大数据应用的关键技术研究

顾加强 刘 锦

（江西经济管理干部学院信息工程系，江西 南昌 330088）

摘要：本文分析了大数据的基本概念与特点，并对支撑大数据应用的大数据采集与集成技术、大数据存储技术、大数据分析及数据挖掘技术、大数据安全技术等关键技术进行了介绍。

关键词：大数据；数据获取；数据存储；数据分析与挖掘；应用

中图分类号：TP391 **文献标识码**：A **文章编号**：1671-4792(2017)7-0056-04

Research on Key Technology of Big Data Application

Gu Jiaqiang Liu Jin

(Department of Information Engineering, Jiangxi Institute of Economic Administrators, Jiangxi Nanchang 330088)

Abstract: This paper analyzes the basic concepts and characteristics of big data and introduces the key technologies of big data collection and integration, big data storage, big data analysis and data mining, and big data security that support big data applications.

Keywords: Big Data; Data Collection; Data Storage; Data Analysis and Integration; Application

1 大数据的概念与特点

数据是用于对客观事物进行记录和描述的可识别的符号，人们使用数据对客观事物的特征、状态以及它们之间的相互关系进行记录。

在计算机系统和数据库管理系统出现之后，人们对数据的分析和使用，大体上经历了四个阶段。第一阶段是：数据电子化阶段，这个阶段主要是将原来纸质材料上的数据录入到计算机系统，从而形成电子化的数据；第二个阶段是：数据处理流程电子化阶段，这一阶段引入了ERP、OA、CRM等理念，实现了工作流程优化，提高了工作效率；第三个阶段是：互联网数据阶段，数据依赖互联网络解决了人与人之间电子化沟通，出现了电子商务等新的数据应用；第四个阶段是：大数据阶段，这一阶段随着“互联网+”的概念提出，标志着数据开始随着互联网渗透到各个行业，大数据与“智慧”密切联系到了一起。

“大数据”的概念从2009年被提出，目前还没有严谨的定义。互联网促进了大数据的产生和发展。大数据的意义，除了让人们掌握庞大的数据信息之外，更包含着对数据进行专业化处理之后产生的附加价值。大数据相对于传统的数据而言具有以下特点：大数据应用过程中的数据量非常巨大(Volume)，大数据一般使用PB(1PB=1024TB)作为容量单位，目前有一些大型互联网公司的数据容量甚至可以采用EB(1EB=1024PB)作为容量单位；数据的类型多样化(Variety)，大数据应用时，除了传统的结构化数据之外，还有数据内容和结构混在一起的半结构化的数据，如XML或HTML，以及大量的无法直接描述出来的声音、视频、图片、地理位置等非结构化的数据；数据的处理速度快(Velocity)，这主要是因为大数据增长的速度很快，需要在很短的时间内对数据进行实时分析，为及时获取有价值的信息，大数据处

理与传统的数据挖掘有着本质的区别;数据的潜在商业价值高(Value),对大数据进行合理的分析和处理,将会带来巨大的商业回报。

互联网在使用过程中产生了海量的大数据,移动互联网的普及让大数据具备更大的商业价值,大数据的价值与它的数据类型、数据容量都有密切的关系,类型越复杂、数据量越大所蕴含的价值越高,但是如果没有合适的工具和方法对数据进行及时分析和处理,随着时间的推移,数据的价值会慢慢流失。大数据资源归集到一定阶段之后,大数据资源的开放利用成为可能,当然大数据的开放使用是有条件的,有些数据是可以向公众开放的,任何人都可以使用,有些数据的使用是有条件的,还有些数据是保密的,这对大数据的应用提出新的、更高的要求。

2 大数据采集与集成技术

传统的数据采集来源单一,数据量也相对较小,追求数据的高度一致性和容错性。在大数据时代,海量大数据中的绝大部分是由互联网或物联网络设备产生,数据的采集通常与数据的使用价值是分离的,即时大数据采集是后续大数据应用的前提。数据采集可以借助相关工具和手段对结构化、半结构化或非结构化的海量数据进行自动化、智能化的识别、转换和初步处理。

目前,进行大数据采集时主要可以采用下列方法:(1)通过使用系统日志进行数据采集的方法。目前已经有很多知名互联网公司开发了不同类型的基于系统日志的大数据采集工具。比如基于 Hadoop 的大数据采集 Chukwa 工具,Facebook 公司开发的大数据采集工具 Scribe 等。这一类大数据采集工具一般都采用分布式体系架构,在自动进行数据采集时,可以达到数百兆字节/每秒的数据采集和传输速度。(2)利用网络爬虫或 API 进行大数据采集。这种方法可以通过使用网络爬虫工具或利用相关互联网企业在数据资料共享平台上公开提供的应用程序接口(API)等方式,来获取各类网站上的相关数据信息。利用这种方法,可以从网页或网站中把各类需要的信息抽取出来,然后在本地按统一格式存储为特定的数据文件。(3)其他大数据采集方式。对于一些保密性要求较高的行业或内部数据,如:企业生产经营过程中的数据、学科研究过程中的数据等,可以通过与相关企业或研究单位进行合作,使用由相关数据产生企业或研究单位提供的特定系统接口功能进行采集数据。无论我们通过什么方式来进行大数据采集,都应该能够提供一种机制,让我们可以从多角度来验证数据的全面性和可信性。

大数据集成是将分散获得的数据进行集中,再进行统一管理的过程,数据集成是大数据应用活动中最复杂的活动之一。大数据集成的本质就是将数据由一种格式转换为另外一种统一的格式,为后续的大数据应用做好准备。目前大数据集成的方法主要有:第一种方法是 ETL 法,该方法在进行数据集成的时候按照数据抽取、数据格式转换、数据装载的步骤(Extract,Transform,and Load,简称 ETL)进行;第二种方法与第一种方法的步骤顺序有些不同,它是 ELT 方法,这个方法是先进行数据抽取,然后进行数据装载,最后进行数据格式转换(Extract,Load and Transform,ELT);第三种方法被称为变化数据抓取的方法(Change Data Capture,简称 CDC),这种方法在实现的时候,通过在相关机器上安装某个用于进行数据采集与集成的第三方应用程序来实现自动收集数据,并自动转换格式等数据集成操作。

3 大数据存储技术

大数据应用要考虑数据存储的问题,对于大数据应用的单位而言,数据包括单位内部因业务而产生的业务数据,还包括通过互联网、物联网或其他途径获取的海量数据。在数据存储时要解决下列问题:重复数据处理问题,数据文件管理和保护问题,数据的高性能共享的问题。消除数据冗余是提高数据存储效率、降低存储开销的重要手段。

传统的数据中心方式,无论是在性能上、效率上、安全性上,还是在投资收益上都不能满足大数据存储的要求。找到一种更好、更快,能够更多存储数

据的存储技术是大数据应用面临的另一个挑战。目前，实现数据存储主要有以下几种方法：

（1）基于嵌入式架构的存储系统方法。这种方法由嵌入式硬件和固化在硬件平台中的软件组成，通常应用于对实时性要求不高的简单场合，对于更复杂的应用场合，可以给嵌入式架构系统配上嵌入式实时操作系统。

（2）基于 X86 架构的存储系统方法。可以采用 DAS（直接附加存储设备）方法，DAS 存储是通过 SCSI 接口将存储设备连接到一台服务器上使用；也可以采用 NAS(网络附加存储服务器)方法，NAS 设备是一种带存储设备的瘦服务器，NAS 直接连接到网络上，其他主机通过网络使用数据；还可以采用 SAN(专用存储区域网络)方法，SAN 为了实现数据存储而专门建立的专用网络。三种方法中性能最好的是 SAN 方法，但这种方法价格昂贵且异地扩展比较困难。

（3）基于云存储和云计算技术方法。云存储与云计算系统是近几年发展起来的一个非常复杂的存储与计算系统，它不但可以实现云存储还能实现云计算，同时解决数据存储与数据计算两大问题。该系统一般由部署在云端的网络服务器、网络存储设备、网络管理软件、应用程序接口、客户端为特定应用开发的应用程序等部件构成，在整个云计算与云存储系统中，以云计算设备和云存储设备为系统核心，通过各应用层相关软件为不同用户提供大数据存储与应用相关服务，企业或个人利用云计算与云存储系统，能够在花费比较小的情况下，实现非常高的系统安全性与可靠性，是目前解决大数据计算与存储问题的一种性价比较高的方案，目前得到广泛应用。

4 大数据分析及数据挖掘技术

数据不经过分析是没有使用价值的，只有经过分析之后变为特定的格式的信息，才能在应用时发挥数据潜在的价值。在大数据使用之前，要对大数据进行分析和数据挖掘，这一活动的目的是从看起来没有什么关联的大量的、复杂的数据中，通过技术手段快速的找出这些数据之间的存在潜在关联。对大数据进行分析的处理流程与传统数据进行分析处理流程基本相同，但由于大数据要处理的数据的结构十分复杂（有结构化数据也有非结构人数据）、数据量非常大，而且对数据处理时效性要求还非常高，所以在进行大数据分析和数据挖掘时，有新的、更高的要求。

对大数据进行分析和挖掘，本质就是从获得的海量数据中，按照某种方法或策略，采用相关技术手段来提取这些数据里所包含的、我们不能直接发现的、但又有应用价值的信息的过程和方法。在进行大数据分析与数据挖掘时，可以采用不同的技术手段，这些技术又分为大数据描述性技术和大数据分析预测性技术。大数据描述性技术主要是利用技术手段分析海量数据与数据之间存在的相关规律的技术；而大数据分析预测性技术主要是利用技术手段对使用历史数据进行分析，从而预测这一行业或领域未来情况的技术。

大数据分析与数据挖掘技术在整个大数据应用过程中十分重要，它实现功能主要有：(1)未来发展的趋势预测功能。这一功能主要是通过对数据本身进行特定分析，然后对数据规律进行总结，继而通过总结来预测未来的发展趋势，也就是我们常说的所谓“智慧”；(2)大数据的总结功能。这一功能主要是利用数学上经常使用的统计学方法来实现，如：在对大数据进行分析时可以用“求方差”或者求“标准差”等方法，来对大量数据进行统计与分析；(3)数据的聚类与分类功能。主是把海量数据按不同类别分解成不同的子集，方便用户理解；(4)数据关联分析与数据偏差检测功能。主要用于发现事物之间的关联性，找出极端的特殊例子。

数据分析与挖掘技术的本质是知识发现的过程，通常可以分为 6 个子过程，分别为定义数据分析与挖掘目标、进行数据取样、进行数据探索、进行数据预处理、模式发现、模式评价。在数据分析与挖掘

前需要对数据分析与挖掘目标进行定义；数据取样是利用样本数据对总体目标进行评价估计的重要方法；数据探索的目的是发现数据中复杂关系；数据预处理是保证数据质量的重要手段；模式发现是利用关联分析、神经网络分析等分析技术对事物或现象进行描述、识别、分类和解释的过程；模型评价是在建立的各种模型中找出大数据分析过程中解决实际问题的所适用的模型。在数据分析和挖掘过程中要用到各种算法，不同的算法有不同的作用和适用场合。云计算技术可以为大数据分析与挖掘提供稳定可靠的算力，已经成为大数据分析及数据挖掘技术重要支撑技术。

5 大数据安全技术

大数据的开放性与数据的安全性是一对相辅相成又相互矛盾的应用需求，数据对外开发肯定会带来安全方面的问题，没有安全保障是不能实现好开放的大数据平台的。在大数据应用过程中，需要使用相关安全性技术和手段来对大数据应用进行保障。可以通过中立区技术隔离核心区与其他区域，通过物理方式隔离内网与外网，通过防火墙隔离共享层与应用接口层。

在大数据安全性方面有两个关键点的控制非常重要，其一是敏感数据的多重防护，敏感数据是指相关隐私数据，其二是大数据生命周期安全管理，这贯穿于整个大数据应用过程，从源数据到应用层及应用系统整个过程。两个方面的安全性需求都可以采用如：4A 认证、数字水印、数据加密、统一日志管理等技术来实现，可以做到事前可管理、事中有监控、事后可追查。

6 结束语

信息化的普及产生了大量的数据，利用相关技术可以把这些数据收集起来，通信网络高速发展、云计算、云存储等技术的成熟使得数据能够被及时存储和处理，互联网、物联网的普及使得数据获取和处理的成本越来越低，让大数据的开发和利用成为可能。

大数据平台的出现，使得公众和商家使用大数据成果变得越来越简单。越来越多企业已经意识到打造开放的大数平台的重要性。阿里巴巴、百度、腾讯等公司都建立了各自的大数据应用平台。阿里巴巴在很早就已经确定了以云计算和大数据为中心的大数据应用战略，2013 年阿里巴巴公司的开发的“御膳房”数据引擎业务已经开始探索将其在运行过程中收集的海量数据转化为生产资料、商品、商家、客服、品牌等主题数据可以通过 API 接口的方式向各类用户开放，淘宝的 API 平台，目前已经成为能够整合上下游产业链的、多方合作共赢的大数据生态圈，第三方应用程序开发者利用淘宝 API 接口和相关开发环境，可以直接访问存储于阿里巴巴的云服务器上的各种数据，使大数据的应用变得更简单、更直观。

参考文献

[1]王秀磊，等.大数据关键技术[J].中兴通信技术，2013，(04).

[2]杨巨龙，等.大数据据技术全解[M].北京：电子工业出版社，2014.

[3]段云峰，等.大数据的互联网思维[M].北京：电子工业出版社，2015.

[4] 韩晶.大数据服务若干关技术研究[D].北京：北京邮电大学，2013.

[5]黄宏程，等.大数据之美[M].北京：电子工业出版社，2016.

[6]程学旗，等.大数据系统和分析技术综述[J].软件学报，2014，(09).

作者简介

顾加强(1975—)，男，江西瑞昌人，工程硕士，副教授，主要研究方向：网络及网络安全、电子商务；

刘锦(1979—)，女，江西都昌人，工程硕士，副教授，主要研究方向：软件开发与电子商务应用。