教育部人文社科项目
深圳市科创委技术攻关项目

大数据分布式并行处理技术

——基于天云星数据库的交通管理大数据处理

向怀坤　陈晓攀
熊志强　刘义宗　著

西安电子科技大学出版社
http://www.xduph.com

教育部人文社科项目
深圳市科创委技术攻关项目

大数据分布式并行处理技术

——基于天云星数据库的交通管理大数据处理

向怀坤　陈晓攀　熊志强　刘义宗　著

西安电子科技大学出版社

内 容 简 介

本书立足于当前公安交通管理领域利用 Hadoop 技术在处理非互联网行业大数据时存在的低效问题，基于天云星数据库(SCSDB)对结构化大数据分布式并行处理技术进行了介绍。全书共 7 章，主要内容包括概论、天云星数据库基础、数据库对象管理、SCSDB 安全管理、SCSDB 备份与还原、数据库监控与调优、数据导入与导出。在介绍理论知识的同时，本书在文中还穿插了公安交通管理大数据处理应用案例。

本书适用于高校计算机科学与技术、交通信息工程及控制、智能交通技术等专业，也可供大数据、软件工程、人工智能等领域的专业技术人员参考。

图书在版编目(CIP)数据

大数据分布式并行处理技术：基于天云星数据库的交通管理大数据处理 / 向怀坤等著. —西安：西安电子科技大学出版社，2018.7

ISBN 978-7-5606-4961-0

Ⅰ. ① 大… Ⅱ. ① 向… Ⅲ. ① 统计数据—分布式数据—处理 Ⅳ. ① O212

中国版本图书馆 CIP 数据核字(2018)第 145300 号

策划编辑　李惠萍

责任编辑　秦媛媛　阎　彬

出版发行　西安电子科技大学出版社(西安市太白南路 2 号)

电　　话　(029)88242885　88201467　　邮　　编　710071

网　　址　www.xduph.com　　电子邮箱　xdupfxb001@163.com

经　　销　新华书店

印刷单位　陕西华沐印刷科技有限责任公司

版　　次　2018 年 7 月第 1 版　2018 年 7 月第 1 次印刷

开　　本　787 毫米×1092 毫米　1/16　印　张　15

字　　数　351 千字

印　　数　1～2000 册

定　　价　36.00 元

ISBN 978-7-5606-4961-0/O

XDUP 5263001-1

前　言

伴随着以互联网、即时通信与智能终端等为代表的新一代信息技术的飞速发展及广泛应用，各行各业累积的数据开始爆炸式增长，在此背景下诞生了大数据(Big Data)概念。随着大数据概念从提出到落地，大数据产业即以一日千里的速度向前发展。全球多家权威机构统计，大数据产业正在迎来黄金发展时期。据互联网数据中心(Internet Data Center，IDC)预计，大数据和分析市场将从2016年的1300亿美元增长到2020年的2030亿美元以上；中国报告大厅发布的大数据行业报告表明，自2017年起，我国大数据产业迎来飞速发展，未来2～3年的市场规模增长率将保持在35%左右。

在各个大数据细分领域中，公安交通管理行业所生成的交通大数据占据了重要地位。我国于20世纪90年代开始大力发展城市智能交通系统，从国家到地方高度重视，到目前为止基本建立起覆盖道路、轨道、水运等交通运输方式在内的多模式智能交通系统(Intelligent Transportation System，ITS)。以城市道路交通管理与控制为例，我国绝大部分城市已经建立了较完善的视频监控、交通检测、信号控制、交通诱导、车辆导航等智能化交通管控系统。这些系统每天将产生超过拍字节(PB)级的交通大数据，如何对这些交通大数据进行“加工”处理，从中挖掘出有用的“知识”，为诸如路况预测、风险规避、交通救援、事故鉴定等业务应用提供“增值”服务，是我国目前交通管理大数据亟待研究解决的重要课题。

与其他行业如电子商务、互联网页等产生的大数据相比，公安交通管理行业不仅包含海量的非结构化数据(如交通图像、违法视频等)，还包含海量的结构化数据。因此，针对公安交通管理大数据的数据采集、数据存储、数据挖掘与数据分析等大数据技术开发，需要进一步结合行业应用展开。针对Hadoop技术在处理非互联网行业(如政府、企业等)大数据时存在的低效问题，本书基于天云星数据库(SCSDB)在城市道路公安交通管理结构化大数据处理中的实战案例，重点对城市道路交通警务结构化大数据处理技术进行了分析论述。

全书分为7章。第1章为概论，主要包括大数据发展概况、大数据技术架构、大数据关键技术以及公安交通管理大数据概述；第2章介绍了天云星数据库基础，主要包括天云星数据库概述、天云星数据库安装、天云星数据库运维和管理；第3章介绍了数据库对象管理，包括数据库对象的命名规则、数据库管理、数据表管理、索引管理、视图管理和序列号管理；第4章介绍了SCSDB安全管理，主要包括SCSDB账户管理、SCSDB权限管理和数据库审计；第5章介绍了SCSDB备份与还原，主要包括SCSDB实时备份机制和SCSDB冷备份；第6章介绍了数据库监控与调优，主要包括系统监控、数据库监控、数据库调优以及公安交通大数据应用案例；第7章介绍了数据导入与导出，主要包括使用SOURCE命令导入数据、使用重定向功能导出数据、使用LOAD DATA命令导入数据、使用易镜进行数据的导入和导出、使用SYNCD进行数据同步以及使用Kettle进行数据抽取。为方便读者阅读，本书最后提供了两个附录文件，一个是SCSDB的数据类型，另一个是公安交通警务大数据案例表结构。

本书由深圳职业技术学院向怀坤博士主持编写，其中第1～4章、第6章由向怀坤、陈晓攀完成；第5章、第7章、附录B由熊志强完成；附录A由刘义宗完成。另外，梁嘉、王峰、黄秀、毛立洁参与了全书插图、表格、数据、案例等的编辑整理工作。

大数据分析技术还在不断发展之中，书内参考了近年来该领域公开发表的大量文献，在此对相关著作者表示诚挚的谢意。由于作者理论水平和实践经验有限，本书内容难免存在欠缺与疏漏，恳请广大读者批评指正。

著　者

2018年3月于深圳

目　录

第1章 概 论

1.1 大数据发展概况

伴随着计算机技术、移动互联网技术、人工智能技术等的飞速发展，各行各业都开始走向与互联网融合的道路，特别是各类应用软件、社交网站、行业网站等不断地融入人们的生活，导致各类数据呈现爆炸式增长趋势，人类因此进入到大数据时代。

大数据的快速发展可追溯到2000年前后。当时互联网网页量快速增长，据不完全统计，每天新增约700万个网页，到2000年年底，全球网页数达到40亿。如此多的网页，让用户检索信息越来越不方便，为此谷歌(Google)、亚马逊(Amazon)等公司率先建立了覆盖数十亿网页的索引库，开始提供较为精确的搜索服务，从而大大提升了人们使用互联网的效率，这是大数据应用的起点。针对当时网页数据不仅数量庞大，而且以非结构化为主，传统的搜索引擎和处理技术难以处理的问题，Google公司率先提出了一套以分布式为特征的全新技术体系，即后来陆续公开的分布式文件系统(Google File System，GFS)、分布式并行计算(MapReduce)和分布式数据库(Big Table)等技术，这些技术奠定了当前大数据技术的基础。

伴随互联网产业的崛起，这种创新的海量数据处理技术在电子商务、定向广告、智能推送、网络社交等领域得到了全面应用，取得了巨大的商业成功。这一现象启发了全社会开始重新审视数据的巨大价值，于是金融、电信等拥有大量数据的行业开始尝试这种新的理念和技术，并取得了初步成效。与此同时，业界也在不断地对Google公司的技术体系进行改进和扩展，使之能适用于更多的场景。2011年，麦肯锡、世界经济论坛等知名机构对这种数据驱动的创新进行了研究总结，随即在全世界掀起了一股大数据热潮。

进入2012年，大数据(Big Data)一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。国际数据中心(IDC)在2012年的数据统计中表明，非结构化数据约占互联网数据总量的75%。IDC官方的《数字宇宙》(Digital Universe)的研究报告预测，到2020年，全世界新建和复制的数据量将会超过40 ZB，增长至2012年的12倍。与此同时，中国的数据量将会在2020年接近9 ZB，相比2012年将增长近22倍。

大数据概念第一次被创造出来是在2008年9月4日Google公司成立10周年之际。此后不久，《自然(Nature)》期刊推出了大数据专辑，包括了8篇大数据专题文章和1篇编者按。虽然大数据已成为全社会的热议话题，但到目前为止，对大数据仍无一个统一的定义。麦肯锡对大数据的定义是：大数据是指那些规模大到传统的数据库软件工具已经无法采集、存储、管理和分析的数据集。研究机构Gartner认为：大数据是指需要新处理模式才能具有

更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。一般来说，大数据是具有体量大、结构多样、时效性强等特征的数据；处理大数据需要采用新型计算架构和智能算法等新技术；大数据的应用强调以新的理念应用于辅助决策、发现新知识，同时更强调在线闭环的业务流程优化。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台计算机进行处理，必须采用分布式架构。大数据的特色在于对海量数据进行分布式数据挖掘，但必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术等。

1.2 大数据技术架构

随着大数据的容量和其复杂性的快速增长，对原有的 IT 架构及其计算与处理都提出了新的挑战。2003 年，Google 的三篇论文奠定了大数据技术的发展基础，经过多年的发展，在大数据处理方面已经诞生了大量分布式数据处理技术和分布式处理框架，从最开始的 Hadoop 分布式系统到后来的内存处理系统 Spark，数据实时处理系统 Storm、Flink 等，这些分布式技术及其生态圈的发展已经形成了一套完整的大数据解决方案和应用架构。

由于大数据来源于互联网、企业系统和物联网等信息系统，因此从总体上来看大数据技术应用需求是通过对大数据处理系统的分析挖掘来产生新的知识，从而支撑行业企业决策或业务的自动智能化运转的。由此可知，不同行业、不同业务其大数据技术构架的设计会有所不同。下面分别从大数据准备、大数据存储、大数据计算、大数据分析和大数据展示这五个层面来简述大数据处理系统的技术构架，如图 1-1 所示。

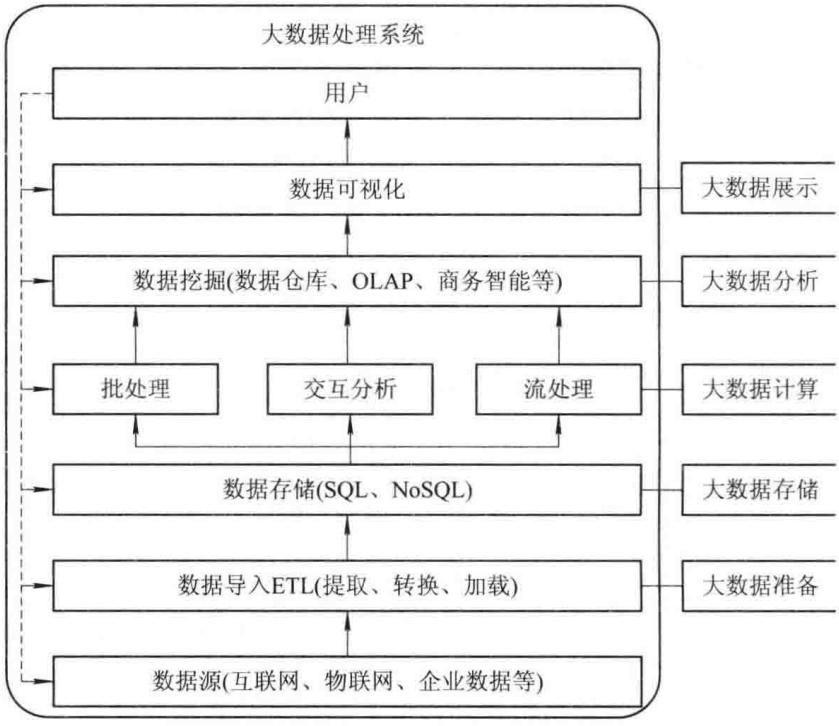

图 1-1 大数据处理技术框架

大数据准备是大数据处理系统的基础部分。大数据的来源有很多渠道，因不同行业应用而不同，如互联网、物联网、电子商务、智慧交通等。在获得海量数据进行存储和处理之前，需要对这些数据进行清洗、整理，如提取、转换和加载等，传统数据处理体系中称为 ETL(Extracting, Transforming, Loading)过程。与以往数据分析相比，大数据的来源多种多样，包括企业内部数据库、互联网数据和物联网数据，不仅数据体量庞大、格式不一，质量也良莠不齐。这就要求进行大数据准备，一方面要规范数据格式，便于后续存储管理，另一方面要在尽可能保留原有语义的情况下去粗取精、消除噪声。

在完成大数据导入即大数据准备后，要求采用高效的数据存储方案对海量数据进行存储管理。当前全球数据量正在以每年超过 50%的速度增长，存储技术的成本和性能都面临非常大的压力。大数据存储系统不仅需要以极低的成本存储海量数据，还要适应多样化的非结构化数据管理需求，具备数据格式上的可扩展性。随着大数据技术的日趋完善，各大公司及开源社区陆续发布了一系列新型数据库来解决海量数据的组织、存储及管理，如 HBase、Spark、Redis、MemcacheDB、Storm 和 SCSDB(天云星数据库)。

大数据计算部分是在大数据存储管理的基础上，重点解决数据计算分析所需要的算法、处理速度与计算资源分配等问题，需要根据处理的数据类型和分析目标，采用适当的算法模型，如批处理、交互分析、流处理等，快速处理各类数据。海量数据处理将消耗大量的计算资源，对于传统的单机系统或并行计算技术而言，速度、可扩展性和成本都难以适应大数据计算分析的新需求，在此背景下分布式并行计算成为大数据的主流计算架构，但在某些特定的场景下，分布式并行计算的实时性和计算效率还需要大幅度提升。

大数据分析是指从纷繁复杂的数据中发现规律，提取新的知识，也就是数据挖掘的主要工作内容，这是实现大数据价值的关键环节。传统的数据挖掘对象多是结构化、单一对象的小数据集，挖掘时更侧重根据先验知识预先通过人工手段建立数学模型，然后依据既定的模型对数据进行分析挖掘。对于非结构化、多源异构的大数据集而言，往往缺乏先验知识，因而很难建立显式数据模型，这就需要研究开发更加智能化的数据挖掘方法。

经过大数据分析后，需要结合具体的行业应用，通过适当的形式展示大数据分析结果。从大数据服务于决策支撑的场景来看，以直观的方式将分析结果呈现给用户，是大数据分析的重要环节。如何让复杂的分析结果易于理解，是这一部分工作面临的主要挑战。在嵌入多业务中的闭环大数据应用中，一般是由机器根据算法直接应用分析结果，无需人工直接干预，这种场景下大数据分析结果的展现则不是必需的。

1.3 大数据关键技术

1.3.1 大数据存储管理技术

面对数据海量化和快速的增长需求，要求大数据存储管理系统的底层硬件架构和文件系统的性价比必须大大高于传统技术，存储容量应可以无限制扩展，且要求有很强的容错能力和并发读写能力。

传统的网络附着存储系统(NAS)和存储区域网络(SAN)等体系，其存储和计算的物理设备分离，相互之间要通过网络接口连接，这导致在进行数据密集型计算时，I/O 容易成为瓶颈。同时，传统的单机文件系统(如 NTFS)和网络文件系统(NFS)要求一个文件系统的数据必须存储在一台物理机器上，且不提供数据冗余性，其可扩展性、容错能力和并发读写能力难以满足大数据需求。

谷歌文件系统(Google File System，GFS)和 Hadoop 的分布式文件系统 HDFS(Hadoop Distributed File System)奠定了大数据存储技术的基础。与传统存储系统相比，GFS/HDFS 将计算和存储节点在物理上结合在一起，从而避免在数据密集计算中形成 I/O 吞吐量的制约，同时这类分布式存储系统的文件系统也采用了分布式架构，能达到较高的并发访问能力，如图 1-2 所示。

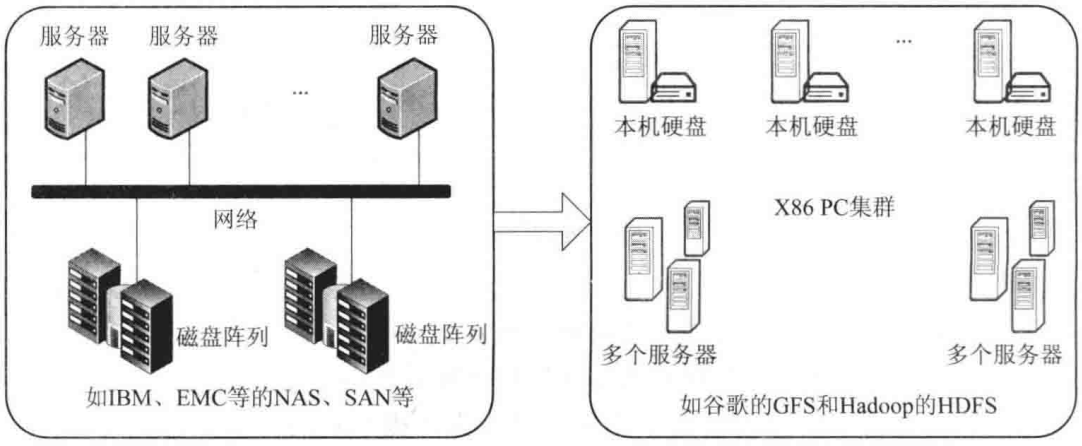

图 1-2　大数据存储架构的变化

随着大数据时代应用范围的不断扩展，GFS 和 HDFS 也面临诸多瓶颈。虽然 GFS 和 HDFS 在大文件的追加(Append)写入和读取时能够获得很高的性能，但对于随机访问(Random Access)和海量小文件的频繁写入等需求而言，其工作性能较低，这导致其在某些应用领域很难得到推广。业界当前和下一步的研究重点主要是在硬件上进行功能拓展，特别是研究开发基于 SSD 等新型存储介质的存储体系架构，同时对现有分布式存储文件系统进行改进，以提高随机并发访问和海量小文件存取等操作的性能。

大数据对存储技术的另一个挑战是处理多样化数据格式的适应能力。格式多样化是大数据的主要特征之一，为此需要大数据存储管理系统能够适应各种结构化和非结构化数据，并对各种类型的数据进行高效处理。数据库的一致性(Consistency)、可用性(Availability)和分区容错性(Partition-Tolerance)不可能都达到最佳，在设计存储管理系统时，需要在这三个方面做出权衡。传统的关系型数据库管理系统以支持事务处理为主，采用了结构化数据表管理方式，满足了一致性要求而牺牲了可用性。为大数据设计的非关系型数据库(NoSQL，即 Not only SQL)如 Google 的 BigTable 和 Hadoop HBase 等通过使用“键-值”(Key-Value)对、文件等非二维表结构，具有很好的包容性，适应了非结构化数据多样化的特点。同时，这类 NoSQL 数据库主要面向分析型业务，但其一致性要求则有所降低。整体来看，未来大数据的存储管理技术将进一步把关系型数据库的操作便捷性和非关系型数据库的灵活性结合起来，研发新的融合型数据存储管理技术。

1.3.2 大数据并行计算技术

大数据的分析挖掘属于一种数据密集型计算模式，需要强大的计算能力。与传统的“数据简单、算法复杂”的高性能计算不同，大数据的计算对计算单元和存储单元间的数据吞吐率要求极高，对性价比和扩展性要求也非常高。传统的依赖大型机和小型机的并行计算系统不仅成本高，数据吞吐量也难以满足大数据计算要求，同时靠提升单机 CPU 的性能、增加内存、扩展磁盘等来实现性能提升的纵向扩展(Scale Up)的方式也难以支撑平滑扩容。

Google 公司在 2004 年公开的 MapReduce 分布式并行计算技术，是新型分布式计算技术的代表。一个 MapReduce 系统由廉价的通用服务器构成，通过添加服务器节点可线性扩展系统的总处理能力(Scale Out)，在成本和可扩展性上都有巨大优势。MapReduce 是 Google 公司内部网页索引、广告等核心系统的基础。之后出现的 Apache Hadoop 是 MapReduce 的开源实现，已经成为目前应用最广泛的大数据计算软件平台。MapReduce 架构能够满足“先存储后处理”的离线批量计算(Batch Processing)需求，但也存在局限性，最大的问题是时延过大，难以适用于机器学习迭代、流处理等实时计算任务，也不适合针对大规模图数据等特定数据结构的快速计算。

针对上述问题，业界在 MapReduce 基础上，提出了多种不同的并行计算技术路线。如 Yahoo 提出的 S4 系统、Twitter 的 Storm 系统是针对“边到达边计算”的实时流计算(Real Time Streaming Process)框架，可在一个时间窗口上对数据流进行在线实时分析，已经在实时广告、微博等系统中得到应用。Google 公司于 2010 年公布的 Dremel 系统，是一种交互分析(Interactive Analysis)引擎，几秒钟就可完成 PB($1PB = 2^{50}B$)级数据查询操作。此外，还出现了将 MapReduce 内存化以提高实时性的 Spark 框架、针对大规模图数据进行了优化的 Pregel 系统等。

对于不同计算场景建立和维护不同计算平台的做法，硬件资源难以复用，管理运维也很不方便，研发适合多种计算模型的通用架构成为业界的普遍诉求。为此，Apache Hadoop 社区在 2013 年 10 月发布的 Hadoop 2.0 中推出了新一代的 MapReduce 架构，见图 1-3。新架构的主要变化是将旧版本 MapReduce 中的资源管理和任务调度功能分离，形成一层与任务无关的资源管理层(YARN)。YARN 对下负责物理资源的统一管理，对上可支持批处理、流处理、图计算等不同模型，为统一大数据平台的建立提供了新平台。基于新的统一资源管理层开发适应特定应用的计算模型，仍将是未来大数据计算技术发展的重点。

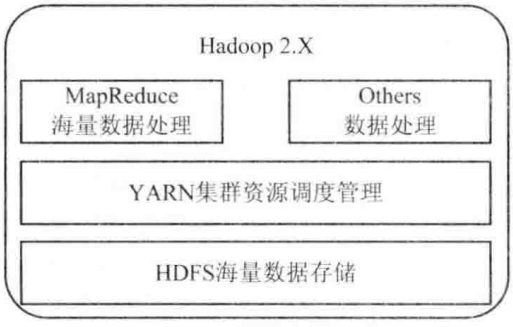

图 1-3 Hadoop 2.X 核心模块

1.3.3 大数据查询和分析技术

大数据查询和分析技术的发展需要在两个方面取得突破：其一是对体量庞大的结构化和半结构化数据进行高效率的深度分析，挖掘隐性知识，如从自然语言构成的文本网页中理解和识别语义、情感、意图等；其二是对非结构化数据进行分析，将海量复杂多源的语音、图像和视频等数据转化为机器可识别的、具有明确语义的信息，进而从中提取出有用

的知识。在人类全部数字化数据中，仅有非常小的一部分(约占总数据量的1%)数值型数据得到了深入分析和挖掘(如回归、分类、聚类)，大型互联网企业对网页索引、社交数据等半结构化数据进行了浅层分析(如排序)，对占总量近 60%的语音、图片、视频等非结构化数据还难以进行有效的分析。

目前的大数据查询和分析主要有两条技术路线：一是凭借先验知识，通过人工建立数学模型来查询和分析数据；二是通过建立人工智能系统，使用大量样本数据进行训练，让机器代替人工获得从数据中提取知识的能力。由于占大数据主要部分的非结构化数据往往模式不明且多变化，因此难以靠人工建立数学模型去进行数据挖掘。通过人工智能和机器学习技术来查询和分析大数据，被业界认为具有很好的发展前景。2006 年 Google 等公司的科学家根据人脑认知过程的分层特性，提出增加人工神经网络层数和神经元节点数量，加大机器学习的规模，构建深度神经网络，可提高训练和识别效果。这一观点在后续试验中得到了证实，最有名的当数 2016 年、2017 年的“人机围棋大战”事件，即 Google 公司的人工智能围棋软件 AlphaGo 轻松战胜世界围棋名将李世石、柯洁等人，体现了深度神经网络的威力。这一事件引起了工业界和学术界的高度关注，使得神经网络技术重新成为数据分析技术的热点。目前，基于深度神经网络的机器学习技术已经在语音识别和图像识别方面取得了很好的效果，但未来深度学习要在大数据分析上广泛应用，还有大量理论和工程问题需要解决，主要包括模型迁移适应能力，以及超大规模神经网络的工程实现等。

1.3.4 大数据可视化技术

大数据可视化技术是指利用计算机图形学及图像处理技术，将数据转换为图形或图像形式显示到屏幕上，并进行交互处理的理论和方法的统称。它涉及计算机视觉、图像处理、计算机辅助设计、计算机图形学等多个领域，成为一项研究数据表示、数据处理、决策分析等问题的综合技术。为实现信息的有效传达，数据可视化应兼顾美学功能，直观地传达出关键的特征，便于挖掘数据背后隐藏的价值。

大规模数据的可视化和绘制主要是基于并行算法设计的技术，合理利用有限的计算资源，高效地处理和分析特定的数据集的特性。很多情况下，大规模数据可视化的技术通常会结合多分辨率表示等方法，以获得足够的互动性能。在面向大规模数据的并行可视化工作中，主要涉及以下四种基本技术。

(1) 数据流线化(Data Streaming)：是将大数据分为相互独立的子块后依次进行处理，其中离核渲染(Out-of-Core Rendering)是数据流线化的一种重要形式，在数据规模远大于计算资源时这是一类主要的可视化手段。它能够处理任意大规模的数据，同时也能提供更有效的缓存使用效率，并减少内存交换，但通常这类方法需要较长的处理时间，不能提供对数据的交互挖掘。

(2) 任务并行化(Task Parallelism)：是指把多个独立的任务模块进行平行处理。该方法要求将一个算法分解为多个独立的子任务，并需要相应的多重计算资源。其并行程度主要受限于算法的可分解粒度以及计算资源中节点的数目。

(3) 管道并行化(Pipeline Parallelism)：是指能同时处理各自面向不同数据子块的多个独立的任务模块。对于任务并行化和管道并行化两类方法，如何达到负载的平衡是其关键点。

(4) 数据并行化(Data Parallelism)：是将数据分块后进行平行处理，通常称为单程序多

数据流模式。该方法能达到高度的平行化，并且在计算节点增加时可以达到较好的可扩展性。对于超大规模的并行可视化处理，节点之间的通信效率往往是重要的制约因素，实践表明，提高数据的本地性可以大大提高系统的运行效率。

以上这些技术在实践中往往相互结合，从而构建出一个更高效的解决方案。虽然数据可视化日益受到关注，可视化技术也日益成熟。然而，当前大数据可视化仍存在许多问题，且面临着巨大的挑战。数据可视化面临的挑战主要指可视化分析过程中数据的呈现方式，包括可视化技术和信息可视化显示。目前，数据简约可视化研究中，高清晰显示、大屏幕显示、高可扩展数据投影、维度降解等技术都试着从不同角度解决这些难题，在可预见的未来，大数据的可视化问题仍会是一个重要的挑战。

1.4 公安交通管理大数据概述

1.4.1 交通行业信息化发展历程

自1975年成立交通部(现为交通运输部，下同)计算机应用研究所至今，交通信息化的发展建设已历时40多年。交通运输行业为适应交通运输发展的需要，对交通信息化发展进行了不断的探索，我国交通信息化经历了从无到有、从有到精、由点到面的发展历程，到目前为止初步建成了日趋完善的交通信息化体系，可分为三个不同特征的发展阶段。

1．单机应用阶段

我国于1989年出台《交通运输经济信息系统(TEIS)——“八五”发展计划》，这被看成是我国交通运输信息化起步的标志。同年交通部计算机应用研究所更名为中国交通信息中心，被赋予其行业信息化管理职能，统筹我国交通行业的信息化建设和发展。20世纪70年代，北京、上海、广州等大城市开始了交通信号控制的研究与开发，单点定周期交通信号控制器和配套的车辆检测、网络通信等设备得到了快速发展。到20世纪80年代后期，我国开始尝试交通智能化管理的基础研究工作，尝试在交通信息采集、驾驶员考试培训、车辆动态识别等领域应用信息技术。

2．部门应用阶段

20世纪90年代前期，我国开始密切关注国际智能交通系统(Intelligent Transportation System，ITS)的发展，于“九五”期间，交通部制定了《公路、水运交通运输信息化“九五”规划和2010年远景目标》，开展交通运输信息网络(CTInet)建设。“十五”期间，无论公路、水路运输还是城市交通均得到了有力的信息化建设支持。随着ITS理念被正式引入我国，1999年经科技部批准，国家智能交通系统工程技术研究中心(National Center of ITS Engineering & Technology, ITSC)(以下简称“国家ITS中心”)正式成立。《公路水路交通信息化“十五”发展规划》提出围绕政府办公、行业监管、现代物流三大领域开展信息化建设。这段时期，我国交通运输行业基础信息网络基本完善，各部门业务应用逐步覆盖。2000年，科技部会同国家计委、经贸委、公安部、铁道部、交通部、建设部、信息产业部等几十个部、委、局联合建立了“全国智能运输系统协调领导小组”及办公室，并成立了ITS专家咨询委员会。2001年，在科技部和交通部的支持下，国家ITS中心完成了“中国国家

ITS 体系框架研究”和“国家 ITS 标准体系研究”等课题。城市交通方面，2002 年 4 月，科技部正式批复“十五”国家科技攻关“智能交通系统关键技术开发和示范工程”重大项目，北京、上海、天津等 10 个城市作为首批试点城市，以城市、城际道路运输为主要实施对象，开展了交通管理与控制系统、智能公交调度、综合交通信息平台等领域的研究与应用示范。

3. 整合应用阶段

随着《公路水路交通运输信息化“十一五”发展规划》(2006 年)的实施，全国开始推行省级公路信息资源整合和服务试点工程。国家 863 计划设立了“现代交通技术领域”，并针对 ITS 技术部署了一批前沿和前瞻性项目，在智能化交通控制技术，交通信息采集、处理及服务技术，车辆运行状态监控与安全预警等领域取得了实质性进展。借 2008 年北京奥运会、2010 年上海世博会、2010 年广州亚运会等大型活动举办之机，科技部于 2006 年启动实施了“国家综合智能交通技术集成应用示范”科技计划项目，交通智能化管理与动态诱导技术、跨区域联网不停车收费技术、远洋船舶及货物运输在线监控等关键技术得到了重点突破，北京、上海、广州、深圳等一线城市在城市公交、交通管理、出行服务等方面均开展了深入的建设，积累了许多实践经验，为城市客货运输提供了更加优质的服务。

《公路水路交通运输信息化“十二五”发展规划》(2011 年)提出了坚持资源共享和业务协同的发展理念；住房和城乡建设部在全国 193 个城市开展“智慧城市”试点工作。2013 年交通运输部杨传堂部长发表《加快推进科技创新为“四个交通”建设提供坚实支撑》的讲话，正式提出了“智慧交通”的发展理念。这一时期，我国交通运输行业信息化建设全面推进，部省联动、共建共享得到加强，资源开发利用水平大幅提升，公众服务水平明显提高。2016 年交通运输部印发《交通运输信息化“十三五”发展规划》，提出紧扣国家战略、结合行业实际、加强顶层设计、深化行业整体应用、依托政企合作和打造服务新生态的发展理念。随着“互联网+”、大数据上升为国家战略，“十三五”期间我国交通运输行政改革正在全面深化，综合运输发展也将全面转型，将着重解决行业基础信息碎片化问题、行业应用整体性问题和行业信息推进策略与保障机制问题，将突出行业基础信息的集聚、共享和开放，形成行业大数据能力，突出应用的综合性、整体性和协同性，提升交通运输综合治理能力，最终形成政府、市场、公众共同参与，多方共赢的交通运输信息化治理体系。

1.4.2 公安交通管理大数据现状

随着我国经济的飞速发展和城市化进程的加速，人、车、路的矛盾日益突出，交通拥堵、交通事故频发等问题早已从一线城市蔓延至二、三线城市，仅凭传统的“人海战术、人力作业、人工运转”的管理模式已无法解决当前交通管理存在的问题。

近几年，是公安交通管理信息化大跃进、信息大爆炸的时代，公安交通管理信息化经历了从无到有，从信息孤岛到大集中、优整合、高共享的建设高潮。2011 年，交通运输部交管局在全国推广应用了“公安交通管理综合应用平台”，统一了业务系统，规范了业务流程，是公安交通管理信息化建设中里程碑式的发展。

为持续实现“科技强警，向科技要警力、向科技要战斗力”，各地公安交通管理部门不断加大交通管理信息化建设力度，各类传感器、高清卡口和信息终端已遍布整个城市。这