Hadoop 研究及挑战综述

何思佑，王亚强
（成都信息工程大学，四川 成都 610225）

摘要：大数据具有以下特征简称为 4 个“V”:数据是海量的（vast）、数据是多样的（variety）、价值是巨大的（value）、流传是快速的（velocity）。要处理如此规模的数据显然不是我们传统的计算机能够单独完成的，这显然成为了当今 IT 技术的一大挑战。在谷歌的推动下，2006 年 Hadoop 最初版本应运而生，其集合了分布式存储系统，分布式编程模型等必要组件，Hadoop 技术的思想和优点已对大数据处理技术产生了深远的影响，但其同样也存在一些缺陷。本文回顾了关于 Hadoop 的突出研究，并指出由于其缺点而正面临的挑战。

关键词：大数据；Hadoop；集群
中图分类号：TP311 **文献标识码**：A **文章编号**：1673-1131（2018）10-0290-02

A review of Hadoop's research and challenges

Abstract:Big data has been defined by International data agency as four V that contain vast, velocity, variety, and giant values. However, dealing with data of this size is our traditional computers cannot do alone, which becomes a challenge for IT technology. Thanks for Google, the original version of Hadoop came into being in 2006. It brought together the necessary components such as distributed file system and distributed programming mode like MapReduce. The ideas and advantages of Hadoop technology have had a profound impact on big data processing technology, but it also has some drawbacks. We reviewed the prominent research on Hadoop and points out the challenges it is facing due to its shortcomings.

0 引言

传统的集中式计算机终端是客户机，而数据由中央存储系统负责管理和储存，终端仅进行网络传输和 I/O 交互，其他所有任务都由主机处理。

上述传统集中式计算机网络的一个比较大的缺点在于网络传输瓶颈，相对分布式集群处理方式来说，前者并不适用于大规模数据[1]的实时处理和传输。

可以说集中式与分布式是两个完全相反的概念，分布式的计算和存储可以在不同的地方进行，而其处理过程均由本地工作站负责，这是分布式网络[2]的一大特点。相比较集中式计算机，其限制大大减少，主要体现在以下几个方面：①多用户访问，②信息共享。③访问快速。以上特点使系统更加灵活，既可以满足独立计算地区用户的需求，也可以作为共享资源为企业服务。综合上述特点，分布式系统十分契合大数据时代[3]的 4 个特点，适用于大数据应用。

经过十多年的发展，Hadoop 已经成为了一个具有完整生态链的分布式大数据处理系统，其拥有存储、计算、查找、任务管理等众多组件于一体的组件且正是由于这些组件被开源社区不断的完善整理获得了前所未有的高容错、高性能、低成本的强大能力。最重要的一点在于 Hadoop 中的框架大部分由 Java 语言编写，众所周知 Java 语言的最大特点就是其可移植性强，这使得将 Hadoop 应用在各个平台间不需要做任何多余处理，对于企业的生产应用非常理想。同时，Hadoop 也支持其他语言编写的程序，如 C++等。

但由于其立项较早，不可能面面俱到这也导致了 Hadoop 未来的发展的一些瓶颈问题，如编程模型的逻辑合理性和复杂性等，这些在后文都将有所探讨。

1 Hadoop 相关研究及现状

Nutch[4]是一个始于 2002 年的开源 Java 实现的搜索引擎，其创始人结合 2003 年谷歌的 GFS 研究提出了 NGFS[5]，随后又整合 mapreduce 于 2006 年正式提出 Hadoop 框架。此后，大量学者和社区开发者参与到 Hadoop 的维护与开发中，也是从那个时候开始 Hadoop 的组件逐步完善。在这之前 Hadoop 已经拥有了分布式计算框架和分布式文件系统，但缺少一个分布式数据库做数据的存储于是 Hbase 被选为最合适的候选者，并在后来确立了其在分布式数据库中的地位。为了简化使用 Hadoop Hive 与 Hbase 配套出现，作为一个简单的数据仓库工具，其优点是学习成本低。相类似的研究还有 Pig 等工具的出现，其作用也是为了简化 Hadoop 的操作，为查询大规模半结构化数据提供方便快捷的接口。有了上述工具已经可以支持基本的大规模数据处理但管理还相对困难于是需要一个“管理者”来管理分布式应用中的命名、状态等问题，ZooKeeper 可以将上述提到的所有工具统一管理起来并向用户提供可靠的服务接口。后续所有研究几乎都是基于上述工具的性能优化展开，如 S Li 等的研究改进了 mapreduce 计算模型，也有人研究在一定资源下如何预估 Hadoop 任务所需时间以及如何优化资源配置的模型。

2 Hadoop 的不足与挑战

2.1 Hadoop 的不足

Hadoop 作为大数据时代分布式系统的先驱，必然存在着许多的不足，其不足主要表现在两个方面：

第一：大数据时代的很多数据处理业务需求是针对流数据的，而流数据的一大特点就是实时性很强，这也是目前 Hadoop 的一大缺点：它并不太擅长处理流数据，时延较高。如果对延迟要求较低的需求 Hadoop 是可以胜任的，但就实际情况而言，大部分企业需求要求是在毫秒级的。Hadoop 的时延较长主要原因在于它的文件系统工作机制，首先它需要接收到流数据然后将其存入 HDFS，再将文件切割为 N 份提供给不同的工作节点，这是一项比较耗时的工作，这项工作完成后 Hadoop 的计算框架才正式启动作业。也许计算作业仅花费 1 秒左右，而文件的存储切分工作就会花去 1 分钟甚至更长的时间，这一点是 Hadoop 的主要缺点，也是将来急需解决的一点。

第二 Hadoop 计算模型 MapReduce 有着很严重的局限性，编程框架将所有操作都抽象成两个步骤即：Map 和 Reduce，虽然这两种操作已经可以满足大部分逻辑需求，但考虑到大数据处理过程中存在的复杂性这种框架的抽象层次还是太低，提供的操作类型较少。并且每台节点计算机都会拥有 Job，Job 负责所有复杂的计算，但大量的 Job 管理成了最难解决的问题：没有中间结果。Hadoop 框架将中间结果隐藏导致交互式数据处理变得十分困难，而所有中间结果也需要使用 HDFS 文件系统进行读写并需要等待所有 MapTask 都完成后才可以开始。

新一代 PGIS 技术在智慧消防中的创新应用

王江腾
（内蒙古自治区赤峰市公安消防支队，内蒙古 赤峰 024000）

摘要：PGIS 警用地理信息系统作为“智慧消防”架构体系中关键基础服务平台之一，为实现一张图指挥、一张图调度、一张图分析、一张图决策、一张图展示、一张图管理起到高端应用的基础支撑。2014 年后，新一代 PGIS 技术在原有基础上升级优化后提供的创新性应用，对于加速推进大数据、物联网、移动互联网等现代科技与消防工作的深度融合，全面提高消防工作科技化、信息化、智能化水平，实现信息化条件下火灾防控和综合应急救援工作转型升级具有重要作用。

关键词：PGIS；智慧消防；地理信息
中图分类号：D631.6 **文献标识码：**A **文章编号：**1673-1131（2018）10-0291-02

0 引言

公安警用地理信息平台简称为“PGIS 平台”，是以公安信息网络为基础，以警用电子地图为核心，地理信息技术为支撑，服务于公安业务管理、信息共享和决策支持的可视化为目标的重要信息化基础设施。该平台的软件由公安部统一组织开发，采用统一的标准规范，基于开放的商用基础 GIS 软件，能够提供统一基础应用服务和工具，可以在部、省、市三级分布式部署的平台软件。“十一五”期间，公安部消防局组织研发消防 GIS，“十二五”期间完成与各地公安 PGIS 平台对接，建立公安机关警种间条块结合的应用模式，实现平台服务和数据资源的最大化共享及应用，进一步推动了 PGIS 平台数据建设，不断丰富警用地理信息资源。

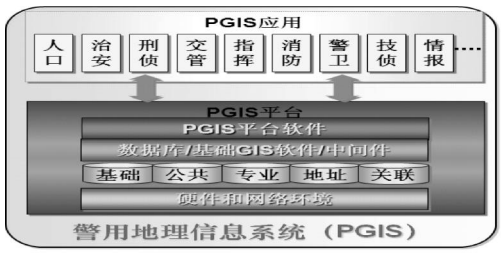

图 1 “十一五”国家科技支撑计划 PGIS 总体架构

2014 年，为解决 PGIS 数据获取体系不完善、图层升级跟不上实战需要、不适应大数据、移动警务等实际应用问题，公安部启动国家科技支撑计划“新一代 PGIS 项目试点和关键技术应用示范”，以进一步更好服务于公安“四项建设”和建设“平安中国”，提高平台的整体完备性、适用性、实战性、先进性、稳定性和可靠性。

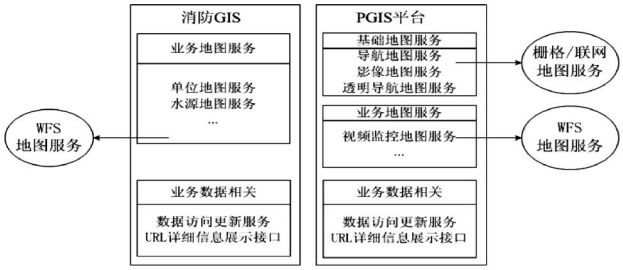

图 2 现行消防 GIS 与 PGIS 对接体制

1 新一代 PGIS 技术发展现状

1.1 建设背景

针对第一代全国 PGIS 平台建设及应用中所存在的发展不平衡、实战应用成效不明显、未形成跨区域资源共享、服务机制、数据质量不高、更新不及时等突出问题，公安部在“十一

总而言之，这里的问题在于太多的文件系统读写操作。

2.2 Hadoop 的挑战

如果要说 Hadoop 真正的挑战那么不得不提到比 Hadoop 还要新型的技术 Spark。Spark 是 UC Berkeley AMP lab（加州大学伯克利分校的 AMP 实验室）所开源的类 Hadoop MapReduce 的通用并行框架，Spark，拥有 Hadoop MapReduce 所具有的优点；但不同于 MapReduce 的是 Job 中间输出结果可以保存在内存中，从而不再需要读写 HDFS，因此 Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法。Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同之处，这些有用的不同之处使 Spark 在某些工作负载方面表现得更加优越，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。Spark 是在 Scala 语言中实现的，它将 Scala 用作其应用程序框架。与 Hadoop 不同，Spark 和 Scala 能够紧密集成，其中的 Scala 可以像操作本地集合对象一样轻松地操作分布式数据集。

在计算速度方面很难抛开 Spark 单独的谈论 Hadoop，并且曾有学者指出 Spark 或许将完全替代 Hadoop，其他相关大数据工具也将不再针对 Hadoop 更新而转向 Spark。上述言论也许才是当今 Hadoop 最大的挑战，但其实上文我们已经分析 Hadoop 最大的弱点在于计算框架，所以最恰当的说法或趋势或许是 RDD 计算模型将会取代 MapReduce 模型，毕竟目前 Spark 依靠的文件系统和数据库工具等都是来自 Hadoop 的 HDFS 还有 Yarn 等，并且 Hadoop 的低成本优势很难被忽视。

3 结语

通过对 Hadoop 发展历史的回顾和优缺点分析以及与 Spark 的对比可以发现 Hadoop 基于其完整的生态圈在大数据处理的地位还是不可撼动的，但同时期具有的局限性也逐渐暴露，后续研究需要从使用便捷性和编程模型合理性上进行改良否则其使用成本将会让许多大数据从业者望而生畏。此外，一个框架的诞生并不是为了使用而使用，Hadoop 和其他生态圈一样不可能直接适合某些实际项目的直接运用。所以大数据时代的技术还需要根据自身的实际项目需求来选择不同的组合开发和改进。只有通过不同新兴技术的改进和合理运用才能使我们从大数据时代获得我们预期的利益。

参考文献：

[1] 亓开元，赵卓峰，房俊，等. 针对高速数据流的大规模数据实时处理方法[J]. 计算机学报，2012，35(3):477-490.
[2] 费宏慧，李健. 大数据的分布式网络入侵实时检测仿真[J]. 计算机仿真，2018(3).
[3] 王元卓，靳小龙，程学旗. 网络大数据:现状与展望[J]. 计算机学报，2013，36(6):1125-1138.
[4] Ghemawat S，Gobioff H，Leung S T. The Google file system[J]. Acm Sigops Operating Systems Review，2003，37(5):29-43.
[5] Cafarella M，Cutting D. Building Nutch[J]. Queue，2004，2(2).