大数据分析研究现状、问题与对策

谭 艳

（广州杰赛科技股份有限公司，广东 广州 510310）

摘 要：大数据的发展速度不断加快，国内外越来越重视大数据的发展，大数据领域中的关键问题是如何科学有效地分析大数据。通过研究可知当前在大数据分析研究过程中还存在一些问题，对其发展造成了一定的影响。所以笔者主要对大数据分析研究现状、问题和对策进行了具体的研究与分析。

关键词：大数据；存储问题；云存储技术

中图分类号：G201 **文献标识码：**A **文章编号：**1003-9767（2017）19-143-02

Status Quo, Problems and Countermeasures of Big Data Analysis

Tan Yan

(Guangzhou Jiesai Science & Technology Co., Ltd., Guangzhou Guangdong 510310, China)

Abstract: The development of big data continues to accelerate, and more and more attention is paid to the development of big data at home and abroad. The key problem in the big data field is how to analyze large data scientifically and effectively. Through the study, we can see that there are still some problems in the process of large data analysis and research, which have a certain impact on its development. Therefore, the author makes a detailed analysis and research on the status quo, problems and countermeasures of the analysis and research of big data.

Key words: big data; storage problem; cloud storage technology

随着 IT 技术的发展速度不断加快，各个领域处理数据的压力不断加大，只依靠人力已无法满足计算分析大量信息的需求。在这样的情况下，美国奥巴马政府发布了《大数据研究和发展倡议》，旨在利用大量复杂数据集合获取知识并提升洞见能力，持续强化其在数据资源领域的优势[1]。该倡议的提出使各个国家政府对大数据的研究和分析力度不断加大，但是在研究和分析中还存在一些问题。所以具体研究大数据分析研究现状、问题与对策具有重要的现实意义。

1 大数据分析研究的现状

1.1 研究大数据的方法

大数据的分析方法是当前大数据分析中主要的研究内容。大数据的分析结果在一定程度上受到分析方法的影响，而且不同的分析方法适用于不同的数据类型。复杂数据的识别技术与传统文本和关系数据的识别技术相比，具有一定的差异性，这使数据分析的难度增加。当前 XML 数据、图数据和复杂网络上实体识别技术原理是复杂数据识别技术所使用的[2]。从大数据系统构架角度来说，在数据分析中主要使用的是九层架构的方式，但是需要进一步研究其具体的应用。

1.2 大数据分析驱动科学萌芽

信息科学技术的发展受到了大数据技术的严重影响，大数据技术改变了许多产业的发展方式，如在社会媒体当中，传统媒体的受众分析和传播策略的研究被大数据所改变。另外，在大数据驱动的背景下也改变了客户生命周期理论。而且大数据技术在发展的过程中也在一定程度上影响着其他行业，在这样的情况下，各个行业必须要与大数据的发展保持一致，不断创新科学技术，从而使其发展具有突破性。

2 大数据面临的问题

2.1 存储问题

数据的存储问题是大数据分析过程中首要解决的问题。从数据量级方面来说，大数据时代数据量由过去的 TB 级达到了 PB 级和 EB 级，这对存储和分析数据具有积极的推动

作者简介：谭艳（1986-），女，重庆人，本科。研究方向：英语。

作用[3]。在分析数据的过程中，会不止一次地存取和调度数据，这样数据的存储就会变成动态的，其增、减和删改等操作会根据数据生命周期的变化和实际的需求产生动态的变化。

从数据存储机构方面来说，不能利用简单传统的机构化数据库存储大量的数据，所以当前人们主要研究与大数据特点相符合的存储方式。此外，另一个需要注意的问题是在大数据动态更新的状况下，怎样保障数据储存和交换过程中的一致性。在分析大数据中数据库领域需要实现数据仓库的可扩展性、向下兼容性和高度容错性等。由此可见，大数据分析的核心问题主要是数据的存储问题。

2.2 可用性较弱的问题

在实际中数据呈现的方式是杂乱无章的，这加大了大数据分析的难度，所以在分析大数据时需要考虑的主要是数据质量因素。数据质量的概念十分宽泛，在这里主要研究数据的可用性问题。数据的一致性、时效性、准确性、实体性和完整性五部分共同组成了数据的可用性。

但是在处理之前，存储和分析大数据的首要条件主要是数据可用性的度量。在采集数据时要研究高效处理过滤数据的方法，从而获取到质量较高的大数据源[4]。从大数据完整性方面来看，要描述和评价数据，需要不断完善数据描述的框架，这对数据采集中的数据描述具有一定的指导作用。从数据的时效性和一致性方面来说，要注重一些客观的数据和时间价值，在大数据的采集中，客观事实与数据的描述必须相符合，而大量级数据背景下的不可或缺的环节主要是数据源的自动检测和修复。这些都是大数据分析研究中的难点问题。

3 解决大数据面临问题的措施

3.1 合理部署云存储技术

目前大数据的发展速度是非常惊人的，处理数据的效率和存储数据的成本都受到大数据存储方式的直接影响[5]。云存储向用户提供的存储服务是以互联网为基础的，这样存储的容量和数据的可用性等各个复杂底层技术的细节不需要用户考虑，其只需要付费就可以从云存储供应商那里获取到无限大的存储空间和企业级的服务质量。

云计算环境下分布式存储的主要基础是数据中心，数据存储中心的划分可以从不同的角度进行。从系统建设方面来说，云储存中心架构的形态主要有三种，即优化的传统数据中心、以云计算为数据中心和两者并存，这是由于历史遗留的存储信息系统的原因造成的。

在提供存储服务的过程中，存储用户和云计算用户是主要的两种云存储用户，这是根据用户服务内容的不同划分的[6]。云存储服务的关键主要是以云计算理论构建的数据中心，如图 1 所示，调度分割并行编程模型下，存储结构化数据和非结构化数据的目标是通过并行数据库和分布式系统来完成的，而云服务接口价格计算资源服务提供给云用户的基础主要是云服务等级协议。

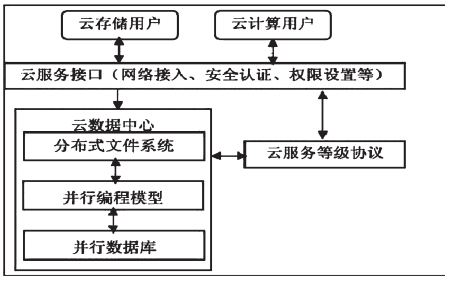

图 1 大数据云存储模型图

3.2 提升数据的可用性

大数据分析的对象主要是大量复杂的数据，其具有不同的产生方式，所以其会涉及各种不同的信息系统等。在分析任何一个大数据项目中，都需要搜集数据，为分析提供保障，分析过程是比较简单的，而重点环节主要是数据的预分析。这里所说的数据可用性主要是解决数据预分析问题。

Web 数据、传感网数据、业务系统数据和科学实验数据是当前主要的大数据的来源，而每一种数据源会有效预处理不同的信息系统和信息结构数据，这是根据不同类型的数据或相同类型的数据进行的。当前获取和整合高质量大数据的理论和技术等五个挑战性的研究问题已被提出来，而其将数据处理可用性领域的问题从各个方面进行了分析，如基础理论和工程技术等，同时探索了大数据可用性的理论和技术。另外，只有为数据的质量提供保障，才能更好进行大数据分析，所以对大数据时代数据质量的保障从流程和管理等方面进行了研究。

4 结 语

随着云计算领域的软件发展速度的不断加快，社会上越来越重视大数据分析，从而逐渐形成分析服务。但是当前大数据分析研究中还存在一些问题，如可用性较弱和存储问题等，所以需要提升其可用性，同时要合理部署云计算，通过这些措施可以更好把握大数据发展的趋势。

参考文献

[1] 官思发，朝乐门．大数据时代信息分析的关键问题、挑战与对策 [J]. 图书情报工作，2015,59(1):12-18,34.

[2] 官思发，孟玺，李宗洁，等．大数据分析研究现状、问题与对策 [J]. 情报杂志，2015,34(3):98-104.

[3] 赵润身．大数据分析研究现状、问题与对策 [A]//2016 智能城市与信息化建设国际学术交流研讨会论文集 III[C].2016:1.

[4] 李炎生．大数据时代我国政府对网络信息监管的问题研究 [D]. 长春：吉林大学，2016,31(6):88-89.

[5] 彭文波．基于大数据分析的专利代理质量研究 [D]. 湘潭：湘潭大学，2016,15(8):45-46.

[6] 于施洋，王建冬，童楠楠．大数据环境下的政府信息服务创新：研究现状与发展对策 [J]. 电子政务，2016,25(10):26-32.