目录

[摘要 2](#_Toc8115)

[一．介绍 3](#_Toc21299)

[二．背景 3](#_Toc17929)

[2.1大数据 3](#_Toc20920)

[2.2 Hadoop和MapReduce技术 5](#_Toc7562)

[2.3 Hadoop在交通大数据中的应用场景及其优势 5](#_Toc23331)

[2.4深度学习 6](#_Toc13846)

[三．Hadoop的优化 8](#_Toc17719)

[3.1 分布式文件系统 9](#_Toc26201)

[3.2分布式计算框架 10](#_Toc19475)

[四．本文提出的方案 11](#_Toc24950)

[4.1伪码及流程图 11](#_Toc10187)

[4.2 TVD-MRDL系统设计 15](#_Toc29379)

[五．实验和结果 16](#_Toc25508)

[5.1性能 18](#_Toc22674)

[5.1.1单机模式下顺序程序与Hadoop的性能比较 18](#_Toc11333)

[5.2可扩展性 19](#_Toc15905)

[5.2.1一个主节点和多个从节点的Hadoop集群的效率和可扩展性比较 19](#_Toc13407)

[5.2.2数据量的可扩展性 20](#_Toc5667)

[5.2.3 TVD-MRDL与传统系统的可扩展性比较 21](#_Toc2962)

[5.3 Hadoop集群成本 21](#_Toc25218)

[5.3.1 TVD-MRDL与顺序方案的资源成本比较 21](#_Toc24998)

[5.3.2 Hadoop从节点扩容时间 23](#_Toc6991)

[5.4 TVD-MRDL与传统系统在数据量方面的正确率、精密度和召回率的比较 23](#_Toc16456)

[5.5关于Hadoop节点数量增加的通信开销 24](#_Toc13808)

[5.6使用深度学习的Hadoop集群的效率和可扩展性的比较 25](#_Toc22506)

[六．综述与结论 28](#_Toc5908)

[七．参考文献 31](#_Toc8319)

# 摘要

Hadoop作为一个开源的云计算平台，以其高可靠性、高扩展性以及在处理和分析海量数据方面的巨大优势而被广泛应用于各个领域。在交通领域，利用Hadoop大数据平台进行数据分析和挖掘可以帮助提取有价值的见解。为了总结Hadoop在交通大数据应用方面的最新研究进展，我们对2012年至今发表的相关文章进行了全面回顾。首先，通过使用VOSviewer软件进行文献计量分析，我们识别了关键词的演变趋势。其次，介绍了Hadoop的核心组件，以及其在交通大数据领域的应用。随后，系统性地识别了最新的研究进展，并对Hadoop及其优化框架在交通领域的主要应用场景进行了分类。根据我们的分析，确定了该领域的研究差距和未来的工作方向。回顾现有研究表明，Hadoop在过去十年中在交通大数据研究中扮演了重要角色。重点关注的领域包括交通基础设施监控、出租车运营管理、出行特征分析、交通流量预测、交通大数据分析平台、交通事件监控与状态判别、车牌识别和最短路径等。此外，对Hadoop的优化框架进行了深入研究，主要包括Hadoop计算模型的优化以及Hadoop与Spark结合的优化。在交通大数据领域取得了多项研究成果，但对Hadoop核心技术的系统研究仍较为有限，Hadoop与交通大数据融合发展的广度和深度还有待提高。未来的建议包括将Hadoop与处理实时数据源的大数据框架如Storm、Flink等相结合，以提高交通大数据的实时处理和分析能力。同时，多源异构交通大数据的研究仍然是一个重点。改进现有的大数据技术，实现交通大数据的分析甚至数据压缩，将为智能交通领域带来新的突破。

本文提出了TVD-MRDL系统，其作为一种驾驶员行为检测系统，旨在加快对多数据集和大数据集的处理时间。它的设计是在Hadoop框架中通过分布式技术检测大规模数据中的交通违规者的危险行为。通过性能分析，TVD-MRDL技术可以通过使用更多的从节点来减少处理时间。在这里，我们提出了一个MapReduce空间，它能够利用分布式架构和广泛的数据来分析交通控制中心的数据，并检测驾驶员的违规行为(不安全行为)。

基于分布式体系结构和MapReduce编程模型，TVD-MRDL系统可以实现交通违规的自动控制。在此基础上，提出了一种基于MapReduce深度学习的大规模驾驶员违规检测方法。事实上，违规检测的分布式学习算法需要大量的时间，并且随着接受域的增加而线性增加。提供了这个条件，所提出的方法似乎适合于大数据学习。一般来说，

TVD-MRDL系统的目的是大大减少错误，提高速度，保持维护成本非常低，所有这些都使其成为交通控制中心检测驾驶员危险行为的合适技术。

论文的其余部分组织如下:在第2节中，介绍了理解所提出的方法所需的背景知识。在第3节中，回顾了在不同研究领域中使用MapReduce技术的研究并阐述了TVD-MRDL系统的总体过程以及所使用的MapReduce技术和深度学习的过程，并描述了每个阶段的组成和功能。在第4节中，对所提出的方法的实施所做的调整以及在违规行为案例中使用该方法的结果进行了检查，最后，在第5节中，根据研究结果得出了总体结论。

# 一．介绍

驾驶员作为普通人，在日常生活中可能因为各种不规范行为，从而导致不同程度的损失。也就是说，驾驶员的冒险举动可能会带来意想不到的事故，进而造成无法挽回的损失。通常情况下，潜在风险与可能导致损害的不可控因素或活动相关。辨识这些风险因素并消除或控制它们，有助于避免损失。在交通领域的事故发生后，人员伤亡的成本和社会成本一般都十分高昂，并且造成该类事故和人员伤亡的责任主要归咎于驾驶员的问题；因此，如何改善并规范驾驶员的操作行为被认为是交通领域最重要、最需解决的问题之一。为了规范司机的交通违规行为，交通警察会出现在入口处，对违规车辆进行罚款。在这种方法中，由于人(警察)是通过观察来控制违章行为的，因此不可避免地会出现错误的发生和误用，包括串通、违章行为不被警察注意、警察缺席、入口处交通拥堵等。

近年来，由于街道和道路上车辆数量的增加，交通管制问题和交通法规的执行受到了更多的关注。在一些国家，已经提出了一些解决方案和系统来自动控制城市交通。此外，在交通流量大的路口和街道上安装不同的摄像头，并提供有效的系统来自动检测驾驶员的危险行为也是必不可少的。考虑到这个问题的重要性，这方面的任何研究都可以对社会做出巨大贡献。所以，我们在本文中，提出了一种通用的系统来跟踪对象和识别事件，并且在本文中，提出了一种用于对象跟踪的监控系统，并使二者合理配合从而达到效果。

由于视频监控摄像机产生的数据速度快、量大、种类多，因此其提供的数据可以被视为大数据。大数据是指数据集由于过于庞大或复杂而无法同时高效地管理、控制和处理的数据集。Hadoop是一种流行的大数据处理技术。Hadoop和MapReduce技术被用于各种领域，如医学、气象、行为检测、数据库，以及用于大数据分析的各种基于MapReduce的机器学习算法的实现。这种方法也可以用于分析来自城市监控摄像头的交通数据。MapReduce是一种用于大规模数据处理的编程模型，它通过分布式数据处理来减少处理时间。因此，本文提出了一种基于MapReduce的交通违章检测深度学习方法(TVD-MRDL)。

# 背景

## 2.1大数据

大数据分析帮助我们获得了新的方法和技术，用于在可扩展的基础上收集、存储和分析非结构化数据。大数据可以从工业、农业、交通运输、医疗保健、公共组织、家庭等方面获得。在过去的20年里，不同领域的数据量有了明显的增长。根据国际数据研究所的数据，2011年全球创造的数据量为1.8ZB(1021字节)，到2020年，将增长到35ZB以上。Meta(目前名为Gartner)的分析师Daugh Lanie在一份研究报告中，通过3v模型(增加的体积-速度-品种)定义了数据增长所带来的机遇和挑战。大数据价值链可以分为四个阶段:数据生产、数据采集、数据存储和数据分析。大数据生产是指大数据的主要来源，包括通过不同技术在各个领域采集到的数据。大数据采集阶段包括数据采集、数据传输到存储基础设施、数据预处理等。在数据存储阶段，使用了各种机制和某些类型的编程模型(如MapReduce)。在大数据分析阶段，采用与大数据分析相关的方法。

交通运输部门越来越多地利用大数据，因为如果不能有效利用，大量数据可能会被浪费。大数据技术可用于从交通数据中挖掘知识并识别管理人类交通行为的一般规律。这有助于改善交通状况并解决各种交通问题。通过查阅相关文献，可以发现以下有趣的主题，以便更深入地了解使用 Hadoop 分析海量流量数据的研究。一些应用研究详细介绍了使用 Hadoop 大数据平台解决流量问题的过程。该过程涉及收集静态和动态数据，例如道路环境、车辆信息以及使用 GPS 和网络信号的车辆速度等实施数据。然后使用 MapReduce 处理收集到的数据，从而促进跨[数据交换](https://www.sciencedirect.com/topics/social-sciences/data-exchange" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关数据交换的更多信息)中心的数据交换。数据存储在Hadoop的HDFS上，以便高效集成和处理。基于 Hadoop 的控制中心将这些海量数据集可视化，以帮助解决复杂的流量问题。Hadoop 因其成熟性、可靠性以及以分布式和并行方式高效处理数据的能力而优于其他系统，为大数据分析提供了强大的架构。因此，利用Hadoop构建智能交通平台可以提高交通数据的价值和有用性。然而，对结构化和非结构化流量数据的研究并不平等。因此，利用Hadoop对非结构化数据进行分析和处理，可以增强交通数据的开发利用，拓展Hadoop在交通大数据分析中的应用场景。Hadoop 是一种广泛使用的大数据分析框架，可以更快、更稳定、更精确地存储和分析大型数据集。它在处理交通数据方面的有效性已得到广泛证明。学术研究表明，MapReduce 计算框架处理数据的速度比标准 Matlab 工具快约 40 倍。然而，仅靠速度并不足以确保交通问题的准确性或解决。算法、模型和系统对于解决各种流量挑战至关重要。因此，算法、模型、系统与Hadoop的全面集成对于建立交通大数据的基础技术至关重要。当前，交通与大数据相结合的新技术不断发展和完善，逐步形成核心技术。[这些核心技术大致可分为路网](https://www.sciencedirect.com/topics/engineering/road-network" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关道路网络的更多信息)大规模并行计算、路径[动态规划](https://www.sciencedirect.com/topics/social-sciences/dynamic-programming" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关动态规划的更多信息)与智能搜索技术、交通系统大数据分析与优化、运行控制与组织调度等四个领域。这些技术中的每一项都经过了广泛的研究，并且充满了模型和技术。然而，缺乏从整体技术角度整合现有研究成果的系统研究，现有研究大多集中在核心技术在特定场景下的应用。因此，开发一个以Hadoop为基本框架、以各种算法模型为关键组件、全面、健壮的交通大数据技术体系具有挑战性。需要持续的研究和开发，以确保核心技术能够产生更全面、更强大的系统。Hadoop与交通大数据融合的研究可以分为两个趋势：融合的广度和深度。集成广度是指Hadoop作为大数据分析平台的广泛使用，优势显着，但并不是唯一的，其他数据分析框架如Spark、Flink也被用来提高实用性。因此，各种大数据分析框架在交通大数据中的应用，体现了交通大数据需要不同程度应用不同的框架。随着大数据技术不断融入交通领域，未来交通大数据将融入更丰富的大数据技术来解决交通相关问题。然而，集成的深度却呈现出普遍性和准确性之间的矛盾。虽然普遍适用的算法或框架在针对不同的问题时可能会表现出局限性或其他矛盾，但解决特定问题的更精确的算法和框架可以为解决问题的方法带来新的突破。虽然大多数文献在Hadoop与交通大数据的融合中使用Java语言编写Map函数和Reduce函数来实现相应的算法功能，但针对大数据领域的具体问题改进Hadoop核心开发[源代码](https://www.sciencedirect.com/topics/engineering/source-coding" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关源代码的更多信息)已经有大量研究。这种以需求为导向的方式完善了Hadoop分布式架构，成为解决特定问题的精准算法和框架，为问题解决方式带来新的突破。需要注意的是，现有的很多算法在解决交通问题时并没有针对交通痛点，而仅仅停留在大数据技术应用的表面。他们没有深入探讨大数据技术的核心发展思路与交通问题解决思路的交集。导致Hadoop与交通大数据的融合还不够深入。针对这一问题，需要开展更多研究，将大数据技术更深入地融入交通运输，探索解决具体交通问题的精准算法和框架。

## 2.2 Hadoop和MapReduce技术

HADOOP 生态系统已变得流行用于分析大数据问题，其中一项数据无法由单个系统管理。Hadoop分布式模式可以通过分布式模式在各种系统上对数据进行分区管理。在这个特定的项目中，我选择以分布式模式执行MapReduce任务，以供人们将来使用。交易预订历史每天都会增加。几年后这将变得难以控制。因此，为了继续分析该技术的发展轨迹并为了技术的进步，我选择了 Hadoop 生态系统来寻找研究问题的答案。在Hadoop的一般信息结构中，信息由Hadoop系统分解并发送到多个服务器(节点)。服务器根据接收到的信息的类型对其进行处理或存储，既可以是处理信息，也可以是存储信息。当系统计划检索信息时，它从不同的服务器接收信息，进行组装，并将其显示为输出。Hadoop的优势在于它自动备份信息的能力。每条信息都存储在几个部分(服务器)中，如果其中一个服务器出现故障，另一个服务器可以承担责任并替换信息。MapReduce是一个软件框架，它为分布式应用的开发提供了一个安全和可扩展的条件。更准确地说，MapReduce包括任务的自动并行化、计算负载和数据平衡、磁盘和网络传输的优化以及机器缺陷的管理。Hadoop 允许我们在成千上万个节点上运行应用程序，这些节点上有成千上万的多媒体工具和应用TB 的数据运行应用程序。Hadoop 由两个主要部分组成，即 MapReduce 和 HDFS（Hadoop 分布式文件系统）。该系统实际上是为在多台服务器上运行而设计的，MapReduce和HDFS分别运行在主服务器和备服务器上，HDFS能够提高节点间的数据传输速率，使系统在一个节点故障的情况下仍能继续不间断运行。这种方法实际上降低了系统发生灾难性故障的风险，即使在大量节点停用的情况下也是如此。MapReduce是一个软件框架，其中的应用程序被分成更小的部分。这些部分(也称为节或块)中的每一个都可以在节点集群集中的每个节点上运行。总的来说，MapReduce是一个用于处理集群上数据的并行编程模型，它包括两个主要阶段，即映射阶段和减少阶段。图1展示了MapReduce模型的总体图。在Map阶段，输入数据被分成更小的部分。Map函数在数据的每一小部分(在从节点上)上运行。在对小部分的响应进行组织之后，在Reduce阶段，将它们组合起来形成所需的输出(通过Master节点)。

## 2.3 Hadoop在交通大数据中的应用场景及其优势

交通大数据的主要来源包括固定检测数据（如地磁线圈）、移动检测数据（如浮动汽车数据）、GPS检测数据和智能手机检测数据。利用Hadoop大数据技术对这些数据进行挖掘和分析，推动Hadoop大数据技术在交通行业的研究和应用。通过查阅文献，我们确定了当前该领域研究主要关注的八个主要应用场景。这些场景包括交通基础设施监控、出租车[运营管理](https://www.sciencedirect.com/topics/social-sciences/operations-management" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关运营管理的更多信息)、出行特征分析、交通流量预测、交通事件监控和状态判别、交通大数据分析平台、车牌识别、最短路径等。在交通领域中，Hadoop的应用场景丰富多样，其中包括交通流量分析：通过收集车辆GPS数据，可以实时监测道路上的车辆流量情况。利用Hadoop进行数据存储和分析，可以帮助交通管理部门优化信号灯控制、规划道路建设等，提升交通效率。路径规划优化：基于历史交通数据和实时路况信息，使用Hadoop进行数据挖掘和分析，可以为驾驶员提供最佳路径规划，避开拥堵路段，节约时间和燃料成本。交通事故预测：通过分析交通数据中的车辆速度、密度等信息，结合天气、道路情况等外部因素，可以利用Hadoop构建预测模型，提前预警可能发生的交通事故，有助于减少事故发生率。停车场管理：结合停车场实时车位信息和车辆进出记录，使用Hadoop进行数据分析和处理，可以实现停车位的智能分配、优化停车流程，提高停车场的利用率。公共交通优化：通过分析乘客乘车记录、公交车实时位置等数据，利用Hadoop进行实时调度和优化，可以提升公共交通系统的运行效率，减少乘客等待时间。这些都展示了Hadoop在交通领域中的多样化应用场景，通过大数据的存储、处理和分析，可以为交通管理、交通规划、交通安全等方面提供更加智能和高效的解决方案。

并且在交通大数据中，Hadoop的应用场景还具有以下优势。存储大规模数据：交通数据通常包含大量车辆轨迹、交通流量等信息，Hadoop的分布式文件系统（HDFS）可以有效存储这些海量数据，并支持高可靠性和高可用性。并行处理：Hadoop的MapReduce框架能够将数据分成小块并在集群中并行处理，加快数据处理速度，适用于交通数据分析中需要大量计算的场景。实时数据处理：结合实时数据处理框架（如Apache Storm、Apache Flink），Hadoop可以实现对交通数据的实时处理和分析，帮助监控交通状况、预测交通拥堵等。数据挖掘和模式识别：通过Hadoop的机器学习库（如Apache Mahout）和数据挖掘工具，可以对交通数据进行挖掘和分析，发现交通规律、优化交通管理策略等。可扩展性和弹性：Hadoop集群可以根据需要进行水平扩展，支持处理不断增长的交通数据量，同时具有自动容错和故障恢复机制，保障数据处理的稳定性。成本效益：Hadoop是开源软件，部署和维护成本相对较低，适合用于处理交通大数据的场景，可以降低数据处理和分析的成本。综上所述，Hadoop在交通大数据中的应用具有存储大规模数据、并行处理、实时数据处理、数据挖掘和模式识别、可扩展性和弹性、成本效益等优势，为交通领域的数据分析和应用提供了强大的支持。

## 2.4深度学习

深度学习是与机器学习相关的一个领域，它是一组算法，其中通过不同层次和层的学习对高阶抽象概念进行建模。深度学习实际上源于人类大脑的运作方式，它需要先进的工具，比如功能强大的显卡，来进行复杂的计算和大量的大数据。这种算法的低数据量会产生较弱的结果和性能;换句话说，深度学习涉及到对大量数据使用神经网络。这种学习方法最重要的优点是它的表征学习，这是一种自动从低阶输入中提取特征的方法。深度学习被广泛应用于包括计算机视觉在内的各个领域，例如图像分类、对象检测、语义分割、图像检索和人体姿势估计，这些都是理解图像的主要因素。

一般来说，深度学习方法根据其衍生的基础方法分为四种不同的类别:卷积神经网络、受限玻尔兹曼机(RBMS)、自动编码器和稀疏编码。对于这些类别中的每一个，都提出了特定的架构。例如，在cnn中，有Alexnet、SPP、VGGnet、Clarfai和Googlenet架构，它们在卷积层和全连接层的数量上通常是不同的。学习方法根据其特征而变化，例如泛化、无监督学习、特征学习、实时的训练、实时预测、生物理解、理论论证、不变性和小训练集。

深度学习是学习[样本数据](https://baike.baidu.com/item/%E6%A0%B7%E6%9C%AC%E6%95%B0%E6%8D%AE/12726279?fromModule=lemma_inlink" \t "_blank)的内在规律和表示层次，这些学习过程中获得的信息对诸如文字、图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。 深度学习是一个复杂的机器学习算法，在语音和图像识别方面取得的效果，远远超过先前相关技术。

深度学习在[搜索技术](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E6%8A%80%E6%9C%AF/1447197?fromModule=lemma_inlink" \t "_blank)、[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98/216477?fromModule=lemma_inlink" \t "_blank)、机器学习、[机器翻译](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E7%BF%BB%E8%AF%91/411793?fromModule=lemma_inlink" \t "_blank)、[自然语言处理](https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/365730?fromModule=lemma_inlink" \t "_blank)、[多媒体学习](https://baike.baidu.com/item/%E5%A4%9A%E5%AA%92%E4%BD%93%E5%AD%A6%E4%B9%A0/10528812?fromModule=lemma_inlink" \t "_blank)、语音、推荐和个性化技术，以及其他相关领域都取得了很多成果。深度学习使机器模仿视听和思考等人类的活动，解决了很多复杂的模式识别难题，使得人工智能相关技术取得了很大进步。

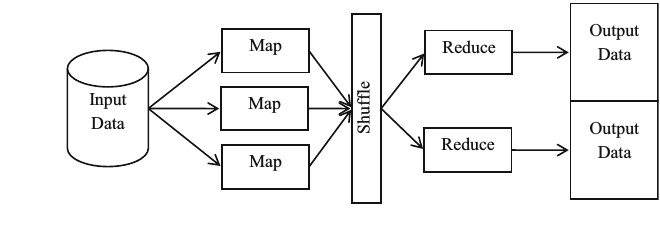


图1 MapReduce编程模型

# Hadoop的优化

Apache Lucene 的 Nutch 子项目创建了 Hadoop 分布式计算框架，这是一个用 Java 编写的开源解决方案。该框架利用集群的计算和存储能力，为大数据批量处理创建可靠、容错、可扩展和可扩展的架构。Hadoop的设计能够实现海量数据的分布式存储和并行处理，使其成为处理大规模数据处理需求的理想解决方案。

Hadoop的显着优势使其成为大[数据分析](https://www.sciencedirect.com/topics/social-sciences/data-analytics" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关数据分析的更多信息)的主流技术。许多公司和组织已将Hadoop 用于各种用例。例如，Facebook 利用 Hadoop 集群进行机器学习和数据分析，而 Yahoo! 将其用于广告和网络搜索研究。淘宝使用 Hadoop 系统处理和存储其电子商务交易数据，百度则使用 Hadoop 进行网络数据挖掘和搜索日志分析。此外，中国移动研究院还使用Hadoop“云”系统进行数据分析，并向公众提供计算服务。

Hadoop 是一种处理数据放置的数据处理工具。然而，其当前的数据放置方法主要侧重于平衡数据分布，并没有考虑不同数据集之间的关系。因此，所有 HDFS 数据都是根据 Hadoop 集群的工作负载需求来放置的。这可能会导致进行 MapReduce 计算时[传输](https://www.sciencedirect.com/topics/engineering/data-transfer" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关数据传输的更多信息)大量数据，从而导致更高的 I/O 开销。为了提高处理效率，出现了几种优化方法。其中一种方法是 CoHadoop，这是 IBM 开发的一种优化机制，可根据应用程序的需求分配数据块。然而，在将大数据提交到HDFS之前，CoHadoop需要根据应用程序的需求对数据进行分区，这会产生很大的处理成本。另一种流行的数据处理工具是 Spark，它是一种类似于 Hadoop MapReduce 的通用并行计算框架。Spark 也使用 MapReduce 算法，但有一些显着的差异。与 Hadoop MapReduce 不同，Spark 可以将作业的中间输出存储在内存中，从而无需读取和写入 HDFS。因此，Spark 特别适合需要迭代处理的数据挖掘和[机器学习算法](https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm" \o "从 ScienceDirect 的 AI 生成的主题页面了解有关机器学习算法的更多信息)。[表1](https://www.sciencedirect.com/science/article/pii/S2095756423001009?ref=pdf_download&fr=RR-2&rr=863263e72f1e5ddc" \l "tbl4)提供了Hadoop和Spark在容错、可扩展性、语言支持、可视化、实时分析、机器学习和SQL支持等各种特性方面的比较。这种比较突出了每个工具的不同应用范围。

这种比较突出了每个工具的不同应用范围。

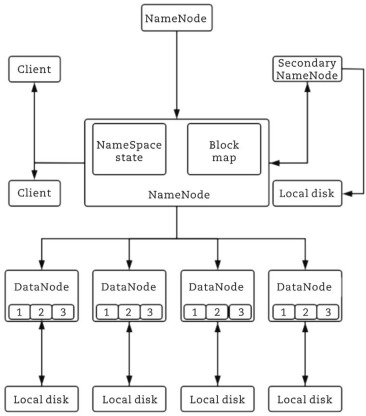
|  | **Hadoop** | Spark |
| --- | --- | --- |
| 容错能力 | 失败不需要重新启动应用程序。 | 无需额外代码或配置即可恢复丢失的工作。 |
| 扩展性 | 具有强大的可扩展潜力，已在数万个节点上用于生产。 | 高度可扩展，能够连续向集群添加 n 个节点。 |
| 语言支持 | 主要支持Java，其他还有C、C++、Ruby、Groovy、Perl、Python。 | 支持 Java、Scala、Python 和 R。 |
| 可视化 | 数据可视化是zoomdata直接连接HDFS以及SQL-on-Hadoop技术的能力。 | 通过 Web 界面提交和执行作业，或集成到 Apache Zeppelin 中。 |
| 实时分析 | MapReduce 无法处理实时数据，因为它旨在对大量数据执行批处理。 | 可以处理实时数据。 |
| SQL支持 | 用户可以使用 Apache Hive 运行 SQL 查询。 | 用户可以使用 Spark-SQL 运行 SQL 查询。 |
| 机器学习 | 需要像 Apache Mahout 这样的机器学习工具。 | 有一套机器学习MLlib。 |

表 1 . Hadoop和Spark之间的比较表。

关于 Hadoop 已经发表了大量研究，重点关注其计算模型优化以及 Hadoop 与 Spark 结合使用时的优化。本节将深入探讨这两个关键问题。

## 3.1 分布式文件系统

Hadoop 分布式文件系统 (HDFS) 是 Google 文件系统 (GFS) 的开源实现。它旨在提供高吞吐量数据访问，非常适合大规模存储和处理并行数据。[HDFS的基本结构如图2](https://www.sciencedirect.com/science/article/pii/S2095756423001009?ref=pdf_download&fr=RR-2&rr=863263e72f1e5ddc" \l "fig3)所示。它采用主从架构，HDFS集群由单个元数据节点（称为NameNode）、多个数据节点（称为DataNode）和一个辅助节点（称为辅助Namenode）组成。



**图2HDFS基本结构。**

在HDFS中，NameNode作为主节点，监督每个从节点的运行。具体来说，它负责跟踪命名空间的更改并管理命名空间本身。HDFS数据操作遵循“一次写入，多次读取”的模型。当一个文件存储在HDFS中时，它通常被分成多个64MB的数据块，每个数据块存储在一个单独的DataNode上。NameNode通过管理数据块并将其映射到相应的DataNode，在此过程中发挥着关键作用。

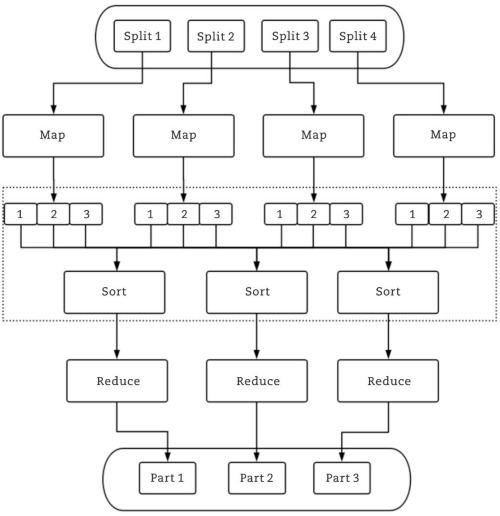
当客户端想要读取或写入文件时，它会与适当的 DataNode 通信以执行请求的操作。DataNode处理客户端的请求，并根据NameNode提供的指令执行创建或删除数据块等操作。要访问 HDFS 中的文件，客户端必须首先从 NameNode 检索文件中每个数据块的位置。然后它可以从适当的 DataNode 检索相应的数据。这两个步骤的过程可确保高效、准确地检索数据。

总体而言，HDFS 的分布式架构和数据处理机制使其成为在并行和分布式环境中管理大量数据的强大工具。

## 3.2分布式计算框架

MapReduce框架最初由Google于2004年提出，是一种分布式并行计算的编程模型。它能够处理大量数据，克服了传统计算方法的低效率。MapReduce 程序由两个主要阶段组成：Map 阶段和Reduce 阶段。

Map函数仅接受<key, value>格式的输入，Hadoop利用InputFormat()方法自动生成输入数据作为<key, value>对以供Map函数处理。键值表示数据切片中每条数据记录的字节偏移量，值表示每一行的内容。同样，Reduce函数也有<key, value>形式的输入和输出。它将 Map 函数的输出作为输入并对其进行操作。核心思想是使用Map过程将任务划分为更小、更易于管理的部分，然后通过Reduce过程组合结果。这种方法利用Hadoop数据平台服务器通过分布式计算并行处理海量数据，提高了处理大规模数据的效率。[图3](https://www.sciencedirect.com/science/article/pii/S2095756423001009?ref=pdf_download&fr=RR-2&rr=863263e72f1e5ddc" \l "fig4)描述了MapReduce程序的典型执行过程。



**图3 MapReduce 工作流程。**

MapReduce执行过程包括以下几个部分。

(1)将海量输入数据分成更小的部分，并将它们分配给不同的机器进行处理。

(2)Map任务工作者将输入数据解析为<key, value>对，并且用户定义的Map函数将它们转换为中间<key, value>对。

(3)根据中间 <key, value> 对的键值对其进行排序和聚合。

(4)将不同的键值及其对应的值集分发到各个机器上执行Reduce操作。

(5)生成Reduce 操作的输出。

# 四．本文提出的方案

## 4.1伪码及流程图

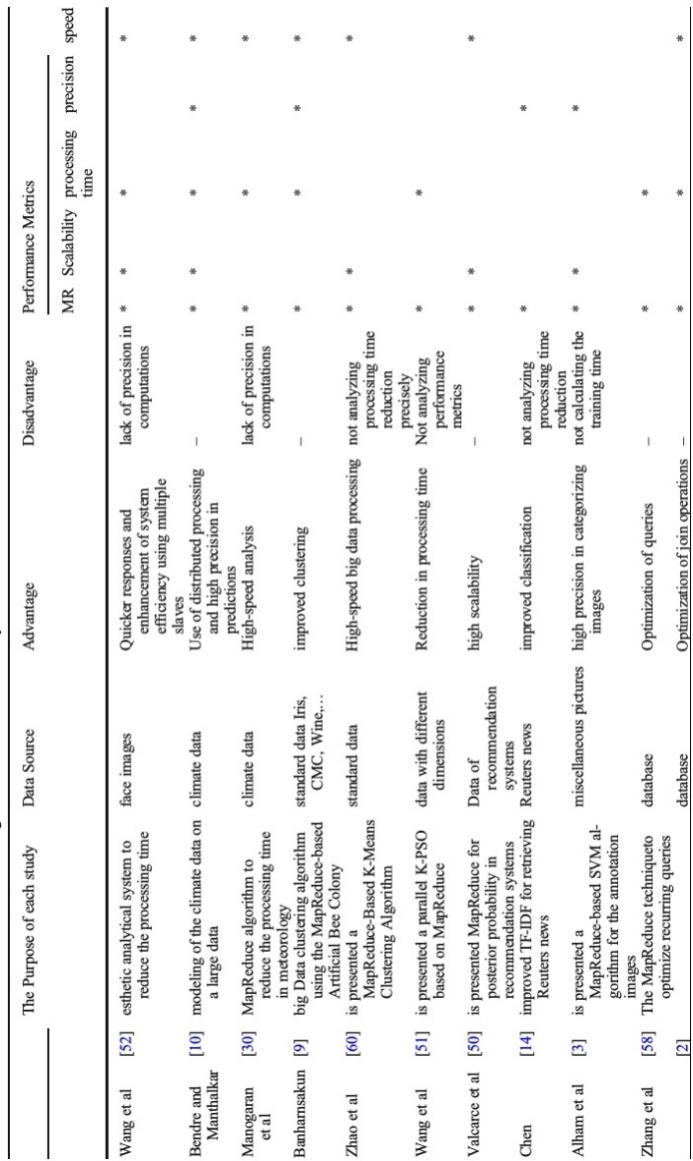
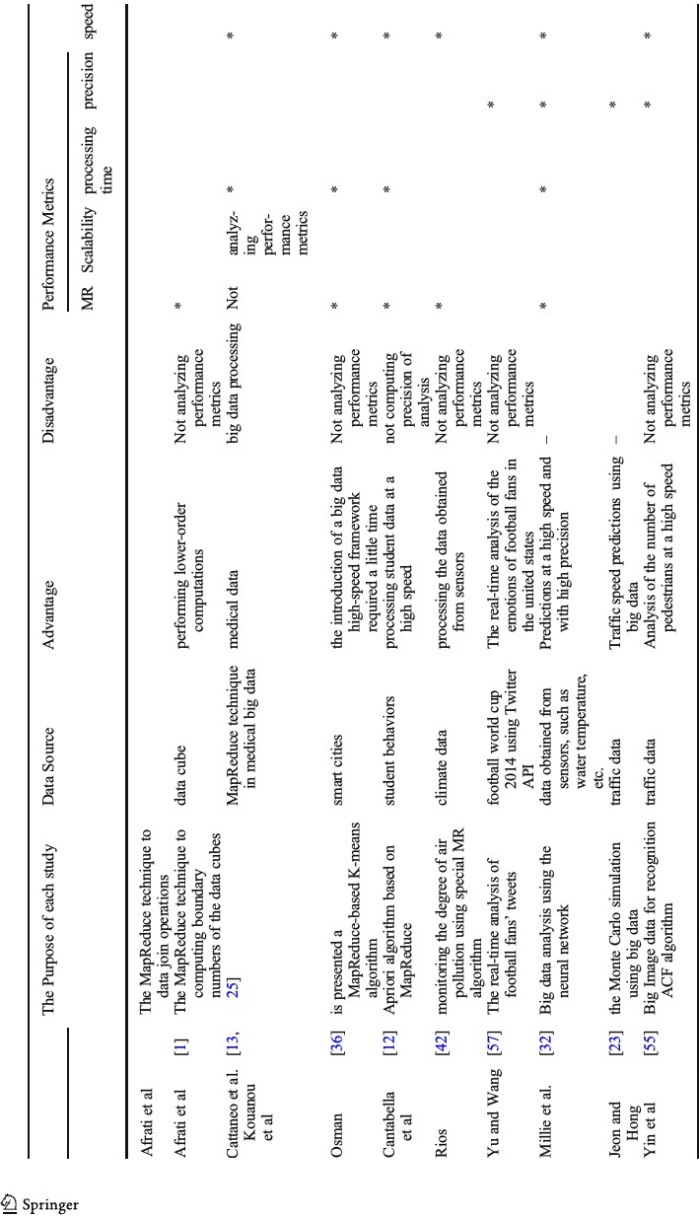
与所提出的方法(TVD-MRDL)相关的流程图如图4所示。该系统包括分析结构化和非结构化数据，以及分析多媒体数据(包括可视文件和视频)。结构化数据是指具有可预测和规则格式的数据，包括记录、字段、键、索引等。非结构化数据是指具有计算机无法识别的结构的不可预测的数据，在一般情况下访问非结构化数据具有不小的难度，特别是当需要对长串数据进行顺序搜索(即解析)以便从中导出数据单元时。大多数类型的非结构化数据都是可用的，而其中最常见的类型是图像。所以在对警方数据的分析中，大多数情况下警方会对司机的危险行为进行拍照，然后使用一定的界面进行标注。这些标签标注了司机的行为类型，并包含了对司机违规行为的完整描述。例如，如果警察发现一辆车停在“禁止停车”的标志下时，他们可以立即拍下照片，并将照片标记为“该车辆违章停放”。

表2

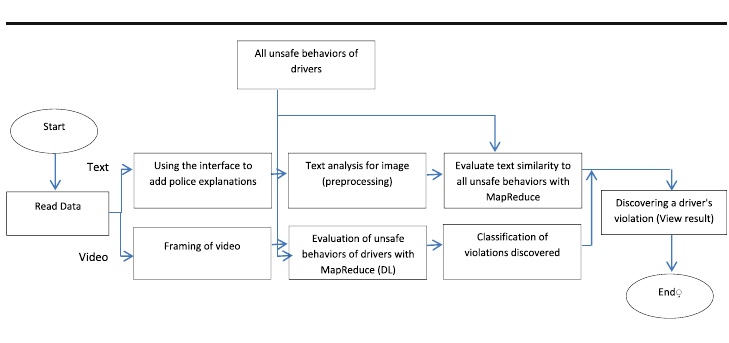


图4 TVD-MRDL流程图

在下一步中，执行预处理操作以提取有用的数据。预处理操作包括省略无用数据(冗余字母)，省略重复数据(警方描述中存在两种相似行为)，以及填充丢失数据(由于难以理解而无法被警方充分描述)。这些操作使警察描述以可用的形式显示出来。在进行预处理操作后，将提取的有用数据与驾驶领域专家收集的所有交通部门预定义的驾驶行为(驾驶违规、停车违规等)(政府或交通部门批准的表格)进行比较。通过相似性描述和行为算法中的MapReduce函数来评估和分析警察描述与所有不安全行为的相似性。

如相似性描述和行为算法所示，为了检测相似性，使用TF-IDF作为相似性准则[59]，如Eq.(1)所示。在Eq.(1)中，d1和d2分别为警察描述和每种预定义驾驶行为的向量，θ为它们的相似性程度。在第2行中，算法从两个输入参数开始，即所有预定义的驾驶行为和预处理的警察描述。在Map函数中，在第3至8行中，提取了警察陈述的频率向量和预定义的行为数据，在第9至16行中，使用Eq. 1中的余弦准则计算了它们的相似性。在第17行中，确定了警察描述与每种预定义的不安全驾驶行为之间的相似性，并将其显示为输出。第一行的Reduce函数接收所有计算值以及它们的相似度，在第3行到第7行中，计算最大的值，最终在第8行中，提供Reduce函数的输出，作为驾驶员行为和违规的类型。通过这种方式，实现了人为错误，例如视觉错误和对行为类型的准确诊断(图4)。

FORMULA

如图2所示，在下一阶段，视频监控摄像头采集并发送到交通控制中心的信息首先被转换成图片或帧，然后，使用MapReduce函数检测驱动程序行为。与视频驾驶行为机械化诊断相关的伪码在驾驶员行为机械化诊断算法中解释(图5)。

在驾驶员行为的机械化诊断算法中可以看到，在第2行中，Map函数接收捕获视频的图像或帧，在第3行和第4行中分别提取图像的长度和宽度。在第5行，使用初始图像的长度和宽度，生成一个大小相同的空帧。在第6行中，将图像灰度化等预处理操作应用于图像，并将创建的灰度图像放置在空帧中。然后，在Map函数的第7行，使用本文中采用的方法提取街道面积(即车道检测)。在第8至12行中，为了确定交通标志的类型，使用本文中使用的方法调用交通标志检测函数。在第14至21行中，使用本文中介绍的函数来识别交通灯及其颜色，并将颜色存储在一个变量中。随后，在第22行，使用本文中提出的方法在第7行提取的街道区域检测人行横道。在第24行中，如果交通灯是红色或黄色，则使用本文中介绍的方法调用人行横道区域内的车辆检测函数。然后，通过在驾驶员行为机械化诊断算法的下一行中编写的几个条件，可以识别每个行为的违规类型。例如，在第27行中，如果车辆检测功能在规定范围内识别出一辆交通信号灯颜色为红色的车辆，违章标记为“闯红灯”，存储在变量中。这个案例就是“闯红灯”违规的一个例子。在该算法的第30-34行，调用了街道区域的车辆检测函数。在第35行和第36行中，如果提取的交通标志的类型是“禁止停车”或与停车相关的任何内容，则将标记为“停车”的违规行为存储在变量中。在Map函数的其余行(第37行至第44行)中，考虑到上述条件，还会识别另外两个违规行为，即“禁止进入”和“禁止停车”。值得注意的是，这里只提到了几种普遍存在的违规行为。最后，所有从被调用函数中提取的函数都被用作Reduce函数的输入。

Map函数中每个函数得到的结果，包括最后提取的图像、交通标志的类型、交通灯的颜色、违规类型(闯红灯、禁止停车、禁止进入、禁止停车等)，都被用作Reduce函数的输入(第2行)。然后，通过第3行到第13行中的几个顺序条件，将不同函数的输出相互比较，最终确定生成帧的违章类型和驾驶员行为。例如，在第3行中，如果在分析的帧中检测到与“交通灯”相关的违规行为，则Reduce函数的输出是最终提取的图像以及交通灯的颜色和所犯违规行为的类型。在第6行中，如果在分析的帧中发生了与“停车”相关的违规行为，则Reduce函数的输出是最终提取的图像以及交通标志的类型和违规类型，即“在交通标志的区域停车”。在第9行中，如果在该帧中发生了与“禁止进入”有关的违规行为，则Reduce函数的输出是最终提取的图像以及交通标志的类型和违规类型，即“禁止进入交通标志区域”。同样，在第12行中，“禁止停车”的违规行为以及发生该违规行为的图像被显示为输出。表3列出了算法m1和算法m2中使用的术语定义。

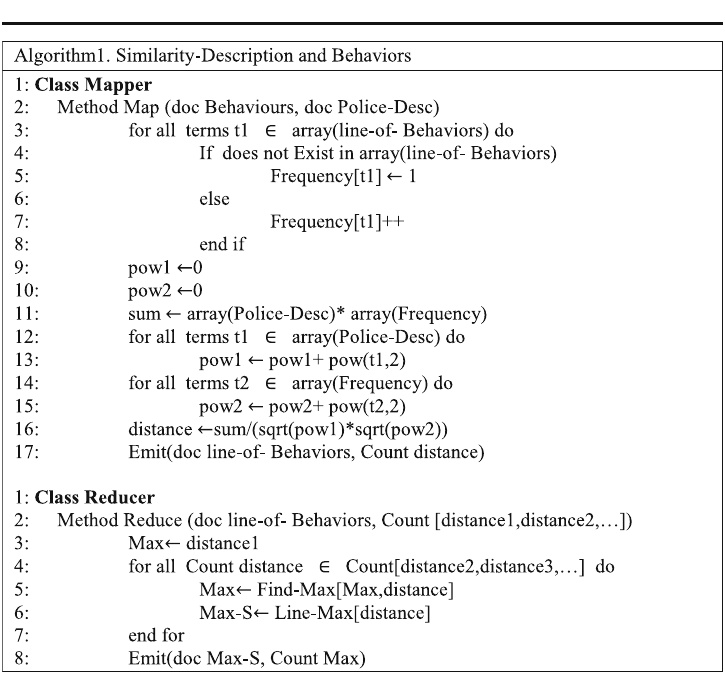
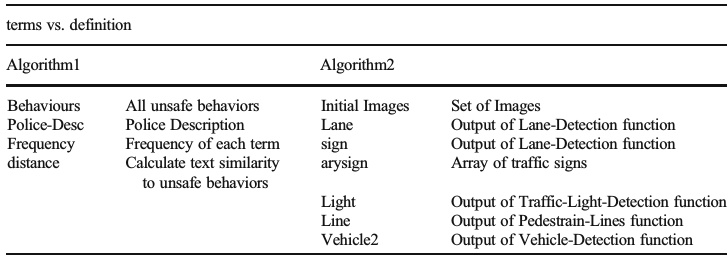


图5 MapReduce函数的伪码，用于确定警察描述与预定义驾驶行为之间的相似度。

## 4.2 TVD-MRDL系统设计

MapReduce交通违规检测的整体过程如图6所示。在图6(a)中，与批准的驾驶违章表相关的文件和警察描述在各种计算机(Slave-Nodes)上被分解。在Map阶段，Map函数应用于所有的输入数据。然后，计算警察描述与预定驾驶行为表的相似度。在Reduce阶段，将不同节点得到的结果进行组合，显示它们的相似程度。在图6(b)中，最初从交通控制中心获得的图像在从属节点上进行划分，然后每个从节点独立分析图像。因此，在Map阶段，在每个从节点中，使用深度学习来运行交通标志检测、交通灯、街道区域、车辆等功能。在 Reduce 阶段，每个系统对输入图像的不同输出进行组合，并引入一个标签，作为在图像中发现的违规行为；但在本文中，使用的是卷积神经网络。一般来说，CNN 由三个主要层组成：卷积层、池化层和全连接层。卷积层、池化层和全连接层。在每个神经网络中有两个阶段，即前馈阶段和反向传播阶段。阶段。反向传播阶段用于训练，前馈阶段用于预测。输入图像被送入网络，这是通过输入图像与每个神经元参数之间的点积完成的。和每个神经元的参数之间的点乘，最后在每一层进行卷积运算。层进行卷积运算。在提议的结构中，输入图像的所有像素共享一个卷积网络。输入图像的所有像素共享一个卷积网络。权重共享大大减少了网络可训练的自由参数数量，从而提高了通用性。最后再通过计算网络输出。图6(c)展示了不同层深度卷积神经网络用于驾驶员违章检测的整体架构。首先，将主图像划分为包含驾驶违规预定义因素(资源、位置、附近动作、附近物体和车辆)的若干图像，然后基于分层训练数据，对每个深度神经网络进行训练。最后，每个神经网络生成一个具有卷积层和连通层的多层网络。在本研究中，卷积神经网络为VGGnet类型，由13至15个卷积层和3个完全连接层组成。它的卷积层和池化层交替放置，在这些层之后，有三个完全连接的层。这为网络提供了一个彻底的增量深度评估。每个深度卷积神经网络都学习驾驶员违规图像中的特征，如交通标志、交通信号灯、街道区域、车辆等，并提取所需的特征，用于对图像进行分类，每个卷积神经网络输出的组合卷积神经网络的输出组合即为驾驶员违章类型。

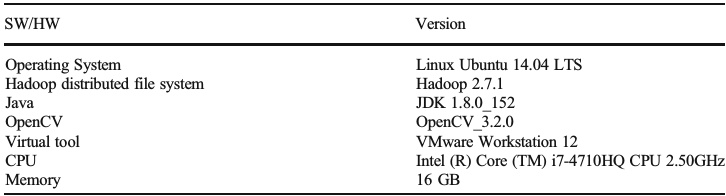


**表3算法m1和算法m2中使用的术语定义**

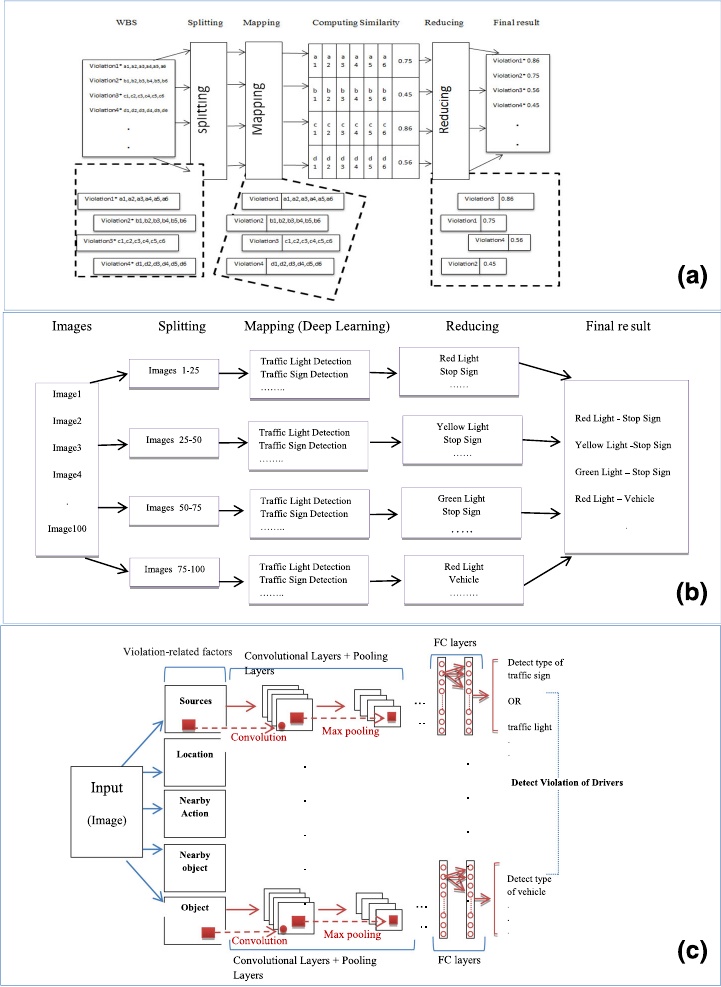
# 五．实验和结果

为了研究预测模型的改进性能，进行了一系列精心设计的实验。在这项实验研究中，所提出的方法是通过Java语言和OpenCV库实现的。测试的计算机系统在VMware环境中使用两个集群实现。第一个集群只有一个节点，第二个集群包含一个主节点和7个从节点。Linux Ubuntu操作系统安装并运行在所有节点上。拟议系统的软硬件规格如表3所示。为了评估 TVD-MRDL 的性能，我们使用了两个训练数据集。训练数据集，包括非结构化文本数据和多媒体数据分析（如视觉文件和视频）。在第一阶段，使用警方收集的一系列文本数据。在第二阶段，使用从监控摄像头收集并发送到交通控制中心的数据（表 4）。

在分析 TVD-MRDL 的结果时，为了考虑各种标准，使用了以下四种情况性能和效率、可扩展性、错误检测的准确性以及通信开销。下面将详细讨论这些方案。



**表4软件和硬件规格**



**图6 TVD-MRDL的整体MapReduce过程。**

**(a)警察描述(b)交通控制中心捕获的图像(c)深度学习**

## 5.1性能

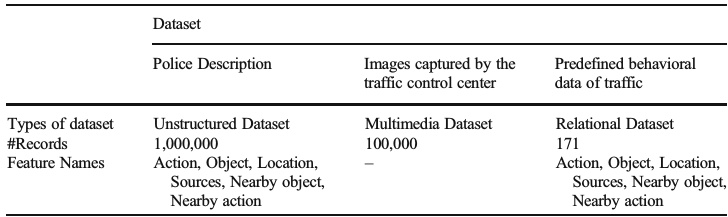
为了比较TVD-MRDL的性能，进行了两次计算实验。在这些实验中，研究了独立模式下Hadoop集群下的顺序程序，并比较了TVD-MRDL、顺序程序和传统系统的性能。这些比较将在接下来的章节中分别进行说明。

### 5.1.1单机模式下顺序程序与Hadoop的性能比较

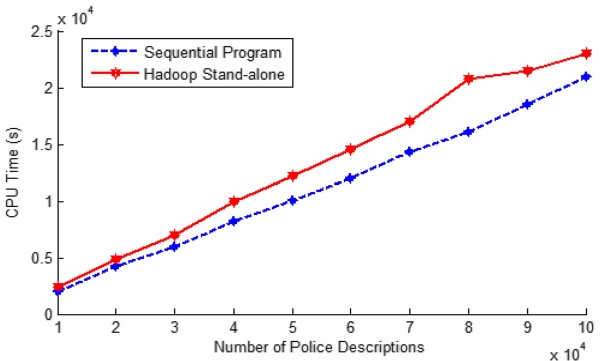
为了比较TVD-MRDL在没有MapReduce技术的顺序程序和独立模式下在Hadoop中的性能，在10,000到100,000和1000到10,000的范围内检查了许多警察描述和监控摄像机图像的处理时间(CPU时间)。从接收到警察描述和图像的时间到得到TVD-MRDL输出的时间计算处理时间。结果如图5和图6所示。

在图6中，蓝线和红线分别作为顺序程序的处理时间和单机模式下Hadoop集群的处理时间，分别用于分析不同数量的非结构化数据(警察描述)。显然，由于高开销，独立模式下的Hadoop需要更多的处理时间。此外，它们的处理时间差异随着文本数据(警察描述)数量的增加而增加。例如，在单机模式下使用Hadoop处理10,000个文本数据的时间为2,400个单元，而处理100,000个数据的时间增加到23,000个单元。因此，顺序程序的性能比独立模式下的Hadoop要好得多，因为在独立模式下的Hadoop中，当数据加载到Hadoop集群中时存在开销。

在图7中，蓝线和红线分别作为顺序程序的处理时间和单机模式下Hadoop的处理时间，分别分析不同数量的非结构化数据(监控摄像头图像)。与分析非结构化数据(文本数据)的处理时间一样，随着多媒体数据(图像)数量的增加，处理时间也会增加。例如，在单机模式下，通过Hadoop处理100张图像时，处理时间为3300个单元，而在处理10,000张图像时，处理时间增加到33,000个单元。因此，对于文本数据，单机模式下使用Hadoop的驾驶员行为检测系统由于开销过大，效率降低。



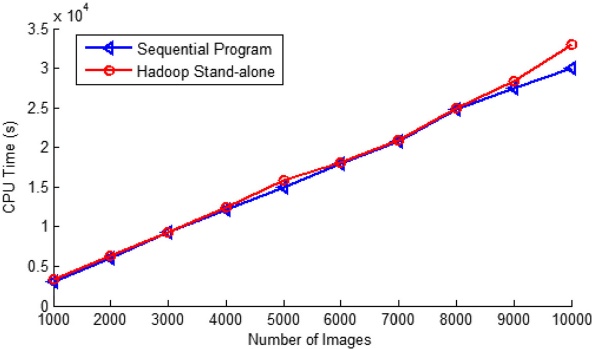
**表5数据集描述**



**图7 TVD-MRDL在顺序程序和Hadoop单机模式下对警察描述的处理时间**

## 5.2可扩展性

为了评估TVD-MRDL的可扩展性，我们进行了四次计算测试，这些测试将在下面的章节中分别进行详细解释。



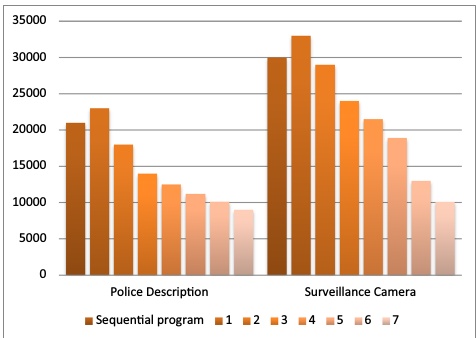
**图8 TVD-MRDL对顺序程序和Hadoop单机模式下监控摄像机图像的处理时间**

### 5.2.1一个主节点和多个从节点的Hadoop集群的效率和可扩展性比较

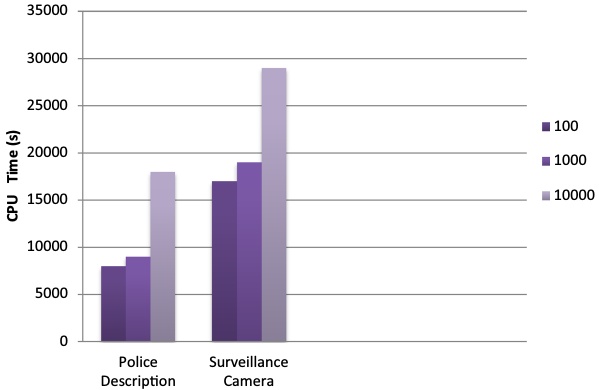
第一次评估的目的是比较通过增加节点数来实现TVD- MRDL算法的可扩展性。在这些实验中，表4所示的两个数据集使用多个节点进行处理。

从图8的结果可以看出，除了单节点集群(单机模式下的Hadoop)外，系统具有近似线性的可扩展性。然而，在对警察描述的分析中，与具有一个从节点的Hadoop集群相比，具有七个从节点的Hadoop集群的处理时间减少了近60.87%。这一改进的原因在于，在MapReduce程序中，系统自动在大型机器集群之间并行执行计算，管理系统故障，并调度系统内通信，通过这种方式，它使网络和磁盘的使用更容易，更有效。因此，通过计算负载在节点间的均衡分配，可以在不受干扰的情况下进行并行执行，减少处理时间。此外，在对监控摄像机图像的分析中，处理时间减少了70%。如4.1节所述，与顺序实现相比，只有一个节点的集群表现出较差的性能，这是由于Hadoop消耗所导致的。

### 5.2.2数据量的可扩展性

该方案旨在估算系统规模如何随着数据量的增加而变化。数据量的变化。为此，整个集群用于处理大量与违章行为有关的警方描述和交通控制中心的相关图像。违章行为和交通控制中心的相关图像。结果如图 8。处理时间与小数据集（低容量数据）的处理时间相同(因为集群超载）。不出所料，大数据量数据集（数据量较多）的处理时间与小数据集（数据量较少）相同（因为集群过载）。(数据量较大）的处理时间是线性的，因为交通控制中心提供的图像数据集是独立处理的。是独立处理的。

**图9 TVD-MRDL处理时间与节点数的关系**



**图10 TVD-MRDL处理时间与数据集数量的关系**

### 5.2.3 TVD-MRDL与传统系统的可扩展性比较

本场景的目的是考察TVD-MRDL、顺序程序和传统系统的可扩展性。程序和传统系统的可扩展性。

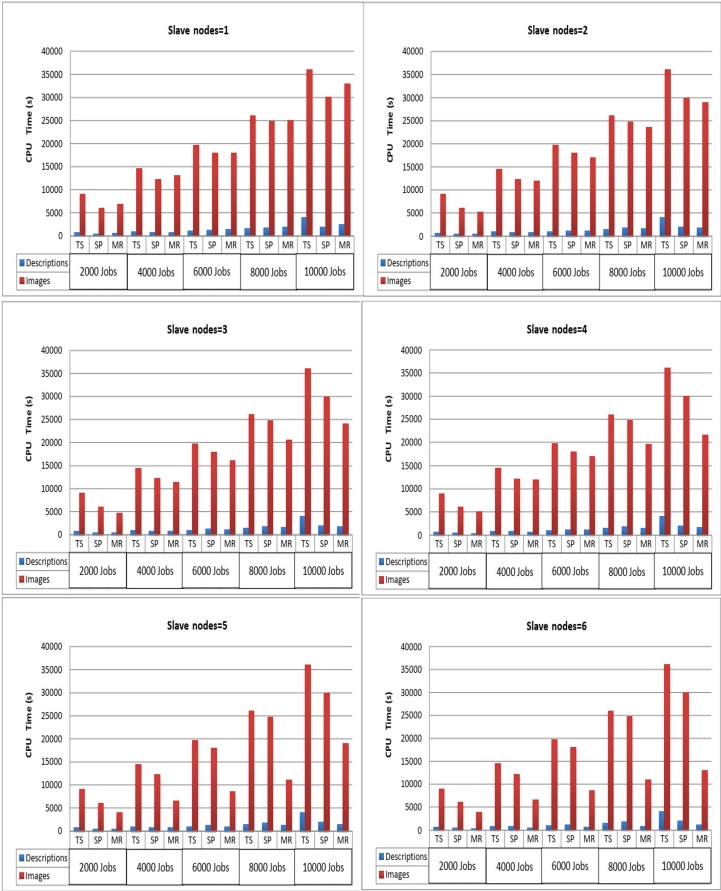
数据的大小和不同从节点的数量。为此，从节点的数量从一个到七个不等。为此，从节点的数量从一个到七个不等，并计算了 2000、4000、6000、8000 和10,000 个数据，这些数据来自警方的描述和摄像头的图像（工作），结果如图 10 所示。

从图11中可以看出，蓝色条代表警察描述的处理时间，红色条代表摄像机图像的处理时间。红色条代表从监控摄像头获取的图像的处理时间。图10在此图中，三种系统的处理时间分别称为 TS（传统系统系统）、SP（顺序程序）和 MR（TVD-MRDL）的处理时间进行了比较。节点的处理时间进行了比较。如图所示，对于给定的数据量，传统系统和具有不同从节点的顺序程序的处理时间是相同的。例如，对于4000个具有不同从节点数量(从1个到7个不等)的文本和可视化数据，传统系统的处理时间与顺序程序的处理时间相似。原因是顺序程序和传统系统中只有一个节点。但是，当使用不同数量的从节点时，各种作业的处理时间是不同的;也就是说，在每个图中，随着从节点数量相对于数据量的增加，处理时间会减少。

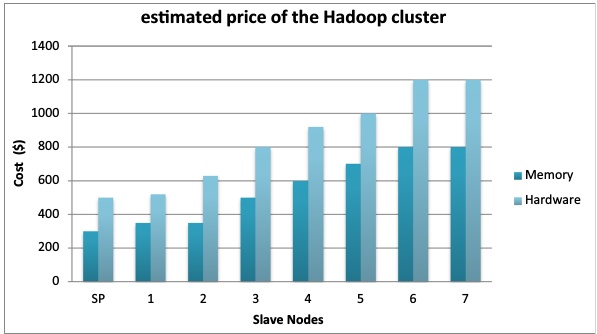
## 5.3 Hadoop集群成本

### 5.3.1 TVD-MRDL与顺序方案的资源成本比较

该方案旨在估算在顺序程序中可能实现的资源成本。在使用虚拟机进行分析期间，TVD-MRDL 被用来建立 Hadoop 集群。在本分析中，比较相同数量和相同强度的虚拟机也很重要。在本分析中，比较 Hadoop 集群内相同数量和相同强度的虚拟机也很重要。我们使用 Hadoop 版本2.7.1 版本，使用的是最弱的 VMware Workstation 类别。这个集群有多个从节点（1 至 7 个），每个节点有 7 个内核、2 GB 内存和 3 个硬盘。该集群的估计成本约为 800 美元（内存）和 1200 美元（硬件）。图 12显示了由从节点（1 至 7 个）和顺序程序组成的 Hadoop 集群的估计成本。



**图10 TVD-MRDL、顺序程序和传统系统在数据量和不同从节点数量方面的不安全行为检测可扩展性**



**图12 Hadoop集群的估计成本**

### 5.3.2 Hadoop从节点扩容时间

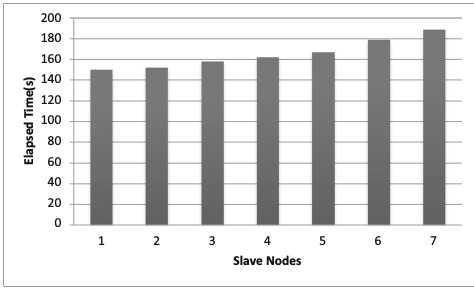
我们研究了一个集群的场景，该集群最初只有一个Hadoop从节点，然后将其扩展到七个从节点。我们对从触发命令到集群扩展完全功能结构之间所需的时间进行了检查。我们重复了5次实验，并得出了在给定节点数下的平均扩容时间（从1到7节点）。图13展示了这些结果，展示了Hadoop集群的一种弹性。在图中，横轴表示扩展集群中的节点数量，纵轴表示给定Hadoop节点在完全功能状态下所需的时间周期。

当前的性能测试并未涵盖Map或Reduce任务数量等各种参数变化对性能的影响。在Hadoop中，有超过190个参数可用于控制MapReduce作业的行为。据保守估计，这些参数中超过25个的设置可能会对作业性能产生重大影响。这些参数的调整和优化对于提高Hadoop集群的性能和效率至关重要

## 5.4 TVD-MRDL与传统系统在数据量方面的正确率、精密度和召回率的比较

本场景的目的是探索在TVD-MRDL、顺序程序(SP)和传统系统(TS)中不同数据量中检测不安全行为的标准。为了探索所提出方法的准确性、精密度和召回率，并将其与传统系统进行比较，图14显示了2000、5000和10000个数据(Job)的不同数量的警察描述和图像。

在图14中，蓝色、红色和绿色条分别代表准确率、精密度和召回率的标准。在此图中，将传统系统与顺序程序和MR程序(TVD-MRDL)进行了比较。很明显，由于系统的机械化，顺序程序和MR中这些标准的值是相同的。而且，顺序程序和程序之间的差异只与处理时间和处理时间的减少有关。从图中可以看出，在MR系统中，准确率(Accuracy)、精度(Precision)和召回率(Recall)这三个标准的值都远远高于传统系统。

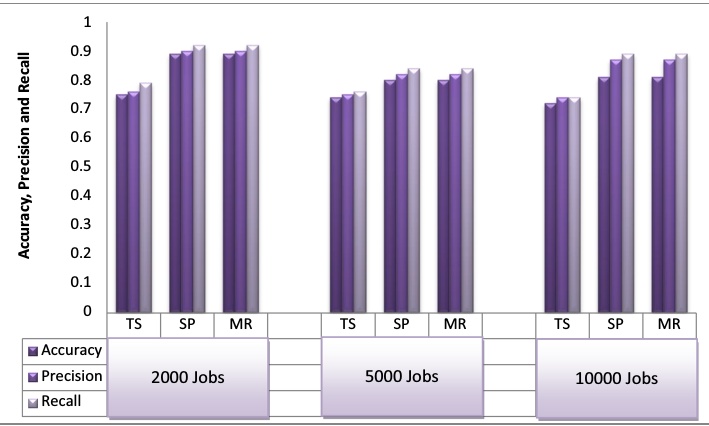


**图13 Hadoop从节点的扩容时间**

## 5.5关于Hadoop节点数量增加的通信开销

一旦Hadoop安装完成，通过SSH激活，Hadoop可以实现无需密码即可使用RCP连接不同节点的功能。在TCP协议中，这种抽象正式被称为客户端协议和数据节点协议。数据节点（从节点）会定期向主节点发送心跳（每三秒一次），以确保主节点知道它们仍处于活动状态。随着节点数量的增加，从节点和主节点之间的通信也会相应增加。

增加节点数量在处理大数据量时可以提高系统性能，这被认为是其主要优势。然而，随着节点数量的增加，通信成本也会随之增加（通信开销时间），同时多个从节点出现故障的概率也会增加，这对系统性能会产生深远的影响。因此，在系统设计中需要权衡增加节点数量所带来的性能提升与通信开销和故障率增加之间的关系，以实现系统的最佳性能和可靠性。



**图14 TVD-MRDL、顺序程序和传统系统在数据量方面的不安全行为检测的准确率、精度和召回率**

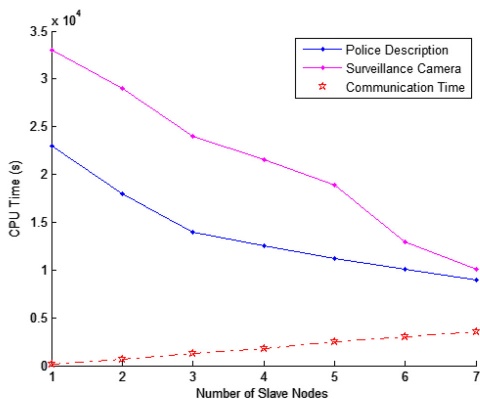
图15描述了不同从节点的通信时间(通信开销-延迟)和计算时间(执行)。选择的数据集是100000个非结构化文本数据和1000个多媒体数据，集群有1到7个从节点。图15显示，执行时间随着从节点数量的增加而减少，而通信时间很低且几乎不变。如图所示，从节点数量越多，通信延迟和信息开销在某种程度上就越明显。因此，可以这样说，随着节点数量的增加，执行时间会显著减少，通信开销也会增加，反之亦然。

## 5.6使用深度学习的Hadoop集群的效率和可扩展性的比较

为了研究TVD-MRDL系统，我们使用深度学习方法，并收集了监控摄像机生成的数据样本，如图16所示。首先，我们将从交通监控摄像头获取的输入数据划分为多个片段。在对监控摄像机图像进行分析时，我们根据专家知识推导出的有效因子对图像进行了分割。这种图像分割的方法被用于定位交通标志、交通信号灯、车辆、周围物体和动作的位置。

在Map阶段，我们根据将在下一阶段使用的函数类型，为每个分割的图像选择适当的卷积神经网络。例如，第一个卷积神经网络可能被设计用于识别与交通标志或交通信号灯相关的类别类型。在Reduce阶段，通过组合卷积网络的结果，我们可以检测并显示驾驶员的违规行为。

由于处理的是高维图像数据，将神经元与其后续的所有神经元连接起来是不现实的。因此，每个神经元只连接到输入图像的一个小区域，这个小区域被称为感受域。感受域的尺寸足够大，可以容纳超参数连接到它们上，从而实现对图像数据的有效处理和分析。这种设计使得神经网络能够更好地理解和学习复杂的图像特征，提高了系统在监控和识别任务中的性能和准确性

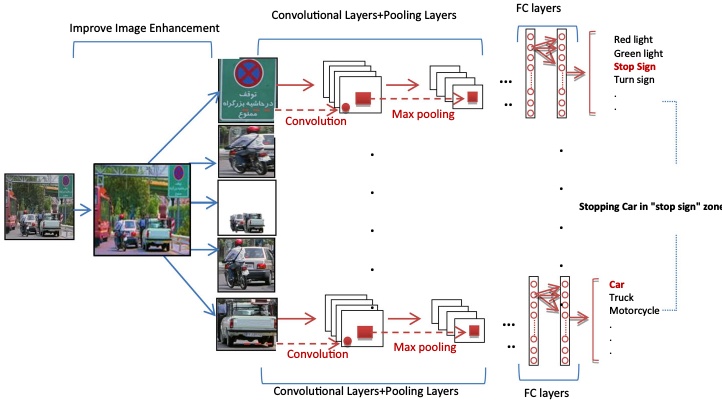


**图15 TVD-MRDL对非结构化文本和多媒体数据在不同从节点数量下的通信时间(通信延迟)和计算时间**

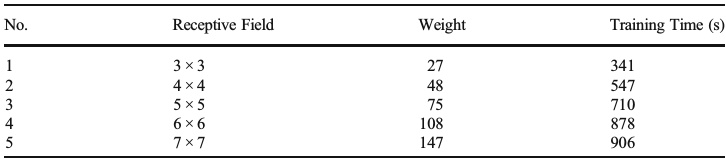
表7展示了不同训练时间和不同维度感受域的数据。例如，表7的第二行表示对于输入图像中的4×4×3（长度、宽度和深度）感受域，卷积层中每个神经元的权重因子为48。这意味着执行权重和输入图像的点积，然后通过非线性函数传递，局部建立关于位置的连接，并生成输出。对于从包含两个节点的非结构化警察数据中获取的图像数据，4×4感受域的学习时间为1189个单位。可以观察到，随着感受域面积的增加，学习时间也会增加。因此，可以得出分布式违例检测算法需要更多学习时间的结论，并且这个时间随着感受域面积的增加而线性增加，这使得它适用于大规模学习任务。

此外，为了比较TVD-MRDL系统在不同条件下的性能，采用MapReduce编程模型对监控摄像机图像的处理时间进行了研究，如图6(b)所示。提出了深度分布式卷积神经网络，如图6(c)所示，并对不同数量的从机进行了探讨。表8提供了图6(b)中MapReduce编程模型中深度学习的结果和检测函数。

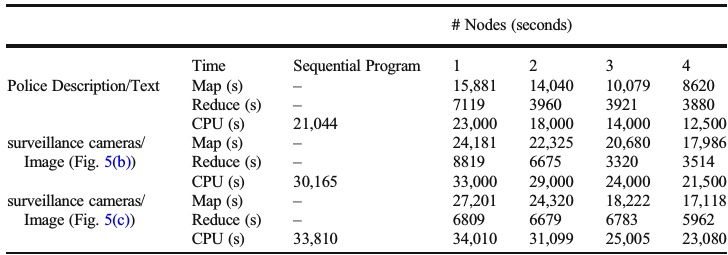
在表8中，Map Time和Reduce Time分别表示Map函数和Reduce函数的处理次数。显然，如果没有MapReduce技术，就没有时间可用于处理Map和Reduce函数。这些结果有助于评估TVD-MRDL系统在不同条件下的性能表现，并为深度学习算法在大规模学习任务中的应用提供了有益的见解。



**图16使用深度学习方法的样本测试数据**



**表7不同感受域的实验结果**



**表8使用图4(b)和图4(c)所示方法的MapReduce和处理时间**

CPU时间是由加载、变换等组成的总作业处理时间。与顺序程序相比，单节点Hadoop由于其高开销，需要更多的时间进行大规模数据处理。此外，通过增加使用深度学习的节点数量，Map函数的处理时间也会大幅减少。例如，在文本数据的处理中，使用深度学习的程序对两个节点的Hadoop集群的训练数据的处理时间等于31,099个单位;而在不使用深度学习的情况下，处理时间则减少到29000个单位。通过在每个节点上使用深度学习将节点数量增加到4个，处理时间减少到23,080个。

根据表6报告的结果，可以得出结论，由于并行执行，每个节点的CPU时间随着slave数量的增加而减少，并且在通过深度学习进行训练的过程中，由于还考虑了训练时间，因此总时间增加。由于数据分布在从机上，Map时间通常很高，CPU时间是根据MapReduce时间报告的。此外，当样本数量增加时，MapReduce技术的效率更高。因此，需要在奴隶数量和数据集的体积之间取得平衡，才能获得更高效的结果。

# 六．综述与结论

由于审核分析过程、特别是对大量图像的审美分析过程非常复杂和耗时，参考文献[52]提出了一种审核分析系统。在本系统中，利用Hadoop和MapReduce技术，通过实现多个slave来减少图像的处理时间。参考文献[10]使用了使用MapReduce框架的时间序列分析和预测分析。参考文献[10]通过MapReduce方法对大数据平台上的气候数据进行建模。利用MapReduce技术，快速获得气候数据分析结果。此外，参考文献[30]提出了一种用于处理气象(气候数据)大数据的MapReduce算法。参考文献[9]介绍了一种基于MapReduce-based人工蜂群(MR-ABC)的大数据聚类算法，结果表明，与k- means聚类相比，基于MapReduce-based聚类的效率有所提高。参考文献[60]提出了一种基于mapreduce的K-Means聚类算法，并证实该算法具有处理大数据集的高效能力。此外，参考文献[51]也研究了基于MapReduce的并行K-PSO。MapReduce也用于推荐系统的后验概率聚类和相关模型[50]。在参考文献[50]中，在不同节点上使用分布式实现观察到良好的可扩展性行为。在参考文献[14]中，引入了一种改进的TF-IDF算法来检索路透社新闻。参考文献[14]的结果表明，基于MapReduce-based的TF- IDF函数在新闻分类、词加权和聚类方面得到了改进。此外，参考文献[3]提出了一种基于MapReduce-based的SVM算法用于图像的标注(标记)。参考文献[3]中提出的分布式算法不仅通过将训练数据集划分为更小的子集并在一组计算机中对划分的子集进行优化，大大减少了训练时间，而且具有较高的准确性和精密度。

在数据库中采用MapReduce技术优化循环查询[58]和数据连接操作[2]，提出了一种基于MapReduce和连接操作的循环查询模型。在参考文献[1]中，使用MapReduce技术进行低阶计算，计算数据立方体的边界数。在医学中，MapReduce技术被用于处理大数据[13,25]。一个新的数据分析框架也可用于智慧城市[36]，讨论了传统系统与大数据系统中知识发现的方法。参考文献[12]探讨了基于大数据框架的学生行为检测。参考文献[42]利用无线传感器网络，通过大数据监测城市空气污染水平。参考文献[57]的目的是利用Twitter API收集2014年FIFA世界杯五场比赛期间美国足球迷的实时推文。参考文献[57]使用情绪分析，研究了美国球迷在推特上的情绪反应，特别是在进球后的情绪变化(无论是他们喜欢的球队还是他们的对手)。在参考文献[32]中，使用了许多传感器来收集所需的数据，包括水温等，并使用神经网络进行预测。由于传感器接收到的数据量很大，因此这些数据被认为是大数据。在参考文献[23]中，基于蒙特卡罗模拟的交通速度预测是利用大数据历史来完成的。在参考文献[55]中，大数据发布用于行人计数。参考文献[55]模拟了新的数据源，如“大数据”，以及对行人的计算分析。对所进行的相关研究进行简要回顾表1给出了使用Hadoop框架和MapReduce技术的大数据在各个研究领域的优缺点和性能指标。

在交通控制领域，使用深度学习进行了以下研究。在参考文献[56]中，提出了一种基于深度模型的新算法来检测交通标志。参考文献[56]中实现该算法的结果表明，检测准确率超过97%。参考文献[5]研究了几种通过深度神经网络检测交通标志的系统，并分析了每个系统的内存分配、准确性和处理时间。在参考文献[7]中，深度学习被用于交通标志检测系统，精度约为100%。此外，在参考文献[6]中，深度学习被用于识别交通标志。参考文献[34]讨论了在驾驶过程中基于GPS跟踪通过深度学习分类技术自动检测红绿灯。因此，在大多数关于交通标志检测的研究中，深度学习作为一种不同的方法，被用于提高准确率。然而，上述文章很少涉及处理时间的问题。据我们所知，在交通控制领域，深度学习主要应用于交通标志的检测。Asadianfam等人[8]在论文中介绍了一种基于大数据的平台，用于识别车辆驾驶员的危险行为。该平台使用了MAPE技术，但没有使用深度学习方法。本研究试图利用深度学习和MapReduce技术对大交通数据(包括监控摄像头数据和警察描述数据)进行作业处理时间的检验，以期识别不安全行为或司机的违规行为。此外，MapReduce技术在交通领域和交通大数据分析方面的研究还比较缺乏。本研究采用基于Hadoop、MapReduce技术和深度学习的TVD-MRDL系统，以期为这方面提供高效的解决方案。

信息技术的进步带来的数据呈指数级增长，使得大数据分析成为必然的结果。随着互联网和通信技术的广泛应用，交通运输行业产生大量实时数据，亟需分析和处理。毫无疑问，Hadoop已经成为大数据分析的主流技术之一，并作为云计算平台受到学术界的广泛研究。因此，研究Hadoop在交通大数据中的应用至关重要。由于交通大数据问题的复杂性，利用Hadoop大数据技术对其进行研究已成为热门话题。为了更好地了解Hadoop在交通大数据中的研究，我们对收集的多篇论文进行了分析，并确定了Hadoop在交通大数据中的8个应用场景。我们还总结了该领域已有研究的发展历程和最新成果。此外，我们还重点研究了 Hadoop 优化方面的文献，并确定了交通领域的最新研究进展。最后，我们确定了当前研究中的差距。根据我们对文献的回顾和文献计量分析，我们得出以下结论。实时数据在交通状态识别、实时交通控制、动态路线引导、实时公交调度等各种交通大数据应用中发挥着至关重要的作用。然而，Hadoop 在处理实时数据方面存在局限性。因此，将Hadoop与其他专为实时数据处理而设计的大数据框架（例如Apache Storm、Apache Flink、Apache Samza和Kafka Streams）集成，可以为交通运输中的实时大数据分析提供有效的解决方案。进一步研究这些框架与交通大数据的集成和开发，可以带来Hadoop大数据技术在交通领域应用的新进展。

大数据背后的基本思想是处理复杂的系统，交通问题就是这种系统的一个典型例子。交通涉及广泛的相互关联的因素，例如交通流量、道路状况、驾驶员行为和天气条件，这些因素高度动态且瞬息万变，因此很难开发有效的解决方案。因此，需要采取一种综合方法来解决交通问题，该方法应考虑到所有相关因素及其相互依赖性。这种方法利用大数据分析来深入了解系统并开发数据驱动的解决方案，以减少拥堵、最大限度地减少事故并提高整体运输效率。然而，现有的大数据技术，例如Hadoop，在处理关系数据时存在局限性，特别是在分析多源和异构流量数据时。应对这一挑战需要集成跨模式、多技术和跨域处理，以实现大数据集的多维关联。Hadoop 分布式文件系统 (HDFS) 在促进 Hadoop 生态系统内的分布式存储方面发挥着不可或缺的作用。然而，利用 HDFS 处理大型数据集需要严格遵守软件和硬件要求。幸运的是，数据压缩技术可以有效地减少存储空间需求，从而减轻其中的一些限制。此外，数据压缩和数据分析过程之间的统计相似性意味着通过人工智能对数据进行编码和解码可以改进数据分析，甚至取代一些HDFS功能。交通运输业可以从这一令人兴奋的前景中获得巨大的利益。然而，它也带来了重大挑战，需要人工智能和交通工程方面的跨学科专业知识来开发和实施高效的数据压缩和分析方法。弥合这两个领域之间的差距对于推进大数据技术在交通领域的应用至关重要。

如今，越来越多的运输系统开始由智能机器视觉程序控制。极低的错误率、极高的速度、极低的维护成本等诸多优势，使得各行各业越来越多地采用图像处理和机器视觉。图像监控系统是交通运输系统管理的主要手段。其优势在于可为决策提供可视化信息。本文利用 MapReduce 算法和深度学习，以机械化控制代替人力。这样一来，首先，系统中消除了由人为失误导致的各种错误，如视觉错误、疲劳等；其次，减少了所需的训练有素的人力资源数量，从而降低了相关成本。这些结果表明，与顺序程序相比，单机模式下的 Hadoop 可将大容量数据的处理时间缩短 70% 以上。此外，在 TVD-MRDL 系统中，通过将从节点的数量从 1 个增加到 7 个，处理时间分别缩短了 60.87% 和 70%。鉴于交通事故的灾难性影响和后果，通过揭示违规行为和交通中断来纠正不正确的驾驶习惯，可以为控制驾驶员的许多非社会行为、避免事故和重大人员伤亡提供机会。所研究的车辆是普通汽车，测试也是在普通汽车上进行的。作为未来的工作，这种方法将用于无人驾驶飞行器的检测。本研究全面、清晰地概述过去数年中将 Hadoop 技术应用于交通大数据的进展。我们的研究旨在提供一种新颖的视角，帮助学者理解该领域的现状并确定未来的研究方向。值得注意的是，由于某些限制，我们的调查仅关注 Hadoop 云计算技术。尽管如此，我们相信我们的研究对于打算对交通大数据中实施的其他云计算技术进行类似分析的研究人员来说可以成为宝贵的资源。此类研究可能会发现更多独特的见解并促进该领域的未来研究。

# 七．参考文献

[1]Afrati FN, Sharma S, Ullman JR, Ullman JD (2018) Computing marginals usingMapReduce. J Comput Syst Sci 94:98–117

[2] Afrati F, StasinopoulosN, Ullman JD, Vassilakopoulos A (2018) Sharesskew: an algorithm to handle skewfor joins in mapreduce. Inf Syst

[3] Alham NK, Li M, Liu Y,Hammoud S (2011) A MapReduce-based distributed SVM algorithm for automaticimage annotation. Comput Math Appl 62(7):2801–2811

[4] Aoyama K (1997) “NextGeneration Universal Traffic Management System (UTMS’21) in Japan,” inIntelligent Transportation System, 1997. ITSC’97., IEEE Conference on, pp.649–654: IEEE

[5] Arcos-García Á,Álvarez-García JA, Soria-Morillo LM (2018) Evaluation of deep neural networksfor traffic sign detection systems. Neurocomputing 316:332–344

[6] Arcos-García Á,Álvarez-García JA, Soria-Morillo LM (2018) Deep neural network for traffic signrecognition systems: an analysis of spatial transformers and stochasticoptimisation methods. Neural Netw 99:158–165

[7]Arcos-García Á, SoilánM, Álvarez-García JA, Riveiro B (2017) Exploiting synergies of mobile mappingsensors and deep learning for traffic sign recognition systems. Expert SystAppl 89:286–295

[8] Asadianfam S, Shamsi M,Rasouli Kenari A (2020) Big data platform of traffic violation detectionsystem: identifying the risky behaviors of vehicle drivers. Multimedia Toolsand Applications 79(33):24645–24684.https://doi.org/10.1007/s11042-020-09099-8

[9] Banharnsakun A (2017) AMapReduce-based artificial bee colony for large-scale data clustering. PatternRecogn Lett 93:78–84

[10] Bendre M, Manthalkar R(2019) Time series decomposition and predictive analytics using MapReduceframework. Expert Syst Appl 116:108–120

[11] Bui-Minh T, Ghita O,Whelan PF, Hoang T, Truong VQ (2012) “Two algorithms for detection of mutuallyoccluding traffic signs,” in Control, Automation and Information Sciences(ICCAIS), 2012 International Conference on, pp. 120–125: IEEE

[12] Cantabella M,Martínez-España R, Ayuso B, Yáñez JA, Muñoz A (2019) Analysis of studentbehavior in learning management systems through a big data framework. FuturGener Comput Syst 90:262–272

[13] Cattaneo G, GiancarloR, Petrillo UF, Roscigno G (2016) “MapReduce in computational biology viaHadoop and spark,” Encyclopedia of Bioinformatics and Computational Biology,pp. 1–9

[14]Chen C-H (2017)Improved TFIDF in big news retrieval: an empirical study. Pattern Recogn Lett93:113–122

[15] Chen M, Mao S, Liu Y (2014) Big data: Asurvey. Mobile networks and applications 19(2):171–209

[16] Chen M, Mao S, Zhang Y,Leung VC (2014) Big data: related technologies, challenges and futureprospects. Springer

[17] De La Escalera A,Armingol JM, Pastor JM, Rodríguez FJ (2004) Visual sign information extractionand identification by deformable models for intelligent vehicles. IEEE TransIntell Transp Syst 5(2):57–68

[18] De Mauro A, Greco M,Grimaldi M, Ritala P (2018) Human resources for big data professions: asystematic classification of job roles and required skill sets. Inf ProcessManag 54(5):807–817

[19] Dean J, Ghemawat S(2008) MapReduce: simplified data processing on large clusters. Commun ACM51(1):107–113

[20]Elotmani S, El Hitmy M(2014) “A light traffic signs recognition system,” in Multimedia Computing andSystems (ICMCS), 2014 International Conference on, pp. 459–464: IEEE

[21]Gantz J, Reinsel D (2011) Extractingvalue from chaos. IDC iview 1142(2011):1–12

[22] Goodfellow I, Bengio Y,Courville A (2016) “Deep learning (adaptive computation and machine learningseries),” Adaptive Computation and Machine Learning series, p. 800

[23] Jeon S, Hong B (2016)Monte Carlo simulation-based traffic speed forecasting using historical bigdata. Futur Gener Comput Syst 65:182–195

Multimedia Tools andApplications

[24] Kasaei SHM, Kasaei SMM(2011) “Extraction and recognition of the vehicle license plate for passingunder outside environment,” in Intelligence and Security Informatics Conference(EISIC), 2011 European, pp. 234–237: IEEE

[25] Kouanou AT, TchiotsopD, Kengne R, Tansaa ZD, Adele NM, Tchinda R (2018) “An optimal big dataworkflow for biomedical image analysis,” Informatics in Medicine Unlocked

[26] Krishnan A, Lewis C,Day D (2009) “Vision system for identifying road signs using triangulation andbundle adjustment,” in Intelligent Transportation Systems. ITSC’09. 12thInternational IEEE Conference on, 2009, pp. 1–6: IEEE

[27] Laney D (2001) 3D datamanagement: Controlling data volume, velocity and variety. META group researchnote 6(70):1

[28]Le TM, Liaw S-Y (2017)Effects of Pros and Cons of Applying Big Data Analytics to Consumers’Responses in an E-Commerce Context.Sustainability 9(5):798

[29] Lotfi E (2011)“Trajectory Clustering and Behaviour Retrieval from Traffic SurveillanceVideos,” Majlesi Journal of Multimedia Processing, vol. 1, no. 2

[30] Manogaran G, Lopez D,Chilamkurti N (2018) In-mapper combiner based MapReduce algorithm forprocessing of big climate data. Futur Gener Comput Syst 86:433–445

[31] McLauchlan P, Beymer D,Coifman B, Mali J (1997) “A real-time computer vision system for measuringtraffic parameters,” in cvpr, p. 495: IEEE

[32] Millie DF, Weckman GR,Young II WA, Ivey JE, Fries DP, Ardjmand E, Fahnenstiel GL (2013),“Coastal‘big Data’and nature-inspiredcomputation: prediction potentials, uncertainties, and knowledge derivation ofneural networks for an algal metric,” Estuarine, Coastal and Shelf Science125:57–67

[33] Moghaddam AM, Ayati E(2014) Introducing a risk estimation index for drivers: a case of Iran. Saf Sci62: 90–97

[34]Munoz-Organero M,Ruiz-Blaquez R, Sánchez-Fernández L (2018) “Automatic detection of trafficlights, street crossings and urban roundabouts combining outlier detection anddeep learning classification techniques based on GPS traces while driving,” Computers,Environment and Urban Systems 68:1–8

[35] Nguyen V, Kim H, Jun S,Boo K (2018) A study on real-time detection method of lane and vehicle for lanechange assistant system using vision system on highway. Engineering science andtechnology, an international journal 21(5):822–833

[36] Osman AMS (2018) “A novel big dataanalytics framework for smart cities,” Futur Gener Comput Syst

[37] Park SH, Jung K, HeaJK, Kim HJ (1999) “Vision-based traffic surveillance system on the internet,”in Computational Intelligence and Multimedia Applications, 1999. ICCIMA’99.Proceedings. Third International Conference on, pp. 201–205: IEEE

[38] Patterson J, Gibson A (2017) DeepLearning: A Practitioner’s Approach. “O’Reilly Media, Inc.”

[39]Phung SL, Le MC,Bouzerdoum A (2016) Pedestrian lane detection in unstructured scenes forassistive navigation. Comput Vis Image Underst 149:186–196

[40] Rahemi Z, Ajorpaz NM,Esfahani MS, Aghajani M (2017) Sensation-seeking and factors related todangerous driving behaviors among Iranian drivers. Personal Individ Differ116:314–318

[41] Rakotonirainy A,Schroeter R, Soro A (2014) Three social car visions to improve driverbehaviour. Pervasive and Mobile Computing 14:147–160

[42] Rios LG (2014), “Bigdata infrastructure for analyzing data generated by wireless sensor networks,”in Big Data (BigData Congress), 2014 IEEE International Congress on, pp.816–823: IEEE

[43] Sallah M, Sarah S,Hussin FA, Yusoff MZ (2011) “Road sign detection and recognition system forreal-time embedded applications,”

[44] Saptharishi M, SpenceOliver C, Diehl CP, Bhat KS, Dolan JM, Trebi-Ollennu A, Khosla PK (2002)Distributed surveillance and reconnaissance using multiple autonomous ATVs:CyberScout. IEEE Trans Robot Autom 18(5):826–836

[45] Schmidhuber J (2015) Deep learning inneural networks: An overview. Neural Netw 61:85–117

[46] Secundo G, Del VecchioP, Dumay J, Passiante G (2017) Intellectual capital in the age of big data:establishing a research agenda. J Intellect Cap 18(2):242–261

[47] Shvachko K, Kuang H,Radia S, Chansler R (2010) “The hadoop distributed file system,” in Massstorage systems and technologies (MSST), 2010 IEEE 26th symposium on, pp. 1–10:Ieee

[48] Stauffer C, Grimson WEL(2000) Learning patterns of activity using real-time tracking. IEEE TransPattern Anal Mach Intell 22(8):747–757

[49] Tao D, Zhang R, Qu X(2017) The role of personality traits and driving experience in self-reportedrisky driving behaviors and accident risk among Chinese drivers. Accid AnalPrev 99:228–235

[50] Valcarce D, Parapar J,Barreiro Á(2018) A MapReduce implementation of posterior probability clusteringand relevance models for recommendation. Eng Appl Artif Intell 75:114–124

[51] Wang J, Yuan D, Jiang M(2012) “Parallel K-PSO based on MapReduce,” in Communication Technology (ICCT),2012 IEEE 14th International Conference on, pp. 1203–1208: IEEE

[52] Wang W, Zhao W, Cai C, Huang J, Xu X, Li L (2015) An efficient image aesthetic analysis system using Hadoop. Signal Process Image Commun 39:499–508

[53] White T (2012) Hadoop: The definitive guide. “O’Reilly Media, Inc.”

[54] Yi S-C, Chen Y-C, Chang C-H (2015) A lane detection approach based on intelligent vision. Comput Electr Eng 42:23–29

[55] Yin L, Cheng Q, Wang Z, Shao Z (2015) ‘Big data’for pedestrian volume: exploring the use of Google street view images for pedestrian counts. Appl Geogr 63:337–345

[56] Yu Y, Li J, Wen C, Guan H, Luo H, Wang C (2016) Bag-of-visual-phrases and hierarchical deep models for traffic sign detection and recognition in mobile laser scanning data. ISPRS J Photogramm Remote Sens 113: 106–123

[57] Yu Y, Wang X (2015) World cup 2014 in the twitter world: a big data analysis of sentiments in US sports fans’ tweets. Comput Hum Behav 48:392–400

[58]Zhang B, Wang X, Zheng Z (2018) The optimization for recurring queries in big data analysis system with MapReduce. Futur Gener Comput Syst 87:549–556

[59] Zhang W, Yoshida T, Tang X (2011) A comparative study of TF\* IDF, LSI and multi-words for text classification. Expert Syst Appl 38(3):2758–2765

[60] Zhao W, Ma H, He Q (2009) “Parallel k-means clustering based on mapreduce,” in IEEE International Conference on Cloud Computing, pp. 674–679: Springer

[61] Y. M. Z. Zhang, "Environmental benefits of bike sharing: A big data-based analysis," Applied Energy, vol. 220, pp. 296-301, 2018.

[62] S. W. &. N. H. Elliot Fishman, "Bike Share: A Synthesis of the Literature Transport Reviews," no. 33:2,, pp. 148-165, 2013.

[63] N. G. H. M. Christine Fricker, "Mean field analysis for inhomogeneous bike sharing systems.," Advances in Complex Systems, vol. 14, no. 03.2011, jul 2012.

[64] R. &. R. T. A. Godavarthy, "Winter Bikesharing in US: User Willingness, andOperator’s Challenges and Best Practices.,"Sustainable Cities and Society, vol. 30.10.1016/j.scs.2017.02.006., 2017.

[65] T. L. &. W. C. J. Hamilton, "Bicycle infrastructure and traffic congestion: Evidence from DC's Capital Bikeshare," Journal of Environmental Economics and Management, Elsevier,, vol. 87(C), pp. 72-93, 2018.

[66] L. C. R. O. M. S. W. Y. Caggiani, "A modeling framework for the dynamic management of free-floating bike-sharing systems," Transportation Research Part C: Emerging Technologies, vol. 84, pp. 159-182, 2018.

[67] J. Y. Z. a. X. X. Yuan, "Discovering regions of different functions in a city using human mobility and POIs," Proceedings of the 18th ACM SIGKDD international conference onKnowledge discovery and data mining, p.186194, 2012.

[68] J. R. H. a. W. v. H. Schuijbroek, "Inventory RebalancingandVehicleRoutinginBikeSharing Systems," TepperSchool of Business., p. 1491., 2013.

[69] L. M. O. Caggiani, "A Modular Soft Computing based Method for Vehicles Repositioning in Bike-sharing Systems.," Procedia - Social and Behavioral Sciences, 2012.

[70] X. ,. W. P. G. J. S. M. A. R. K. M. J. A. D. S. Han, "Bike sharing data analytics for silicon valley in USA," 2017 IEEE SmartWorld Ubiquitous Intelligence and Computing,Advanced and Trusted Computed, ScalableComputing and Communications, Cloud andBig Data Computing, Internet of People and

Smart City Innovation,SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017 - , pp. 1-9, 2018.

[71] F. Soriguera, V. Casado, and E. Jimnez, “A simulation model forpublic bike-sharing systems,” Transportation Research Procedia, vol. 33,pp. 139 – 146, 2018, xIII Conference on Transport Engineering,CIT2018.

[72] A. Gandomi and M. Haider, “Beyond the hype: Big dataconcepts, methods, and analytics,” International Journal of InformationManagement, vol. 35, no. 2, pp. 137 – 144, 2015.