

龚明明, 叶伦强. 地震信息网络数据的动态存储方法研究[J]. 地震工程学报, 2020, 42(4): 1043-1048. doi: 10.3969/j.issn.1000-0844.2020.04.1043

GONG Mingming, YE Lunqiang. A Dynamic Storage Method of Seismic Information Network Data[J]. China Earthquake Engineering Journal, 2020, 42(4): 1043-1048. doi: 10.3969/j.issn.1000-0844.2020.04.1043

地震信息网络数据的动态存储方法研究

龚明明¹, 叶伦强²

(1. 湖南外国语职业学院 信息科学与工程系, 湖南 长沙 410116;

2. 西南民族大学 信息与教育技术中心, 四川 成都 610225)

摘要: 为了提高地震信息网络数据存储的效率和稳定性, 本文提出一种地震信息网络数据的动态存储方法。首先, 设计和分析地震信息网络数据的动态存储系统中的硬件部分, 基于 Hadoop 的分布式集群大数据动态存储系统, 整合多台存储服务器, 为一个集群系统, 并平衡存储服务器负载情况, 获取性能指标的比例值, 通过加权公式获取综合性能参数, 根据综合负载指标和服务器综合指标参数获取最佳存储服务器。对本文系统、高速数据存储系统和网络编码云存储系统进行大数据动态存储对比实验。实验结果表明: 地震信息网络数据的动态存储耗时最短, 存储效率最快; 具有较高存储稳定性, 且不会出现过载问题。

关键词: 动态存储; 地震信息; 网络数据; 集群; 负载指标; 大数据

中图分类号: P631

文献标志码: A

文章编号: 1000-0844(2020)04-1043-06

DOI: 10.3969/j.issn.1000-0844.2020.04.1043

A Dynamic Storage Method of Seismic Information Network Data

GONG Mingming¹, YE Lunqiang²

(1. Department of Information Science and Engineering, Hunan College of Foreign Studies, Changsha 410116, Hunan, China;

2. Information and Educational Technology Center, Southwest Minzu University, Chengdu 610225, Sichuan, China)

Abstract: This study proposed a dynamic storage of seismic information network data to improve the storage efficiency and stability of seismic information network data. First, the hardware part of the dynamic storage system of seismic information network data was designed and analyzed. Then, a dynamic storage system of a distributed big data cluster based on Hadoop was constructed. In this system, multiple storage servers were integrated into a cluster system, and the load of storage servers was balanced to obtain the proportion of performance indicators and comprehensive performance parameters through the weighted formula. Comparative experiments were performed on the proposed system, the high-speed data storage system, and the network coding cloud storage system. Experimental results showed that the dynamic storage system of the seismic information network data had the shortest time and the fastest storage efficiency. The storage

收稿日期: 2019-11-08

基金项目: 湖南省教育厅科学研究项目(16C1061); 西南民族大学中央高校基本科研业务费专项资金资助项目(2019NQ54); 国家自然科学基金重点项目(570314002)

第一作者简介: 龚明明(1982-), 女, 湖南常德人, 硕士, 讲师, 研究方向: 计算机网络、物联网。E-mail: Lixia19806@163.com。

stability was high, and no overload problem was detected.

Keywords: dynamic storage; seismic information; network data; cluster; load index; big data

0 引言

在地震信息网络数据采集中,地震仪器的性能对地震信息网络数据采集的质量起到决定性作用,而存储技术决定了地震仪器的性能^[1]。磁带是地震信息网络数据采集中存储的主要介质,对其稳定性和高效性有很高的要求。而在数据采集完成后,采用高性能的计算机和软件对数据进行处理,最后对海量的数据进行存储。因此许多计算机存储技术在地震信息网络数据处理中得到应用,并在快速发展中为数据的处理提供了良好的条件,数据存储的效率和速度也有了稳步的提升^[2-3]。

地震信息网络数据存储技术是地震数据特点与计算机存储技术相结合的产物。提高存储容量和访问速度一直是技术研究的目标^[4]。上世纪 80 年代,最先进的地震仪器是法国和美国制造的 SN388 和 DFS-V。国产地震仪器为西安石油仪表厂生产的 DZ-751 型。采用 SEG-B 和 SEG-C 的记录格式,用 9 轨磁带存储关于地震的数据资料。存储容量约为 150 M^[5]。在仪器性能方面来说,与国外仪器相比,国产仪器与其还存在较大差距,地震资料采集后,国内仪器的误差率较高。而且,地震处理还处于初级阶段。计算机和地震处理软件目前还是从国外引入。在这一阶段,地震处理的主要存储介质是轨道磁带。磁带主要作为运行地震处理作业输入输出的存储介质,花费了大量的时间,效率低下。上世纪 90 年代初,国内地震仪器主要是西安石油仪表厂的 GYZ-4000 和 BGP-WF-1006。到了 90 年代末,国内仪器基本停止研发。国外已经生产了各种先进的地震仪器,SN-388、SYSTEM TWO 是当时具有代表性的产品。2000 年以来,油田地震采集仪器主要在国外生产,而后国内地震处理软件 CRISYS 在地震现场数据采集质量控制中发挥了重要作用,因此该软件得到了广泛的应用^[6]。2006 年,中国石油天然气集团将地震数据采集记录系统列为重点工程,并在 2007 年开始研发。到 2010 年,国产地震仪 ES109 成功测试。

地震相关的数据信息存储涉及地震勘探的各个阶段,针对地震网络信息存储的研究,有些学者分析数据存储的机能性以及地震数据格式,但很少涉及动态存储技术。因此,本文提出一种地震信息网络数据

的动态存储方法,以提高地震信息网络存储的存储效率和存储稳定性。

1 硬件设计

地震信息网络数据数量大的特点,这也直接导致其对数据存储、采集和检索都有着较高要求。通常大数据收集速率为 MB/s 和 GB/s,存储容量多为 TB 或 PB^[7-9]。Hadoop 是分布式系统的基础架构,可以存储 TB 或 pB 级的大数据,同时可以检测和处理硬件故障。相似的系统还有 Flink、Spark。与 Hadoop 系统相比,Flink 在一定程度上提高了性能,Spark 则是实现了多种数据的转换。基于 Hadoop 的分布式集群大数据动态存储系统的总体结构如图 1 所示,由中央控制集群、大数据采集集群、大数据存储集群、高速以太网连接模块 4 部分组成。

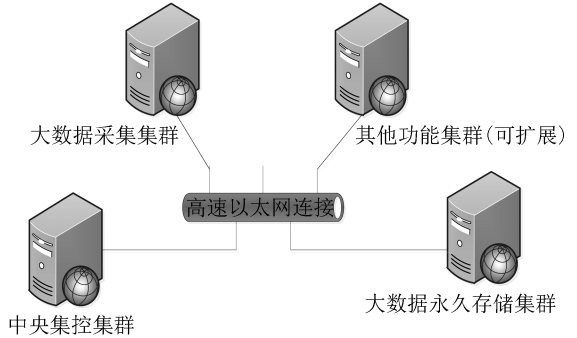


图 1 基于 Hadoop 的分布式集群大数据动态存储系统的总体结构

Fig.1 Overall structure of the dynamic storage system of distributed cluster big data based on Hadoop

1.1 中央集控集群

系统运行时的管理控制模块是中央控制集群。中央控制集群的主要作用是通过控制多个系统,将指令准确下达到各系统中,提高单位时间内执行的任务数,在整个系统中起到的大脑。系统获取用户需求后,中央控制集群命令高速以太网连接模块指挥大数据采集集群进行采集^[10]。中央控制集群有着实时监控数据状态、解决异常状况、开启、关闭和更新集群的功能,分为主控单元和信令调度单元两部分,主控单元负责指示信令信道,其电路结构如图 2 所示。

1.2 大数据采集集群

大数据采集集群是系统的内部入口,通过设置

并发数据采集子模块,来提高数据采集的速度;在快速采集动态数据后,将采集的数据定期导入永久存储集群进行存储^[11],其总体结构如图 3 所示。

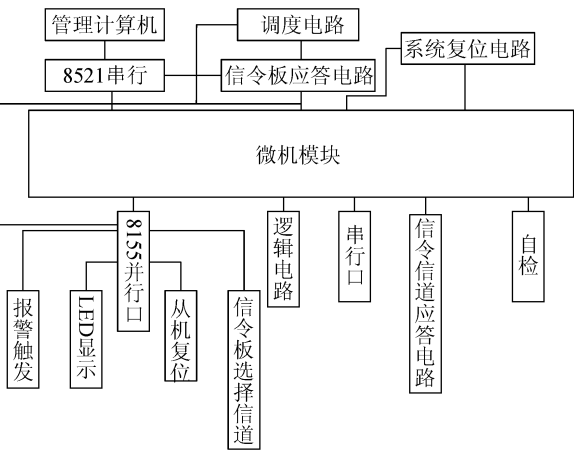


图 2 中央控制集群主控单元电路结构图

Fig.2 Circuit structure diagram of main control unit of central control cluster

和负载状态,并获得索引参数。根据存储服务器性能与负载指标,计算其综合指标。

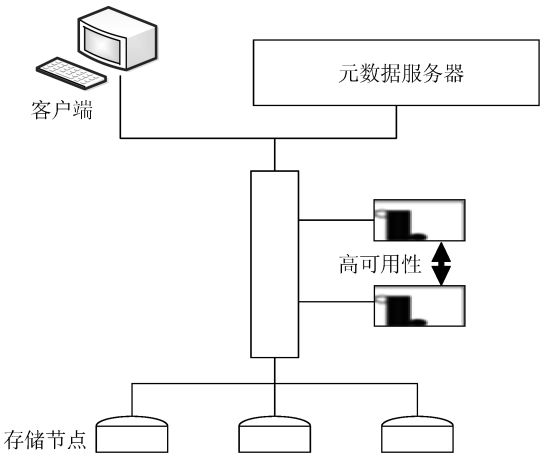


图 4 大数据持久存储集群的总体结构

Fig.4 The overall structure of big data persistent storage cluster

2.1 综合性能参数

如果假设大数据永久存储集群中共有 m 个存储服务器,聚合形式为 $F=\{F_1,F_2,\cdots,F_m\}$ 。存储服务器综合性能指标的计算过程为:a.获得存储服务器输出的性能指标;b.计算每个存储服务器的性能指标,得到最大值;c.比较最大值,获得性能指标的比例值,得到绩效指标的比例值。利用加权公式,得到综合性能参数:

(1) 存储服务器的 CPU 处理速率比值 CP_a 为:

$$CP_a = \frac{A_a}{\max(A_1,A_2,\cdots,A_m)} \tag{1}$$

式中: A_a 代表第 a 台存服务器 CPU 处理数据的速率, $\max(A_1,A_2,\cdots,A_m)$ 则表示存储服务器 CPU 处理数据的最大速率。

(2) 存储服务器的内存比值 NP_a 为:

$$NP_a = \frac{N_a}{\max(N_1,N_2,\cdots,N_m)} \tag{2}$$

式中: N_a 为第 a 台存储服务器的内存容量, $\max(N_1,N_2,\cdots,N_m)$ 则是服务器内存最大容量。

(3) 磁盘读写速率比值 SP_a 为:

$$SP_a = \frac{S_a}{\max(S_1,S_2,\cdots,S_m)} \tag{3}$$

式中: S_a 代表第 a 台存储服务器的磁盘读写速率, $\max(S_1,S_2,\cdots,S_m)$ 用来表示磁盘读写最大速率。

(4) 网络带宽比值 DP_a 为:

$$DP_a = \frac{D_a}{\max(D_1,D_2,\cdots,D_m)} \tag{4}$$

图 3 大数据采集集群的总体结构

Fig.3 The overall structure of big data collection cluster

1.3 大数据永久存储集群

大数据永久存储集群实现了数据的定期存储和更新,在一定程度上可以视为“仓库”。具体结构如图 4 所示,其总体结构是分为元数据服务器、存储节点和客户端三部分^[12]的分布式结构。这种结构将大数据划分为不同的小数据并缩小到单个节点可以处理的程度,在数据存储节点之间均匀分布。

2 软件设计

大数据动态存储系统^[13]是一个基于 Hadoop 的分布式集群,主要负责衡存储服务器的负载,以此确保服务器可以稳定安全地完成用户的存储指令,到达减少运行时间,避免系统崩溃的目的。在运行过程中,负载均衡调度器决定着存储服务器的性能

式中: D_a 为第 a 台存储服务器的网络带宽, $\max(D_1, D_2, \dots, D_m)$ 是网络最大带宽。

上述过程中的值域均为 $[0, 1]$, 判断存储服务器性能的好坏的依据式比例值的大小, 并以此来分析存储服务器的综合性能, 再根据其综合性能为其分配权重, 将 V_c, V_N, V_s, V_D 设为加权系数, 值域为 $[0, 1]$, 值的大小表示该性能指标的关键度, 则第 a 台存储服务器的综合性能参数 $ZP(B_a)$ 为:

$$ZP(B_a) = V_c * CP_a + V_N * NP_a + V_s * SP_a + V_D * DP_a \quad (5)$$

2.2 服务器负载参数

用 Y_a 表示在某一时刻, 第 a 台存储服务器的实时连接用户的数量, 则 FY_a 表示服务器的最大负载, 那么该时刻第 a 台服务器负载参数 $EP(B_a)$:

$$EP(B_a) = \frac{Y_a}{FY_a} \quad (6)$$

2.3 服务器综合索引参数

利用 2.1 和 2.2 部分所计算出的综合性能参数和服务器负载参数, 计算求得第 a 台存储服务器综合指标参数 $G(B_a)$:

$$G(B_a) = \frac{ZP(B_a)}{EP(B_a)} \quad (7)$$

根据上式, 可以求出服务器的综合索引参数, 随着索引参数增大, 综合性能参数也逐渐增大, 综合负载参数则逐渐变小, 这样就可以看作是最好的存储服务器。

3 实验分析

实验分别采用本文所述系统、文献[14]系统、文献[15]系统三种系统进行对比。文献[14]中为使用了大量 flash 存储器的高速数据存储系统, 其优势主要表现在硬件的优化方面。文献[15]中为能够使每个源符号仅参加其所在行的编码的网络编码云存储系统, 但这种系统开销很大, 更新过程响应慢。本文共选取 400 个数据包, 分别对三种系统的存储响应时间进行测试, 结果如图 5 所示。

由图 5 可得, 相比于文献[14]和文献[15]的系统, 本文系统对不同数量的数据包响应时间最少, 对 400 个数量包的响应时间也仅为 40 ms, 而文献[14]和文献[15]的系统分别需要 90 ms 和 105 ms, 可以证明本文系统在大数据存储方面效率较高。继续使用以上三种系统, 对比测试在具体某一时间, 用户使用特定强度向系统发送大数据动态存储指令时, 三种系统服务器实时负载情况如图 6 所示。

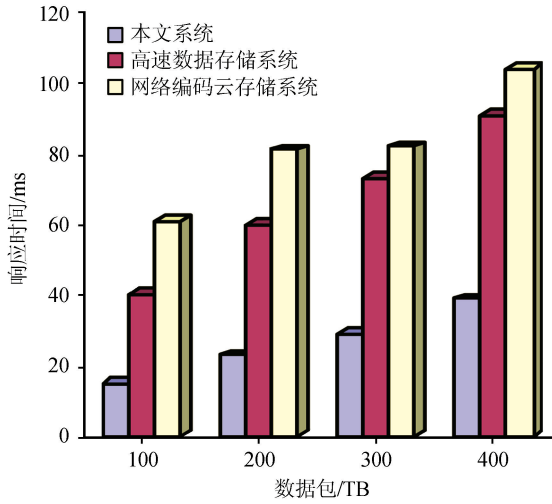


图 5 3 种不同系统存储响应时间对比
Fig.5 Storage response time comparison of three different systems

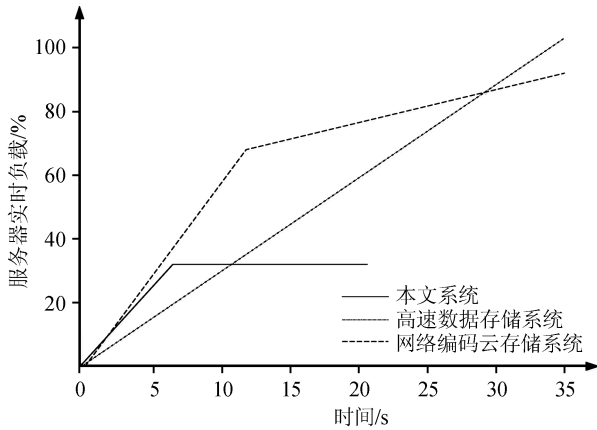


图 6 服务器实时负载对比结果
Fig.6 Comparison results of real time load of three servers

图 6 的对比结果显示, 本文所述系统服务器的实时负载在 7 s 后趋于稳定, 实时负载量基本稳定在 30%, 文献[14]和文献[15]的系统的实时负载都呈逐渐上升趋势, 最大值分别可以达到 100% 和 90%, 以上分析能够说明本文系统的在负载均衡性能方面明显优于另外两种系统。本文通过优化所设计的系统引擎, 增强了操作符的性能, 负载处理更具优势。在上述实验基础上, 对三种系统的存储性能进行进一步测试。选取 4 个不同容量的数据集, 采用三种系统其进行 7 次动态存储, 记录耗时, 所得结果如下。

通过以上三个表的比较分析, 对不同容量的数据集, 本文系统的平均时间消耗为 $(90 \text{ ms} \cdot 100 \text{ ms} + 120 \text{ ms} + 130 \text{ ms}) / 4 = 110 \text{ ms}$, 而文献[14]和文献[15]系统平均耗时分别为 158 ms 和 256.5 ms。

本文所述系统中动态存储的时间最短,效率最高。在实验过程中对三个系统的鲁棒性进行了测试和比较。结果分别如表 4、表 5 和表 6 所列。

表 1 本文系统存储耗时(单位:ms)

Table 1 Storage time of this system (Unit: ms)				
实验次数 /次	数据集 10 GB	数据集 50 GB	数据集 100 GB	数据集 150 GB
1	90	100	120	130
2	93	110	120	128
3	93	100	119	128
4	94	105	119	128
5	87	111	117	128
6	90	100	117	128
7	87	115	119	128
均值	90	100	119	128

表 2 高速数据存储系统存储耗时(单位:ms)

Table 2 Storage time of high-speed data storage system (Unit: ms)				
实验次数 /次	数据集 10 GB	数据集 50 GB	数据集 100 GB	数据集 150 GB
1	125	148	169	190
2	124	143	169	189
3	124	143	170	188
4	125	143	173	188
5	126	148	173	190
6	125	149	173	190
7	126	148	178	195
均值	125	148	170	190

表 3 网络编码云存储系统存储耗时(单位:ms)

Table 3 Storage time of network coding cloud storage system (Unit: ms)				
实验次数 /次	数据集 10 GB	数据集 50 GB	数据集 100 GB	数据集 150 GB
1	205	244	266	311
2	207	241	264	312
3	205	246	265	314
4	207	247	265	314
5	207	248	265	314
6	208	244	266	322
7	210	241	267	321
均值	207	245	278	316

表 4 本文系统鲁棒性测试结果(单位:%)

Table 4 Test results of system robustness in this paper (Unit:%)				
实验次数 /次	数据集 10 GB	数据集 50 GB	数据集 100 GB	数据集 150 GB
1	98.94	98.97	99.96	99.76
2	98.94	98.97	99.97	99.75
3	99.44	98.88	99.86	99.76
4	99.44	99.86	99.85	99.97
5	99.64	99.76	99.76	99.86
6	99.77	99.65	99.65	99.77
7	99.97	99.95	99.53	99.77

表 5 高速数据存储系统鲁棒性测试结果(单位:%)

Table 5 Robustness test results of high speed data storage system (Unit:%)				
实验次数 /次	数据集 10 GB	数据集 50 GB	数据集 100 GB	数据集 150 GB
1	67.45	66.55	66.65	65.31
2	67.34	66.67	66.37	65.51
3	67.55	66.65	66.16	65.33
4	67.54	66.55	66.36	67.24
5	67.45	66.85	66.33	67.33
6	67.88	67.22	66.49	67.35
7	67.85	67.43	67.16	67.56

表 6 网格编码云存储系统鲁棒性测试结果(单位:%)

Table 6 Robustness test results of grid coding cloud storage system (Unit:%)				
实验次数 /次	数据集 10 GB	数据集 50 GB	数据集 100 GB	数据集 150 GB
1	74.53	73.64	73.53	72.54
2	75.53	73.45	73.42	73.44
3	74.44	75.75	73.66	73.56
4	74.55	77.75	73.64	74.23
5	74.14	73.43	72.27	75.25
6	74.16	71.45	72.67	75.43
7	74.87	74.44	74.73	75.44

分析上表中的数据可以看出,当存储数据量为 10 GB 时,本文系统的最大鲁棒性为 99.97%,而文献[14]和文献[15]系统分别仅有 67.88%和 74.87%,比本文系统的最高鲁棒性分别少了 32.11%和 25.10%,当存储数据量为 50 GB 时,本文系统的最大鲁棒性为 99.95%,比其他两个系统分别高 32.52%和 22.21%。当存储数据集为 100 GB 和 150 GB 时,本文系统的最大鲁棒性均为 99.97%。当存储数据量增大时,本文系统的最大鲁棒性也高于其他两个系统的最大鲁棒性,系统的稳定性始终高于 99%,具有较高的稳定性且无过载问题。

4 结语

本文提出的地震信息网络动态存储方法,不仅缩短了大数据动态存储所花费的时间,还在一定程度上提高了动态存储的稳定性。本文提出的动态存储方法能够有效提升地震仪器的性能,在地震勘探的有效开展方面发挥重要作用。

参考文献(References)

[1] 段建华,王云宏,王保利.随采地震监测数据采集控制软件开发[J].煤田地质与勘探,2019,47(3):35-40.
DUAN Jianhua,WANG Yunhong,WANG Baoli.Development of Data Acquisition and Control Software for Seismicmonitoring with Mining[J].Coal Geology & Exploration,2019,47(3):

- 35-40.
- [2] 金维浚,马珊珊,张丽莉,等.基于地震数据构造曲率属性预测裂缝[J].地质科学,2019(2):345-355.
JIN Weijun, MA Shanshan, ZHANG Lili, et al. Fracture Prediction Based on Structural Curvature Attribute of Seismic Data [J]. Chinese Journal of Geology, 2019(2): 345-355.
- [3] 娄莉,李勇,等.改进光流算法在地震图像纹理分析的应用研究[J].哈尔滨工业大学学报,2019,51(11):47-54.
LOU Li, LI Yong, et al. Application Research of an Improved Optical Flow Algorithm in Seismic Image Texture Analysis [J]. Journal of Harbin Institute of Technology, 2019, 51(11): 47-54.
- [4] 史子乐,黄艳林,李静,等.陆上地震数据采集节点设备现状与发展[J].石油管材与仪器,2019,5(1):14-18.
SHI Zile, HUANG Yanlin, LI Jing, et al. Status and Development Trend of Onshore Seismic Nodal Instrument [J]. Petroleum Instruments, 2019, 5(1): 14-18.
- [5] 姚秀光,刘东烈,郭金城,等.高频 GPS 数据处理方法比较及在地震学中的应用[J].测绘地理信息,2018,43(5):42-46.
YAO Xiuguang, LIU Donglie, GUO Jincheng, et al. Comparison of Data Processing Methods of High-Frequency GPS and Its Application to Seismology [J]. Journal of Geomatics, 2018, 43(5): 42-46.
- [6] 李芳菊.基于 Hadoop 的网络行为大数据安全实体识别系统设计[J].现代电子技术,2019,42(17):7-9.
LI Jufang. Design of Big Data Security Entity Recognition System for Network Behavior based on Hadoop [J]. Modern Electronics Technique, 2019, 42(17): 7-9.
- [7] 张圣杰.分布式大数据管理系统的设计与实现研究[J].信息通信,2019(2):217-218.
ZHANG Shengjie. Design and Implementation of Distributed Big Data Management System [J]. Information & Communications, 2019(2): 217-218.
- [8] 刘倍雄,张毅,陈孟祥.工业大数据时序数据存储安全预警研究[J].机械设计与制造工程,2019,48(5):62-66.
LIU Beixiong, ZHANG Yi, CHEN Mengxiang. Research on Safety Early Warning of Time Series Data Storage for Industrial Big Data [J]. Machine Design and Manufacturing Engineering, 2019, 48(5): 62-66.
- [9] 张俊玲.大数据时代计算机网络信息安全与防护措施研究[J].信息技术与信息化,2019(4):130-132.
ZHANG Junling. Research on Computer Network Information Security and Protection Measures in the Era of Big Data [J]. Information Technology & Informatization, 2019(4): 130-132.
- [10] 季益龙,杨传斌.一种校园混合云存储系统的设计与实现[J].浙江师范大学学报(自然科学版),2019,42(2):156-160.
JI Yilong, YANG Chuangbing. On the Design and Implementation of a Campus Based Mixed Cloud Storage System [J]. Journal of Zhejiang Normal University Natural Sciences, 2019, 42(2): 156-160.
- [11] 陈敬德,盛戈崑,吴继健,等.大数据技术在智能电网中的应用现状及展望[J].高压电器,2018,54(1):35-43.
CHEN Jingde, SHENG Gehao, WU Jijian, et al. Application Status and Prospect of Big Data Technology in Smart Grid [J]. High Voltage Apparatus, 2018, 54(1): 35-43.
- [12] 潘力.激光数据存储系统的设计与实现[J].激光杂志,2018,39(5):3-4.
PAN Li. Design of Dynamic Storage System for Lidar Data Under Cloud Computing [J]. Laser Journal, 2018, 39(5): 3-4.
- [13] 侯捷.云计算海量激光雷达数据动态存储系统设计[J].科技通报,2018,34(9):201-205.
HOU Jie. Design of Dynamic Data Storage System for Massive Lidar Data Under Cloud Computing [J]. Bulletin of Science and Technology, 2018, 34(9): 201-205.
- [14] 易志强,韩宾,江虹,等.基于 FPGA 的多通道同步实时高速数据采集系统设计[J].电子技术应用,2019,45(6):6-8.
YI Zhiqiang, HAN Bin, JIANG Hong, et al. Design of Multi-channel Synchronous Real-time High-speed Data Acquisition System based on FPGA [J]. Application of Electronic Technique, 2019, 45(6): 6-8.
- [15] 方欣,方睿,刘雪涛,等.基于链表多分支路径树的云存储数据完整性验证机制[J].成都信息工程大学学报,2018,33(1):44-49.
FANG Xin, FANG Rui, LIU Xuetao, et al. Data Integrity Verification Mechanism based on SLBT Tree [J]. Journal of Chengdu University Of Information Technology, 2018, 33(1): 44-49.
- [16] 孙晓莉,田淑静,万荣.基于 RS 和 GIS 的土地利用动态变化分析——以云南国土资源职业学院周边区域为例[J].贵州大学学报:自然科学版,2018,35(2):117-120.
SUN Xiaoli, TIAN Shujing, WAN Rong. Analysis of Land Use Dynamic Change Based on RS and GIS: A Case Study of Yunnan Land and Resources Vocational College Surrounding Area [J]. Journal of Guizhou University (Natural Science), 2018, 35(2): 117-120.
- [17] 张苏,周小成,黄洪宇,等.基于 SVR 的 GF1 号遥感影像森林蓄积量估测[J].贵州大学学报(自然科学版),2019,36(3):21-26.
ZHANG Su, ZHOU Xiaocheng, HUANG Hongyu, et al. Estimation of Forest Volume of GF1 Remote Sensing Image Based on SVR [J]. Journal of Guizhou University (Natural Science), 2019, 36(3): 21-26.
- [18] 易秀龙.基于逐步区间删失数据 Rayleigh 分布的参数估计[J].贵州大学学报(自然科学版),2018,35(6):32-36.
YI Xiulong. Parameter Estimation of Rayleigh Distribution Based on Progressive Interval Censored Data [J]. Journal of Guizhou University (Natural Science), 2018, 35(6): 32-36.