数据管理在云计算领域的应用

学号：		姓名：	

摘要：

随着科技产业的高速发展，互联网渗入到人们生产生活的方方面面，近年来，许多领域都产生了海量的数据，大数据时代的风暴已经席卷全球。而对这些数据的分析和处理，云计算提供了充沛的运算能力，但数据产生的速度越来越快，多样性更加丰富，云计算对数据管理的需求十分旺盛，已经成为整个领域所关注并研究的热点问题。本文就数据管理在云计算领域的应用进行了深入阐述，具体分析了数据管理相关方法在云计算的数据冗余缩减、数据存储、和数据分析过程中的应用，并结合当今应用需求和新技术对数据管理和云计算的发展趋势、应用前景进行展望。

关键词：

数据管理 云计算 数据获取 数据存储 数据分析

媒体数据管理大作业 Only For Preview

1 引言

如今，互联网和计算机技术已经融入到了我们社会生活的各个方面。来自互联网、物联网、电子商务、科学研究等领域所产生的数据呈现指数增长的趋势，为了处理这样日益增长的数据，云计算的服务市场规模逐年增加。预计在未来几年仍然保持较快增长态势，到 2022 年市场规模将达到 1171.6 亿元。在国内，由于政策、信息安全以及本地应用习惯等原因，阿里云当前占据了最大的市场份额。阿里云在中国云计算市场居于第一名的领先地位，占据了约 43.2%的市场份额，处于明显领先的地位。从全球市场上，2018 年阿里云全球公有云市场份额第三，已超过 Google、IBM 的云业务。[1]

虽然海量的数据使得我们可以处理、分析并利用的数据大量增加，但大量的数据却给云计算中的数据采集、存储、维护、处理等带来了极大的挑战，传统的数据处理方法已经不能够满足云计算的处理需求，数据产生的速度和多样性，给数据管理、分析处理以及网络带宽等方面都带来了前所未有的挑战，并已经成为整个领域所关注并研究的热点问题。

本文首先对云计算技术进行简要介绍，研究云计算环境下数据管理的需求和问题。之后对云计算领域中数据管理相关方面使用的技术进行分析，针对云计算

过程中面临的几个重要的数据管理问题：数据冗余缩减、分布式数据存储、媒体数据分析这些方面进行详细讨论。最后结合当今应用需求和新技术对数据管理和云计算的发展趋势、应用前景进行展望。

2 云计算架构和数据管理的关系

2007 年 Google 率先提出了云计算的概念和理念，云计算是一种新兴的商业计算模型，可以通过互联网的方式，以便利的、按需付费的方式来获取用户所需的计算资源（网络带宽、存储空间、服务器、应用）等，这些资源都是分布在一个共享的、可配置的，由大量计算机所构成的资源池中，并可以提供 IT 资源的弹性伸缩服务，且能以最省力、最方便和无人干预的方式进行实时资源获取和释放。

通过采用云计算技术，系统内各类资源能被有效整合,并基于虚拟化技术,来实现内部计算能力、存储空间、网络带宽等资源的共享,分配比较灵活,可以适应当前飞速发展的业务需求,同时降低 IT 总体持有的服务成本

2.1 云计算的体系结构

云计算技术体系结构的一般结构图，如图 1 所示。云计算的技术体系结构一般可以分为四层[2] 分别是:

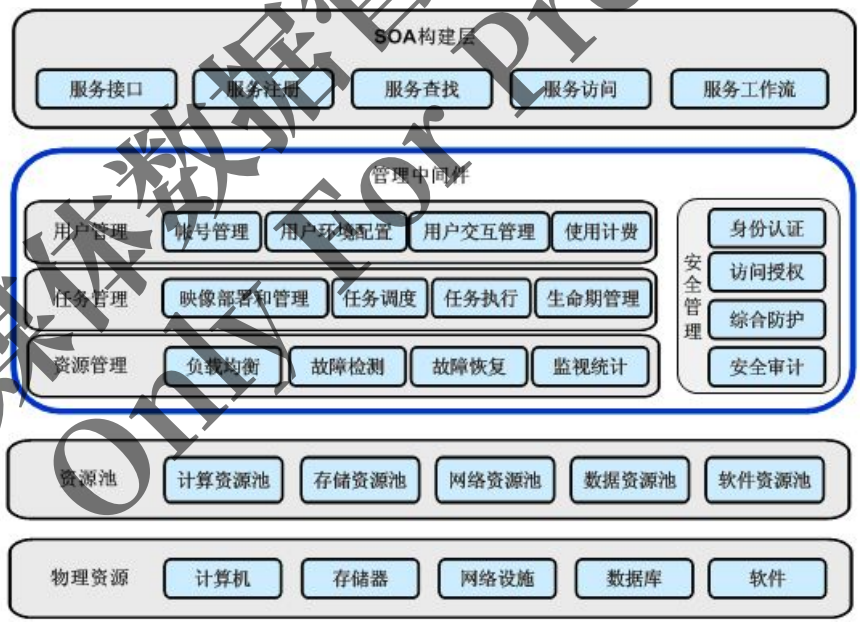

图 1：云计算体系结构

1. 物理资源层：包括各种计算机、存储器、数据库、软件，以及网络设施等，用以确保用户的各种请求。
2. 资源池层：将云平台中的各类资源进行有效整合，尽量达到同构或近似同构的资源池。在该层中，重点是所有物理资源的有效集成以及管理等。
3. 管理中间件层：该层的主要任务是对云平台中的所有资源进行高效管理，同

媒体数据管理大作业 Only For Preview

时，可以对任务进行调度，以达到所有的资源可以为用户提供安全、可靠，以及高效的优质服务。

4. SOA 构建层：云计算的能力被封装成标准的 Web Services 服务，并嵌入到 SOA 体系中被使用和管理。例如:服务注册管理、服务查找管理、服务访问管理和工作流管理等。

2.2 云计算环境下的数据管理需求

近年来，随着互联网产业和技术的迅猛发展，传统用户从数据的消费者已经变为了数据的生产者，这种以用户为中心的数据生成模式使得互联网中的数据呈现了指数级增长的趋势，给云计算系统的管理带来了严峻的挑战。

由于数据类型的多样性和复杂性，因此，来自采集端的不同数据库的数据必须被导入到一个集中式的大规模分布式存储集群中，才可以对这些数据进行有效的分析。在导入的过程中，可以根据需要有针对性的对数据进行一定的预处理工作。在数据的导入和预处理阶段，我们面临的主要问题是数据量的巨大性，每秒钟数据的导入量可以达到百兆、千兆级。如何有效的进行据的导入也是值得考虑的问题。

除了传统的如文本、HTML 数据等，越来越多的音频、视频等多媒体数据也需要进行处理和分析，在媒体数据分析的过程中，也需要使用到数据管理的相关技术。

3 云计算使用的数据管理技术

1.1 数据冗余缩减

云计算需要处理大量的数据，而在这些数据中可能存在一些无意义的数据，数据冗余无疑会增加传输开销，浪费存储空间，导致数据不一致，降低可靠性。影响后续的数据分析，因此需要对数据进行预处理，在不损失数据价值的前提下进行冗余的缩减和数据压缩，减少云计算过程的存储和计算开销。因此许多研究提出了数据冗余减少机制，例如冗余检测[3]和数据压缩[4]等。

在云计算的安全监控领域，广泛部署的摄像头收集的图像和视频数据存在着大量的时间、空间和统计上的冗余。视频压缩技术被用于减少视频数据的冗余，许多重要的标准(如 MPEG-2，MPEG-4，H.263，H.264/AVC) 已被应用以减少存储和传输的负担[5]。Tsai 等在文献[6]中研究了通过视频传感器网络进行智能视频监控的视频压缩技术，如图 2。通过发现场景中背景和前景目标相联系的情境冗余，他们提出了一种新的冗余减少方法。

媒体数据管理大作业 Only For Preview

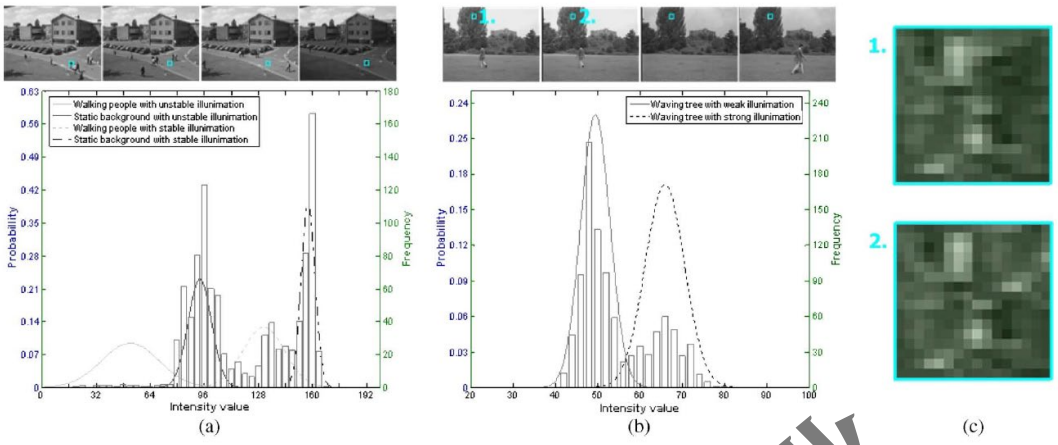

图 2：智能视频监控的视频压缩技术

在一般的数据传输和存储过程中，数据去重(data deduplication) 技术[7]用于消除重复数据的副本。在存储去重过程中，一个唯一的数据块或数据段将分配一个标识并存储，该标识会加入一个标识列表。当去重过程继续时，一个标识已存在于标识列表中的新数据块将被认为是冗余的块。该数据块将被一个指向已存储数据块指针的引用替代。通过这种方式，任何给定的数据块只有一个实例存在。去重技术能够显著地减少存储空间，对分布式存储系统具有非常重要的作用。

除了前面提到的数据预处理方法，还有一些对特定数据对象进行预处理的技术，如特征提取技术，在多媒体搜索[8] 和 DNS 分析[9] 中起着重要的作用。 这些数据对象通常具有高维特征矢量。 数据变形技术则通常用于处理分布式数据源产生的异构数据，对处理商业数据非常有用。Gunter 在文献[10] 中提出了 MapLan，对瑞士国家银行的调查信息进行影射和变形。Wang 等在[11] 中提出了一种在分布式存储系统中异构感知的数据重生成机制，如图 3。在异构链路上传递最少的数据以保持数据的完整性。

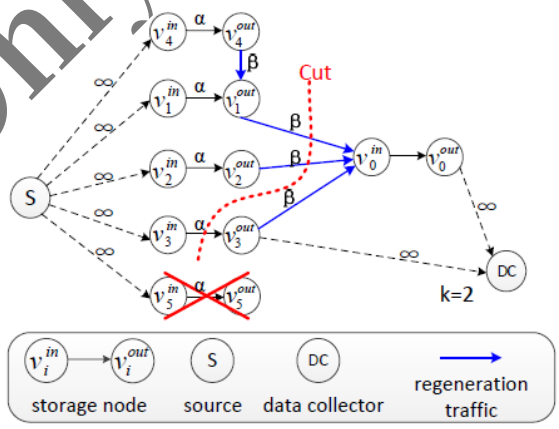

图 3：异构链路上的分布式数据传递

这些方法能够用于不同的数据集和应用环境，提升性能，但同时也带来一定风险。例如，数据压缩方法在进行数据压缩和解压缩时带来了额外的计算负担，

媒体数据管理大作业 Only For Preview

因此需要在冗余减少带来的好处和增加的负担之间进行折中。

1.2 分布式数据存储

文件系统是分布式计算的基础，Google 为大型分布式数据密集型应用设计和实现了一个可扩展的分布式文件系统 GFS[12]。GFS 运行在廉价的商用服务器上，为大量用户提供容错和高性能服务。GFS 适用于大文件存储和读操作远多于写操作的应用。但是 GFS 具有单点失效和处理小文件效率低下的缺点 Colossus 改进了 GFS 并克服了这些缺陷。其他的企业和研究者们开发了各自的文件存储解决方案以适应不同的大数据存储需求。HDFS 是 GFS 的开源产物；Microsoft 开发了 Cosmos 支持其搜索和广告业务[13]; Facebook 实现了 Haystack 存储海量的小照片[14]; 淘宝则设计了两种类似的小文件分布式文件系统: TFS1 和 FastFS。

在文件系统的基础之上，云计算领域中广泛使用数据库技术来方便数据的存储和管理。由于云计算中，数据可能分布在多个不同地域，不同架构的服务器中，传统的关系型数据库系统已经难以解决云计算中多样和海量的数据需求，因此云计算领域普遍采用模式自由、易于复制、API 简单、支持海量数据的 NoSQL 数据库。

键值存储是一种简单的数据存储模型,数据以键值对的形式储存，键是唯一的。近年出现的键值存储数据库受到 Amazon 公司的 Dynamo 影响特别大[15]。在 Dynamo 中， 数据被分割存储在不同的服务器集群中， 并复制为多个副本。Dynamo 的分割机制基于一致性哈希技术[16]，将负载分散在存储主机上。哈希函数的输出范围被看作是一个固定的循环空间或“环”。如图 4 所示，系统中的每个节点将随机分配该空间中的一个值，表示它在环中的位置。通过哈希标识数据项的键，可以获得该数据项在环中对应的节点。

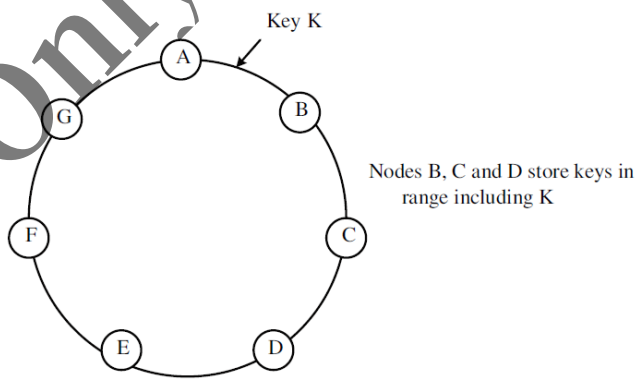

图 4：哈希环用于数据分隔

由于每条唯一的数据项存在多个副本，Dynamo 允许以异步的方式更新副本并提供最终一致性。每次更新被认为是数据的一个新的不可改变的版本。 一个

媒体数据管理大作业 Only For Preview

对象的多个版本可以在系统中共存。

除此之外，主流的 NoSQL 数据库还有以 Google 的 Bigtable 数据库为代表的列式存储数据库。以及支持存储更加负载的数据结构的文档数据库，如 MomgoDB 等。这些数据库都支持将数据进行分片和复制，实现存储的负载均衡。

3.3 媒体数据分析

在分布式计算所处理的海量数据中，除了传统的结构化数据（如文本数据、HTML 数据）之外，更有大量的如图像和视频之类的多媒体数据，媒体数据分析是指从多媒体数据中提取有用的信息，理解多媒体数据中包含的语义信息。进行进一步的分析和处理。

音频摘要可以简单地从原始数据中提取突出的词语或语句，合成为新的数据表达;视频摘要则将视频中最重要或最具代表性的序列进行动态或静态的合成。静态视频摘要使用连续的一系列关键帧或上下文敏感的关键帧表示原视频，这些方法比较简单，并已被用于 Yahoo，AltaVista 和 Google， 但是它们的回放体验较差。动态视频摘要技术则使用一系列的视频片段表示原始视频，并利用底层视频特征进行平滑以使得最终的摘要显得更自然[17]

多媒体标注是指给图像和视频分配一些标签，可以在语法或语义级别上描述它们的内容。在标签的帮助下，很容易实现多媒体内容的管理、摘要和检索。由于人工标注非常耗时并且工作量大，没有人工干预的自动多媒体标注得到了极大的关注。尽管取得了一些重要的进展，目前的自动标注方法性能并不能令人满意。一些研究开始同时利用人和计算机对多媒体进行标注[18]。

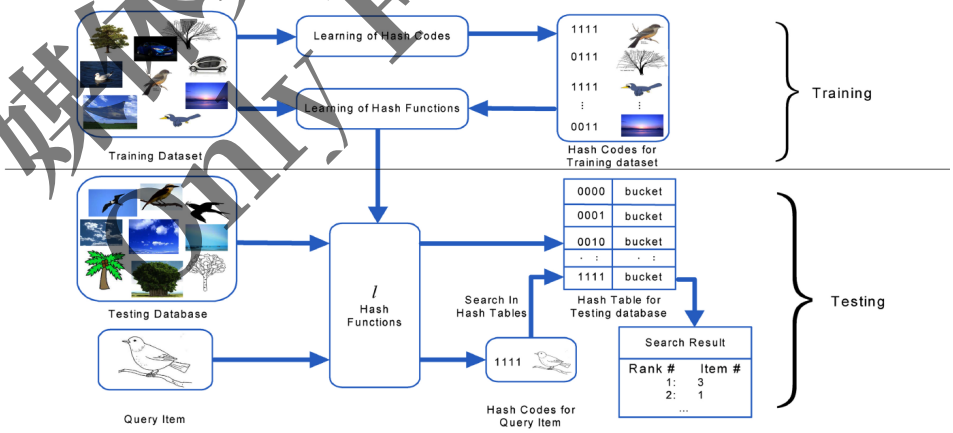

图 5：哈希方法的相似多媒体检索

多媒体索引和检索处理的是多媒体信息的描述、存储和组织，并帮助人们快速方便地发现多媒体资源。一个通用的视频检索框架包括 4 个步骤:结构分析，特征提取，数据挖掘、分类和标注，以及查询和检索。Shao 等[19]提出一种基于内容的视频检索方法，通过时间和空间定位从数据库中有效地检索相关行为的视

媒体数据管理大作业 Only For Preview

频。在大规模图像检索方面，Chen 等[20] 提出一种基于图哈希的方法(spectral embedded hashing)。 Song 等[21] 提出一种基于哈希方法的近似相似多媒体检索，如图 5 所示，通过机器学习方法有效地学习一组哈希函数来给数据产生哈希码。

4 发展前景展望

随着云计算产业、基础设计的不断完善，“云原生”的概念也逐渐被企业和开发者接受，衍生出了如 Kubernetes、OpenShift 等分布式生态系统，软件从诞生之日起就生在云上，长在云上，开发人员从一开始就通过容器和 Kubernetes 部署云原生应用。此外有不上的硬件厂商针对性的研发了适用于云计算的专属配套硬件，这些服务器在设计之初就是为云端负载而生，通过硬件加速云计算的数据管理和运行在云环境中的应用程序。

云计算将顺应产业互联网大潮，下沉行业场景，向垂直化产业化纵深发展。典型的垂直云代表有视频云、金融云、游戏云、政务云、工业云等。以视频云为例，它是将视频采集、存储、编码转换、推流、视频识别等一系列以视频为核心的技术能力整合为一站式垂直云服务，涉及到海量的视频数据管理。配合摄像头硬件和边缘计算节点进军广阔的线下安防监控市场。再如金融云，可针对金融保险机构特殊的合规和安全需要，提供物理隔离的基础设施，还可提供支付、结算、风控、审计等业务组件，需要通过数据管理相关技术提供数据的安全保障。

同时 5G 技术也将成为支撑云计算的关键技术，同时，由于 5G 大幅的增加了数据传输的带宽，提高了数据传输速度，并大大降低了数据的延迟，势必也将为数据管理注入新鲜的技术。

5 总结

网络技术、分布式计算以及虚拟化等技术催生了云计算模式，而面对云计算中多样且大量的数据，涌现出一大批高效处理各类数据的方法，涵盖了云计算中数据获取、数据存储、数据分析的方方面面，较好的解决了云计算发展过程中的一些问题，而海量数据管理时的计算需求由没有边际和上限困扰的云计算来承载。因此数据管理与云计算对彼此的需求正盛，为未来留下一片澄澈的蓝海。

本文首先介绍了云计算的基本概念，并从数据获取、数据存储、数据分析三个方面分析了数据管理在云计算领域的应用，在数据获取阶段，讨论了数据冗余缩减和数据去重技术；在数据存储阶段，分析了 GFS 分布式文件系统，并以键值型数据库分析了 NoSQL 数据存储，其使用哈希技术对大量数据进行分块，实现存储负载均衡，在数据分析方面，详细介绍了媒体数据的处理分析方法。最后对云计算和数据管理的前景进行了探讨和展望。

媒体数据管理大作业
Only For Preview

参考文献：

[1] 智研咨询. 2020-2026 年中国企业 IT 产业运营现状及发展前景分析报告[EB/OL].(2020-4-29)

[2] 刘鹏.云计算[M].北京:电子工业出版社,2010

[3] Zhang Y, Callan J, Minka T. Novelty and redundancy detection in adaptive filtering[C]//Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 2002: 81-88.

[4] Salomon D. Data compression: the complete reference[M]. Springer Science & Business Media, 2004.

[5] Symes P D. Digital video compression[M]. McGraw Hill Professional, 2004.

[6] Tsai T H, Lin C Y. Exploring contextual redundancy in improving object-based video coding for video sensor networks surveillance[J]. IEEE transactions on multimedia, 2011, 14(3): 669-682.

[7] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002: 269-278.

[8] Huang Z, Shen H, Liu J, et al. Effective data co-reduction for multimedia similarity search[C]//Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. 2011: 1021-1032.

[9] Kamath U, Compton J, Islamaj-Dogan R, et al. An evolutionary algorithm approach for feature generation from sequence data and its application to DNA splice site prediction. IEEE/ACM Trans Comput Biol Bioinform, 2012, 9: 1387–1398

[10] Günter M. Introducing MapLan to map banking survey data into a time series database[C]//Proceedings of the 15th International Conference on Extending Database Technology. 2012: 528-533.

[11] Wang Y, Wei D, Yin X, et al. Heterogeneity-aware data regeneration in distributed storage systems[C]//IEEE INFOCOM 2014-IEEE Conference on Computer Communications. IEEE, 2014: 1878-1886.

[12] Ghemawat S, Gobioff H, Leung S T. The Google file system[C]//Proceedings of the nineteenth ACM symposium on Operating systems principles. 2003: 29-43.

[13] Chaiken R, Jenkins B, Larson P Å, et al. SCOPE: easy and efficient parallel processing of massive data sets[J]. Proceedings of the VLDB Endowment, 2008, 1(2): 1265-1276.

媒体数据管理大作业
Only For Preview

[14] Beaver D, Kumar S, Li H C, et al. Finding a Needle in Haystack: Facebooks Photo Storage[C]//OSDI. 2010, 10(2010): 1-8.

[15] Hastorun D, Jampani M, Kakulapati G, et al. Dynamo: Amazon's highly available key-value store[C]//In Proc. SOSP. 2007.

[16] Karger D, Lehman E, Leighton T, et al. Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web[C]//Proceedings of the twenty-ninth annual ACM symposium on Theory of computing. 1997: 654-663.

[17] Ding D, Metze F, Rawat S, et al. Beyond audio and video retrieval: towards multimedia summarization[C]//Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. 2012: 1-8.

[18] Wang M, Ni B, Hua X S, et al. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration[J]. ACM Computing Surveys (CSUR), 2012, 44(4): 1-24.

[19] Shao L, Jones S, Li X. Efficient search and localization of human actions in video databases[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 24(3): 504-512.

[20] Chen L, Xu D, Tsang I W H, et al. Spectral embedded hashing for scalable image retrieval[J]. IEEE Transactions on Cybernetics, 2013, 44(7): 1180-1190.

[21] Song J, Yang Y, Li X, et al. Robust hashing with local models for approximate similarity search[J]. IEEE transactions on cybernetics, 2014, 44(7): 1225-1236.

媒体数据管理大作业
Only For Preview