在我们身边。试想我们未来的社会完全进入到大数据和云计算的时代，到那时大数据已经完全可以预测我们的所想，在我们想买一件衣服时，快递已经送上家门；在我们想要出门时，司机已经在门口等候，这些不是没有可能发生。

参考文献

[1]. 统计与管理,2013年6月,热点透视,河北统计局,李永宏,《大数据与云计算》

[2]. 大数据浪潮：打破电商信息孤岛的边界, <http://www.ebrun.com/20130805/79075.shtml>

[3]. 中国电子商务,2014年9月,信息化建设,武汉职业技术学院计算机学院,方东菊,《大数据与云计算》

[4]. 福建电脑,2014年第7期,大庆师范学院计算机科学与信息技术学院,王涛,邵国强,《基于云计算的大数据分析》

[5]. NIST Special Publication 800-145, September 2011, Peter Mell, Timothy Grance, 《The NIST Definition of Cloud Computing》

[6]. communications of the acm, April 2010, Michael Armbrust, Armando Fox, Rean Griffith,Anthony D. Joseph, Randy Katz, Andy Konwinski,Gunho Lee, Dav id Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia, 《A View of Cloud Computing》