

文章编号: 2096-1472(2017)-11-54-03

基于大数据分析的海量信息软件系统设计与开发

李锦华

(菏泽学院计算机学院, 山东 菏泽 274015)

摘要: 基于大数据分析前提下所开展的海量信息处理任务, 在保障效率的同时更要确定准确程度。在此基础上本文重点分析了基于大数据环境下的海量信息软件系统逻辑构成, 以及设计过程中的具体软件、硬件结构组成模式, 为系统设计开发和任务进行建立适合的环境, 全面提升数据信息处理速度, 实现大数据分析环境下更高效稳定的功能。

关键词: 大数据分析; 海量信息; 信息处理; 软件设计; 系统构建

中图分类号: TP391 **文献标识码:** A

Design and Development of the Mass Information Software System Based on Big Data Analysis

LI Jinhua

(College of Computer, Heze University, Heze 274015, China)

Abstract: Based on big data analysis, the information processing tasks should be more determined and accurate in the process of ensuring efficiency. On this basis, this article focuses on analyzing the logical construction of the mass information software system based on the big data environment, and the specific structure model of software and hardware in the process of design. The suitable environment is established for system design and development tasks. The system greatly raises the data information processing speed, and implements functions with better efficiency and stability in the big data analysis environment.

Keywords: big data analysis; mass information; information processing; software design; system construction

1 引言(Introduction)

大数据分析是一种商业计算模式, 其来源于分布式计算、并行计算和网格计算。Hadoop作为云计算的核心技术, 目前在工业界得到了广泛的应用。Hadoop是Apache开源组织按照MapReduce的工作原理设计的一种开源的分布式处理框架, 也是云计算环境下最著名的开源软件。在Hadoop系统中, 应用程序可以并行运行在由大规模廉价硬件构成的分布式系统中。Hadoop在内部实现了容错和扩展机制, 可以构建高可靠性和高扩展性的分布式系统。在集群中运行分布式应用程序时, MapReduce编程模型简单易用。Hadoop提供的MapReduce编程模型是谷歌MapReduce的开源实现。在MapReduce编程模型中, 开发者只需要编写Map和Reduce函数, 而任务调度、容错等机制由底层实现。因此, 即使开发者没有分布式系统的经验也能编写出高效的分布式应用程序。下文将以此为例进行重点论述。

2 基于大数据环境下的海量信息软件系统逻辑构成(The logical construction of the mass information software system based on the big data environment)

2.1 逻辑分层

根据海量信息软件的处理需求, 将大数据环境下的软件

信息传输划分为三个层次, 依次进行逻辑数据划分。首先对数据进行分隔, 使不同功能层数据在传输中通过分隔来达到相互独立的状态, 避免数据之间产生干扰。分层后产生多个区域, 在区域基础上对相关的软件控制程序进行补充完善, 从而达到最佳使用效果。针对运行使用期间可能会产生的相关技术隐患问题, 更应该探讨出控制方案, 并在系统内深入完善, 了解系统对运行环境的控制需求, 以及现场可能会产生的相关隐患问题。逻辑分层与控制系统软件分隔是相互联系的, 当分隔软件所传输的数据可以达到精准度标准, 海量信息处理任务也能高效完成^[1]。进行海量信息处理, 需要同时进行多项任务。进行相互控制并观察彼此之间的影响能力, 同时完成多项数据任务还需要在逻辑分层中体现出系统控制程序分层。根据所得到的处理信息进行相关指令控制, 从而达到最佳控制管理效果。最后是数据库分层, 用来与所处理的信息进行对接, 避免信息传输期间受到数据库连接干扰影响。

2.2 海量信息处理中的数据抽取

基于大数据环境下所开展的海量信息处理, 首先要针对数据进行抽取, 控制其中所存在的安全隐患, 并结合技术性方法从更深层次探讨需要完善的内容。数据抽取需要在程序控制环境下来进行, 并观察各项数据传输中所面临的具体环

境，通过系统程序加密控制来降低环境因素造成的干扰。数据抽取同样是建立在大数据处理结构基础上，通过各个系统程序之间进行相互控制。大数据分析系统结构示意图如图1所示。

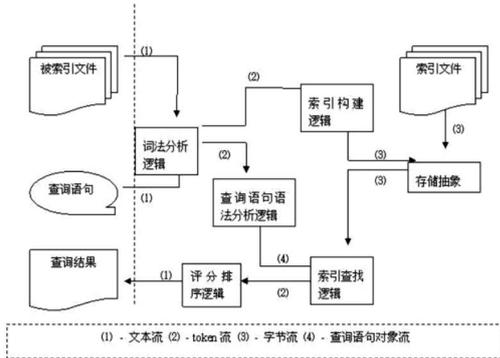


图1 基于物联网的大数据分析系统结构示意图

Fig.1 The big data analysis system structure diagram

图1中的信息抽取构建模式，可以形成多种方案，并在现场建立起更合适的环境。关于设计期间可能会产生的相关问题，充分探讨优化解决方案，最终应用效果也能得到明显提升，通过结构之间高效配合来实现最终的程序控制和信息方案提取。信息提取任务完成后，接下来的分析任务也能得到高效开展，并不会造成彼此之间的干扰影响。

2.3 数据跟踪定位

大数据处理环境下信息获取需要针对位置进行定位，这样才能发现其中是否存在影响数据运行的相关因素，为技术性方案开展建立适合的环境。数据跟踪定位后可确定数据信息获取的主要部分，并在此基础上进行信息海量处理，在系统内部同时完成多项信息处理任务，为最终的管理任务开展建立适合的环境。系统运行期间数据信息处理速度还会受到网络环境的影响。利用大数据分析技术会结合网络平台来进行，数据跟踪期间能够同时完成多项处理任务之间的对接。通过数据准确跟踪定位来提升海量信息处理的整体效率，从而达到最佳处理效果，并为各项管理控制计划开展打下稳定基础。数据分析处理不仅仅体现在方案自身控制方面，更与平台运行稳定性之间存在必然联系，数据跟踪定位可帮助快速判断请求合法性，避免非法请求进入到系统中，影响到最终的功能实现。

3 海量信息存储处理软件的综合设计(Comprehensive design of the mass information storage and processing software)

3.1 文件系统设计中的节点选择

对于文件系统设计需要判断节点选择模式，观察在系统中是否存在可能会影响到数据信息处理效率的因素。根据数据处理信息需求，将节点划分成为数据节点、管理节点与监控节点，可以根据数据信息处理方向以及使用需求来对内部信息进行选择，从而达到最佳处理效果。系统设计中的节点选择可帮助区分文件，对文件内部所存储的信息进行高效识别处理，从而达到最佳处理控制效果。面对海量且分布的文件系统，首先针对其节点部分进行设计构建，基于节点基础上实现对数据的来源识别以及处理对接，避免在最终控制

能力上受到影响。为确保海量信息处理任务能够安全稳定实现，可以增加设计DataNode节点，实现海量数据的存储和查询任务。处于网络环境下的数据存储以及传输均存在安全隐患，很容易受到黑客或病毒攻击而出现损坏或丢失的现象，设计DataNode节点后相邻区域的数据会共同参与到搜索中，从而实现数据节点更高效的反馈，最大程度提升数据处理反应速度。

3.2 文件存储及副本设计

文件存储部分设计需要确定不同文件的主副本，对文件的查看以及数据存储权限范围进行控制设计。文件存储信息后搜索路径也随之得到确定，接下来进行的副本读取也要在此路径内进行，避免出现信息获取后的误差问题。文件内部数据信息存储以及副本读取需要远程控制功能来实现，在此基础上重点探讨提升优化方案的有效措施，具体系统结构模式如图2所示。处于信息读取环境可能会产生的各项综合处理模式，会直接将数据传输到对应的数据库中，从而实现数据信息之间的转变，为接下来所开展的各项数据信息建立稳定的现场环境。图2中显示的内容可以发现，这种安全隐患环境中可能会产生的问题，要综合多方面因素来探讨并探究。客户端产生请求后通过服务器会传导到相应的文件模块中，在此范围内进行信息资源获取。对于副本部分的读取控制，需要对副本进行结合文件夹的信息读取控制，观察其中所存在的问题，并在此基础上充分探讨优化解决方案，全面提升综合控制效果，完成副本与文件夹之间更稳定的对接任务^[2]。

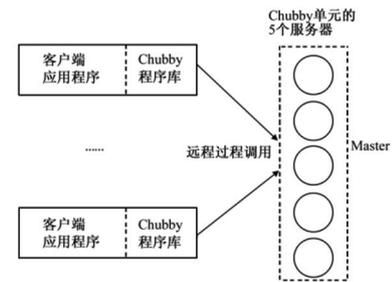


图2 文件存储远程数据读取程序控制模块

Fig.2 The control module of file storage and remote data reading program

3.3 海量信息恢复系统设计

处于大数据分析环境下，针对海量信息进行存储控制，一旦发生信息丢失问题，将会影响到接下来的各项处理功能的实现。因此设计中的一项重点任务是针对软件内容进行设计和开发，观察系统构建中需要继续深入解决的内容，并结合技术性方法来综合解决，构建恢复系统，建立数据库记忆功能，能够自动或者根据操作者对软件系统的使用需求来备份数据，并对其中的重要数据进行自动加密保存，当使用期间受到网络病毒攻击出现数据丢失现象后，也能通过海量信息恢复系统来自动完成恢复，确保文件信息正常使用功能不会受到影响。针对数据信息处理期间可能会产生的相关问题，更应该探讨出有效解决方案，结合内部信息处理模式来进行方案构建^[3]。建立数据信息处理自动恢复模式，并在系统内自动完成备份处理任务，恢复系统设计期间还需要考虑大

数据处理环境下的存储能力，避免在功能实现上出现干扰，并通过各个结构之间的相互配合来综合提升处理速度。大数据处理模式下的海量信息恢复系统结构如图3所示。

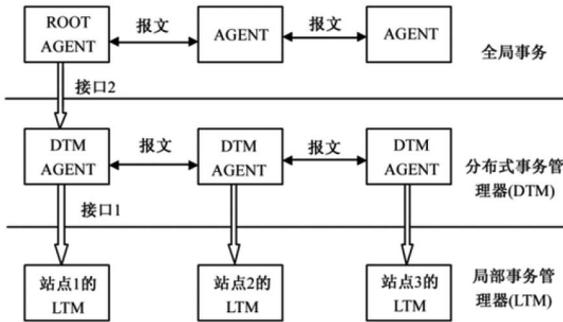


图3 大数据处理模式下的海量信息恢复系统结构

Fig.3 Mass information recovery system structure under big data processing mode

3.4 目录存储与负载均衡设计

海量信息处理不仅是针对文件夹内部数据来进行，其中更为重要的是对数据进行均衡分析，对其内部所包含的目录进行快速扫描记录。目录检索完成后接下来进行的内部数据处理分析效率也更高，并且能够根据所进行的项目扫描来完成负载数据的划分，实现数据之间的对接^[4]。负载均衡设计同样需要建立在节点基础上，以节点为分界衡量对象，实现目录与其内部文件之间的对接，并在此基础上充分构建适合数据分析处理的软件，综合提升目录扫描处理效果。设计访问用户的分区，直接进行访问请求与文件夹之间的对接，任务得到高效处理后，接下来的各项处理分析任务也能更加理想，建立起综合联系系统。根据不同IP用户来进行DNS分流设计，从而构建出内部控制体系，综合提升海量数据信息存储速度，并在达到处理对接速度标准的前提下降低误差产生几率，为构建综合指标建立稳定的现场环境。处于数据信息分析模式下的目录存储，更应该体现出变化性，观察是否在信息内存在需要继续强化完善的内容，并建立起适合的现场环境，为管理任务进行建立有效的环境^[5]。目录在检索过程中自动更新，也能够避免在检索内容上受到影响，文件夹内部信息处理效率也会得到明显提升。

4 大数据分析环境下的海量信息软件系统开发实现(Development and implementation of the mass information software system in the big data analysis environment)

4.1 脚本测试代码构建

对于脚本测试过程中的代码构建和实现，更应该体现出系统内部程序控制。根据硬件结构划分模式来进行软件程序汇编，对海量数据软件系统的功能脚本进行检验，观察其功能实现是否可以达到预先设计的标准。Tomcat为脚本检验代码构建的服务器载体，分析系统运行使用中所面对的处理环境，以及信息内部是否存在影响最终功能实现的因素，有关于设计期间比较常见的运行功能实现检验，同样可以在此服务器支持下来实现。采用Windows 10操作系统作为运行载体，面对文件夹中的信息进行处理检索，对所设计构建的方

案可行性进行检验，有关于设计期间确定的重点功能，在此范围内更应该体现出积极性，并深入观察其中可能会造成隐患的因素，同样采用规避措施解决^[5]。代码构建需要借助SAT模型来实现，面对不同的文件夹更应该体现出其中的功能差异性，从而避免数据传输出现彼此干扰的现象。

4.2 海量信息处理系统模拟

确定海量信息软件构成模式后，为提升设计方案可行性，可以首先针对其中的设计方案进行模拟，检验所设计系统中存在的漏洞，并通过程序优化设计的方法来统计漏洞数量和所在位置，从而通过程序设计来将其优化处理。确定一个模拟数据库，其中的信息数量要达到标准，这样接下来所开展的各项设计方案才不会因此受到影响。利用软件来对模拟数据库中所存储的内容进行检验，判断是否在数据传输和存储过程中出现过错误情况。在检验系统内自动记录这部分错误参数，有关于数据传输中受网络病毒影响所受到的攻击，更应该体现出其中的深层次控制方案，实现控制方案与系统数据效率提升方面的相互连接，并构建出更合理的设计模式。系统模拟中可以对文件夹安全程度进行检验，模拟黑客攻击过程，体现出内部控制，并构建出合理的现场控制环境，了解软件系统运行中的风险类型后，接下来所开展的各项风险控制任务也能更高效完成^[6]。通过设计方案强化的方法来帮助降低软件系统运行风险性，效率提升也更科学稳定。

5 结论(Conclusion)

综上所述，云计算和物联网的迅速发展，越来越多的个人和企业选择将自己的业务迁移到大规模的数据中心，以此来降低本地的硬件成本和系统维护费用。数据中心存储的数据量十分庞大，管理系统的复杂性较高。从存储设备级别上看，数据中心为了控制成本，大量采用廉价存储设备，致使数据极易因硬件设备故障而丢失。这些都对海量数据存储性能、可靠性等方面带来了挑战。云存储是解决海量数据存储的最有效手段。本文提出的一种海量大数据云存储系统设计，是应用系统需要实现的关键技术。

参考文献(References)

- [1] 史金梅,夏伟.基于大数据分析的学生最优选课方案模型的设计与实现[J].现代电子技术,2017,40(14):30-32.
- [2] 马莲姑,杜锋,黄寿孟.基于大数据分析的MOOC智能自主学习系统设计[J].现代电子技术,2017,40(20):12-13.
- [3] 杨玉新,马伟,赵阳.基于大数据分析的电网精准规划信息系统设计[J].现代电子技术,2017,40(7):155-158.
- [4] 龚为纲,张严,蔡恒进.海外自媒体中涉华舆情传播机制的大数据分析——基于Reddit平台的海量舆情信息[J].学术论坛,2017,40(3):21-31.
- [5] 张斌,彭其渊.基于大数据的铁路客户关系管理系统设计研究[J].铁道运输与经济,2017,39(6):42-48.
- [6] 杨宗亮,张玉茜,李建飞.一种基于地理空间大数据的网络舆情监测软件架构[J].测绘通报,2017(3):96-100.

作者简介:

李锦华(1979-),女,硕士,讲师.研究领域:计算机应用技术.