 DOI: https://doi.org/10.12345/bdai.v5i7.19739

Big Data Processing and Analysis Study Based on Hadoop

Jinwang Zhang Huang He Xiyun Wang

Guangdong Innovative Technical College, Dongguan, Guangdong, 523000, China

Abstract

In the rapid development of information technology, big data has become an important force in promoting industrial upgrading and social progress. Hadoop is an open-source big data processing framework with high scalability and strong distributed processing capabilities, making it an important tool for professionals to process and analyze big data. For the purpose of improving the effectiveness of big data processing, this paper will use Hadoop technology to build a big data platform. In practical research, the author first introduces big data and its processing technology, then analyzes the big data processing flow based on Hadoop, and finally builds a big data platform based on Hadoop, hoping to provide some reference for relevant personnel to process and analyze big data.

Keywords

big data; Hadoop; processing; analysis

基于 Hadoop 的大数据处理与分析研究

张金旺 何煌 汪细云

广东创新科技职业学院，中国·广东 东莞 523000

摘 要

在信息技术快速发展过程中，大数据已经成为促进产业升级与社会进步的重要力量。Hadoop属于一种开源的大数据处理框架，其具备扩展性高、分布式处理能力强的特性，使得其成为相关人员处理与分析大数据的重要手段。出于提升大数据处理有效性的目的，论文将运用Hadoop技术搭建大数据平台。在实际研究中，笔者首先介绍大数据及其处理技术，然后分析基于Hadoop的大数据处理流程，最后搭建基于Hadoop的大数据平台，希望能够为相关人员处理与分析大数据提供一定借鉴意义。

关键词

大数据；Hadoop；处理；分析

1 引言

在信息爆炸的新时代中，数据的产生、传输、储存以及分析速度快速提升。大数据不但涉及企业运营、社交媒体等方方面面，更是在交通、医疗、教育等行业中扮演着重要角色。但是因为大数据存在多样性与复杂性的特点，导致数据处理与分析面临很大困难。在此情况下，相关人员需要使用Hadoop这种开源大数据处理框架实施大数据处理与分析。

2 大数据及其处理技术概述

大数据也被叫做巨量资料，就是资料量拥有非常大的规模以至于无法通过主流软件工具在规定时间内完成整理、管理、处理以及收集工作。近年来，大数据已经渗透到我国的各行各业，在此过程中其表现出真实性、多样、低价值密度、高速、大量的特点，而且其能够利用分布式可扩展储存的方式完成管理查询数据的工作，现阶段尽管很多研究机构已经收集大量数据，但是因为缺乏科学的分析手段，而且数据仓库在维护过程中需要支付较高成本，在此情况下很多企业开始应用基于 Hadoop 结构的分布式文件系统 [1]。Hadoop 是一个关键的分布式系统框架，专为大规模数据处理设计，其核心原理包括 MapReduce 计算模型和 HDFS 存储机制。HDFS 采用主从结构，具备出色的故障容忍能力。该系统能够在普通个人计算机上广泛部署，形成多节点的数据存储网络，有效地对海量数据集进行分块管理和存储。此外，HDFS 支持一次写入、多次读取的策略，保证了数据的一致性，适应了当前大数据时代对高吞吐量的需求。MapReduce 是谷歌开发的一种分布式程序设计范式，其通过“映射”和“化简”两个步骤处理大数据。映射阶段在不改动原始数据的情况下，将大文件分割成小文件，转换成独立的元素进行逐步处理，并生成中间结果列表。随后，化简阶段根据函数

【基金项目】基于深度学习的三维人脸重建抗遮挡网络研究（项目编号：2022TSZD03）。

【作者简介】张金旺（1982–），男，中国广东茂名人，本科，讲师，网络工程师，数据分析师，从事大数据技术与人工智能应用技术、高职教学教育等研究。

值对映射产生的文件实施整合或缩减，提炼出不同结构或无关数据的关键特征，并将最终结果存储到特定位置。

3 基于 Hadoop 的大数据处理流程

Hadoop 属于分布式系统基础架构，开发者为 Apache 基金会所，其主要是借助集群的作用高效进行运转与储存。Hadoop 可构成一个分布式文件系统，能完成海量数据储存工作，同时可构建出 MapReduce 编程模型，可满足同时运算大规模数据集的要求。在实际应用过程中，Hadoop 需要通过以下几个步骤处理大数据：①数据采集：利用数据库接口、网络爬虫等技术手段，充分收集各种渠道中产生的数据；②数据预处理：主要是采取标准化、清洗、转换等手段处理收集完成的原始数据，为后续顺利进行分析与处理工作提供方便；③数据存储：在 GDFS 中存储经过预处理的数据，为后续开展分布式处理打下坚实基础；④数据分析：运用 MapReduce 等编程模型，分析与处理储存在 HDFS 中的数据，分析出其中有价值的信息；⑤输出结果：借助图片、表格等方式展示最终分析出的结果降低用户理解与使用难度 [2]。

4 基于 Hadoop 的大数据平台

4.1 系统架构设计

基于 Hadoop 的大数据平台主要是利用物联网云平台产品架构落实设计工作，其间需要经历数据分析、数据解析、数据预警、数据清洗等环节，整个平台主要由三部分组成，分别为平台通用层、泛接入层以及泛应用层，图 1 为其系统架构。

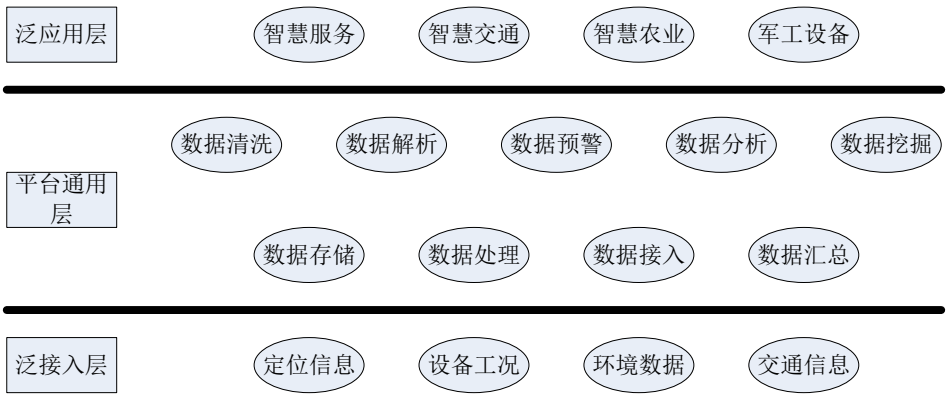

图 1 系统架构图

4.1.1 泛接入层

在搭建泛接入层的过程中，主要利用了常规服务器和集成化设备，刻意避开成本高昂的高性能计算硬件和高端存储设施，从而大幅降低了经济投入。另外，在构筑泛接入层的阶段，借助通用的接入通信规范，可以高效地实现不同设备的互联互通，推动物联网大数据在跨领域的广泛应用，进而减少平台部署和运营的经济负担。

4.1.2 平台通用层

在该平台中平台通用层属于关键组成部分，其能够将数据监管、数据处理、数据存储、数据接入等多种服务提供给用户，能够高效处理且集中地处理海量数据。

4.1.3 泛应用层

泛应用层在实际应用过程中，不需要进行数据分析与计算工作，用户只需将想要查询的信息输入到接口中就能够在短时间内调取平台中已经处理完成的各项数据，并且利用图表的方式直观、呈现出相关数据，用户可快速理解数据中呈现的问题。

4.2 系统功能设计

为了最大限度地利用 Hadoop 的技术，保证大数据平台具有高度的实用和稳定，相关人员可以构造出一种功能架构，并利用 Java 语言进行程序设计，建立 B/S 体系结构，保证各个模块的正常运行。

4.2.1 数据接入模块

在开发数据引入系统时，相关人员主要应用了以下三项关键技术手段：

①均衡负载：该系统的设计思路主要集中在运用 Linux 虚拟服务器以及先进的负载均衡算法。Linux 虚拟服务器在网络层进行操作时，可以有效提高服务器集群的工作效率。为了保证系统的流畅运行，需要在同一局域网内部署虚拟服务器和网关服务器，并确保它们处于同一 IP 地址段。同时，为了提高系统的稳定性和降低单点故障的可能性，可以将 Linux 虚拟服务器升级为硬件负载均衡设备，以保持数据引入的高效率。

②数据关口：这一部件的主要作用涵盖了数据上报的管理、指令的发布以及数据路由的设置。通过采用可变的配置方式，能够高效地完成数据路由的设计目标。

③消息代理：该中间件负责传递多种数据信息，包括原始数据、追踪数据和警示信息，确保各类信息能够顺畅传递 [3]。

4.2.2 数据存储模块

在 Hadoop 中，分布式文件系统发挥着至关重要的作用，能够有效地满足海量数据的处理需求。将该系统运用到该平台中，可以对备份与离线的数据进行高效的管理，从而使该

平台具备了以下三大主要功能：

①自动实施新文件副本复制，使得存取数据的工作变得更加可靠。

②新增的服务器节点能无缝整合到现有系统中，增强了系统的可扩展性和兼容性。

③通过分布式算法均衡地处理数据访问请求，从而保证了高效的数据流量。

4.2.3 数据处理模块

数据处理模块由以下两个组件构成：

①即时处理。鉴于此平台对于大数据处理的速度有严格的需求，期望能在短时间内迅速完成运算。当前，即时处理组件的设计主要借助 Storm 和 SparkStreaming 框架。Storm 以其亚毫秒级的消息处理延迟而著称，而 SparkStreaming 则可在秒级内处理消息。为了提升平台的运行效率，相关人员可选择采用 Storm 来构建即时处理组件。

②批量计算。在设计这一部分时，相关人员应采用 Hadoop 的核心技术 MapReduce 框架，它特别适合处理大规模的数据集离线运算。为发挥 MapReduce 架构的高可扩展性、高容错性及动态自适应能力，对其进行二次包装与优化，使其能够更好地实现高效率地工作，同时也为下一步的任务调度工作奠定基础。

4.2.4 数据导入与交换模块

数据导入与交换模块主要由两部分构成：

①数据接收与整合。借助数据接口的功能，该部分能够高效地把采集自终端的原始数据和即时处理的数据导入，并分别保存在分布式的文件系统及数据库里。

②数据交互平台。该系统为使用者提供了支持多种语言的接口和规范化的开发接口。多种语言的接口使得使用者能够在数据传输过程中享受到高效率和便捷性，数据的交换过程包含了实时数据的获取和向终端设备发送指令等多个环节。

4.2.5 监控报警模块

监控报警模块由以下两个核心部分组成：

①系统观测：本系统采纳 Hadoop 框架，针对多个服务器节点的运作状态进行不间断的动态跟踪，从而为系统的后续优化及资源的合理配置提供了重要的数据基础和决策依据。在系统刚开始运行阶段，借助收集到的监控信息，能够对系统的性能实施有效提升。

②预警系统：该系统从硬件层面着手，利用插件智能化地跟踪服务器核心指标，如 CPU 使用率、硬盘工作负载以及内存使用情况；在软件层面，系统启动后能够即刻将进程信息录入并储存于指定文件夹中，若发现进程运行异常，系统将立即发出警报；而在业务层面，平台上的各功能单元均支持用户自定义预警标准。

4.3 系统性能优化

在构建大数据平台的过程中，系统效能提升是保障平台高效、稳定运行的核心步骤。以下将探讨几个关键的效能优化策略和手段。

4.3.1 硬件配置改善

①内存升级：扩充服务器的内存容量，确保大数据处理时拥有充足的内存资源，防止因内存不足造成处理速度减慢或系统故障。

②存储增强：选用高速存储设备，例如固态硬盘（SSD），以提升数据读写速率，缩短处理数据的时间。

③网络强化：优化网络带宽和稳定性，确保数据传输时不出现因网络问题导致的数据丢失或延迟。

4.3.2 软件配置调整

①JVM 参数定制：对 Java 虚拟机（JVM）进行参数配置，比如调整堆内存大小、优化垃圾回收机制，以提升 Java 应用程序的执行效率。

② Hadoop 集群调整：对 Hadoop 集群进行性能优化，包括调整 MapReduce 任务的数量、数据块大小等设置，以增强集群的处理能力和稳定性。

③数据库优化：对数据库进行改进，如创建有效的索引、优化查询命令，提升数据检索的速度。

4.3.3 算法改进

①数据压缩技术：运用高效的压缩算法，如 Snappy、LZO，压缩存储的数据，减少存储需求，提升数据传输效能。

②重复数据消除：在预处理阶段应用数据去重算法，去除重复数据，减少数据冗余，提高处理效率。

③并行计算策略：针对大数据处理的并行计算需求，选择适用的并行计算算法，如 MapReduce、Spark，加快数据处理速度。

4.3.4 监控与优化工具

①监控系统：使用监控工具实时监控系统性能，如 Zabbix、Prometheus，以便迅速识别并解决性能问题。

②优化工具：借助性能优化工具对系统进行调优，如 Hadoop 性能优化工具、JVM 优化工具，全面提升系统性能。

5 结语

综上所述，本次研究以 Hadoop 技术为基础构建了大数据处理与分析平台，该平台不仅满足了企业对于实时数据处理和批量计算的需求，还通过数据导入与交换模块、数据接入模块、数据存储模块、监控报警模块等组件，确保了数据的完整性和安全性，为企业提供了全方位的数据支持。此项研究证明 Hadoop 技术在大数据处理与分析中具备突出优势，值得相关人员展开深入研究，在未来相关人员可尝试在其中融入机器学习等新兴技术。

参考文献

[1] 赵子晨,杨锋,郭玉辉,等.基于Hadoop技术的加速器大数据安全存储与高效分析系统设计[J].现代电子技术,2024,47(8):9-17.

[2] 汤笛,吴长梦涛,张欣悦,等.基于Hadoop平台的灾害大数据处理及可视化[J].电脑与电信,2024(4):80-84.

[3] 李敏,文燕,叶煜.基于Hadoop的设施蔬菜产销大数据架构分析[J].四川农业科技,2024(3):29-33.