International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.331-340
http://dx.doi.org/10.14257/ijdta.2016.9.12.29

Design and Implementation of Hadoop-based Customer Marketing Big Data Processing System

Yang Liu

Institue of Applied Electronics, ChongQing College of Electronic Engineering, ChongQing 401331, China
E-mail: liuyang1983cq@163.com

Abstract

The era of big data has become the core competitiveness of enterprises, which is an important business capital with supporting data. Companies can truly benefit undefeated. Multi data also have cultural change, organizational models and even business. This requires that companies use to analyze consumer data on consumer behavior. This article study and implementation of a distributed architecture platform Hadoop big data applications (Web data mining and processing platform), the use of Hadoop massive data processing capacity and strong elastic computing capacity expansion. Data processing and analysis will be raised to new heights. For this type of paper, at the same time focus on the distributed cluster data storage, processing capacity optimization and key technology research. The reliability and validity were analyzed to determine the data collected in this study. The questionnaire is effective, which can be used in subsequent studies. Then correlation analysis and regression analysis model and hypothesis were tested. Finally, the model of precision marketing strategy put forward recommendations.

Keywords: *Index system; C2C mode; perceived risk; data mining; precision marketing*

1. Introduction

Big data has changed people's way of life and mode of thinking, it is to generate a source of new inventions and creativity, and create a new era [1]. Data are given a new value, no meaning except that the data itself contains. It developed a potential value of the data contained, and that it applies to the future, which beyond its original basic role as a powerful weapon in the future [2, 3]. This is a radical change of epoch-making significance. It enables enterprises to re-recognize the data, and the use of data has become more clear thinking, use of data analysis of the results obtained, it applied to the development of enterprises among the innovative business model [4, 5]. The era of big data, data has become the core competitiveness of enterprises is an important business capital, with supporting data, companies can truly benefit undefeated. Multi data will also have a cultural change, organizational models and even way business is done. This requires that companies use to analyze consumer data on consumer behavior [6].

A basic big data application system must provide the basic data mining algorithm interface. In the Hadoop platform can not avoid reduce is a Subsidiary; data layer in how to HDFS and HBase optimization also need to focus on; also, a need for high reliability the system architecture for high performance requirements under, hardware and network optimization aspects must also be considered [7]. In short, to build a reliable, high-performance large-data applications, we need to study based on and solve many problems. These problems are also of great significance, a mature system platform that can provide great support after the deployment and expansion of the basis of its applications [8].

Big Data is a new industry based on modern scientific and technological progress generated, which fully reflects the development of modern technology. It is an

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2016 SERSC

application-oriented service. Enterprises should pay full attention to big data brings value, the size and quality of statistical data in the enterprise, large data mining and analysis, development and large data consistent marketing strategy. Enterprise Big Data should be viewed as a means of production, be fully utilized, and the analysis of the data used in the development of precise aspects of business marketing strategy. Big data enables the study of consumer behavior more in-depth and efficient, able to accurately predict consumer demand and deeper consumer insight. Precise marketing, there are also opportunities and challenges, to take full advantage of big data to improve marketing efficiency. This article is the study and implementation of a distributed architecture platform Hadoop big data applications (Web data mining and processing platform), the use of Hadoop massive data processing capacity and strong elastic computing capacity expansion, data processing and analysis will be raised to a new heights. For this type of paper at the same time focus on the distributed cluster data storage, processing capacity optimization and key technology research and analysis.

2. Application of System Architecture Analysis and Overall Design

2.1 HDFS Architecture

HDFS called the Hadoop Distributed FileSystem. HDFS data access to streaming mode to store large files (currently there are already stored data reaches PB level Hadoop clusters) [9, 10]. Hadoop does not require expensive and highly reliable hardware that can run on commodity hardware cluster. Therefore, in large Hadoop cluster node failure is not uncommon. However, because HDFS is highly fault-tolerant design, fault was rarely perceived by the user [11, 12].

Hadoop used Master-Slave structure model. HDFS by NameNode (Master) and a plurality of DataNode (Slave) components, NameNode server is mainly responsible for the data, the file system and access the data, Client control. DataNode then stored as data from node responsible. HDFS architecture was shown in Figure 1:

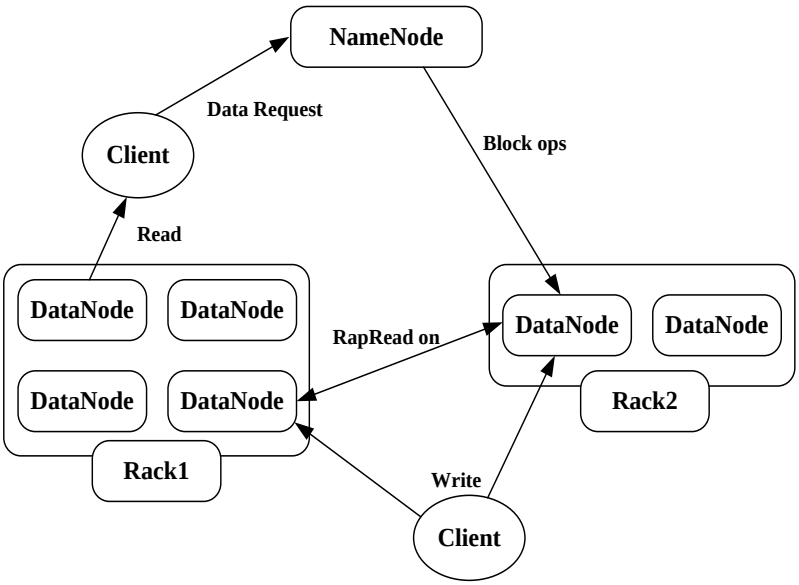

Figure 1. HDFS Architecture

HDFS by the Java developers, it can be deployed on any Java-enabled machine. As can be seen from the chart, the general cluster by a single machine running NameNode, each of the other machine is running 1 ~ N a DataNode. User data will not be saved via

 Copyright ⓒ 2016 SERSC

NameNode, NameNode only responsible for managing the file system and metadata specific data is the responsibility of the DataNode [13].

2.2 Consumer Behavior Insights

Big data regardless of time, regardless of location will be unrestricted access to data can be collected to the desired data in the case of unavailable data, which no limit is ever convenient ways and means can not be done [14, 15]. Big data era such consumer insights developed to a new stage in the past was no conclusion can be obtained using a large data analysis, data acquisition, storage, analysis and utilization to provide different levels and a deeper degree of recommendation (Figure 2). For example, network search platform has a huge consumer search for information, companies can conduct searches based on consumer search platform to provide consumer insight, which is based on a new technology for big data insights through the analysis of consumer search data real-time insight into consumer demand and analyzing the demand, and ultimately effective marketing strategy [16].

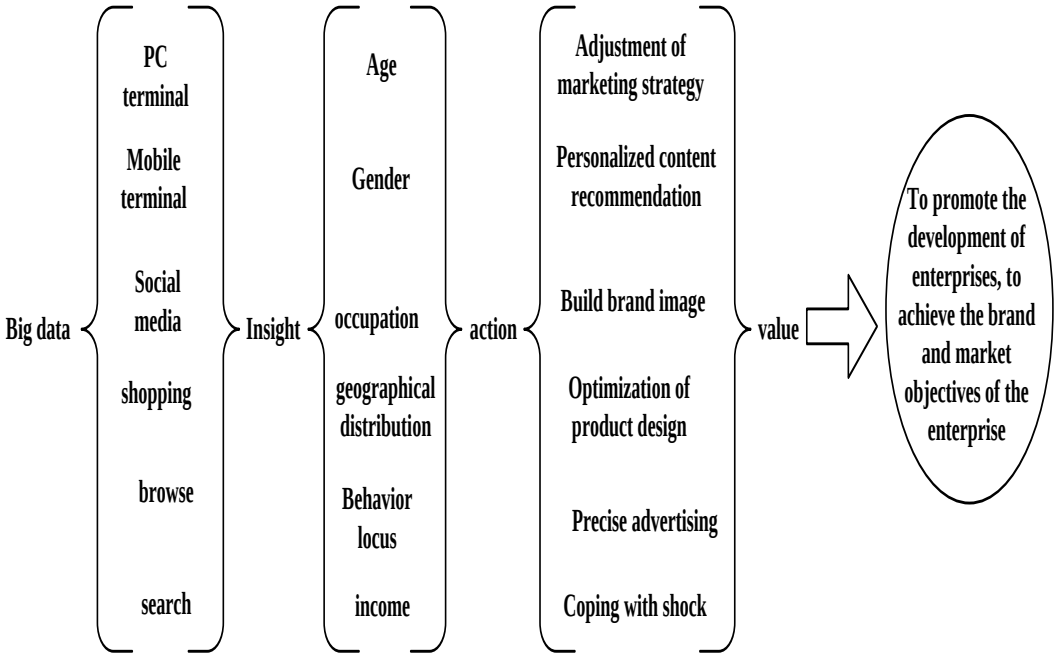

Figure 2. Consumer Behavior Insights Large Data Process

Comprehensive data acquisition is a fundamental characteristic of big data insights. Through these comprehensive data, we can ensure comprehensive data analysis, so that it can ensure that the results of data analysis and the actual situation is more close. As long as scientific and rational use of these vast amounts of data, it can not limit the time and space to be manifested in various ways by these data. Data analysis based on real conclusions to be consumer insight, and insight into the results and other business enterprises combined results conducive to enterprise development business strategy.

Experience emphasizing that consumers should be able to enjoy a variety of experiences obtained in the consumption process, if you can not personally participate in activities to go, but only as a spectator, you can not feel the joy of shopping and satisfaction. Consumers in the process of experience, it is possible to simultaneously use a variety of sensory feelings active process, thus enhancing the participation of so impressed with the product. Enterprises can also experience the process of recording consumers to experience-based marketing.

Copyright ⓒ 2016 SERSC

2.3 Research on Consumer Decision-making Mechanism

One theory for the analysis of consumer decision-making is often used in consumer value theory. More represent are delivered value theory of raised. It is delivered value of total consumer revenue minus total costs. Only when the value is positive, consumers will choose to buy, and transferring the greater value. Its willingness to make a purchase will be more intense. You can use the following function expressed its algorithm:

$$P = Max(CDV_i) = Max(TCV_i - TCC_i), (i = 1, 2, \ldots\ldots, n) \quad (1)$$

$$TCV = f(x_1, x_2, x_3, \ldots\ldots) \quad (2)$$

$$TCC = f(y_1, y_2, y_3, \ldots\ldots) \quad (3)$$

Wherein, P represents total revenue decision-making program, i represents the number of possible influencing factors, CDV indicates customer delivered value, TCV represents the total customer value, TCC represents the total cost of the customer. x1, x2, x3 represent factors that affect overall customer value , y1, y2, y3 represents factors influence the total cost of the customer.

The past and present into a summary, as the total cost of the consumer, then the consumer benefits and costs are compared to measure the results of this decision. With the formula expression can be expressed as:

$$P = F(X) - F(Y) \quad (4)$$

Wherein, F (x) represents consumers on account of something psychological, F (Y) represents delivered value to consumers. P is greater than zero when consumers tend to choose to buy the P in the vicinity of zero will think then choose whether to buy, and when P is much smaller than zero, consumers will not buy. Therefore, the study of mental accounting purchasing decisions of consumers had an important role.

As shown in Figure 1, consumer behavior with different social classes is significantly different. First, in terms of shopping, consumers of different social classes, in the choice of shopping venues tend to correspond with their status of shopping venues. Second, product selection in which the different social classes, for goods and brand preference is different, especially the performance of the clothing, furniture and residential. Status can be displayed in the product. The top consumer of information acquisition and transmission than the lower staff, they are often in the forefront of fashion and trends.

Big data environment is based on a new type of digital resources, information and internet. Use big data personalize and precision marketing is one of the characteristics of the era of big data, especially for online shopping undecided population. The need for more accurate recommendation system, and enjoy a virtual experience. However, when consumers use the big data marketing, consumers will feel the risk felt. It produces large data processing benefits, such as disclosure of personal information, privacy. In addition, consumer confidence in the large data environments, C2C model is also an important issue, C2C mode will have an important influence consumer willingness to buy. Therefore, the present study of three intermediate variables included in the model perceived usefulness, perceived risk and trust.

 Copyright ⓒ 2016 SERSC

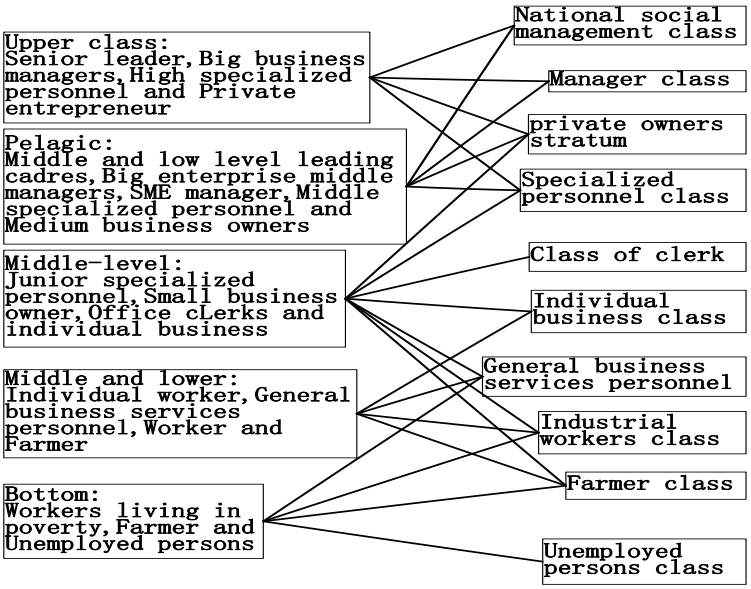

Figure 3. Social Classes and Social Classification Diagram

3. Modeling and Description

3.1 C2C Mode of Consumer Behavior Research

Big data environment mode is changing with the times. Research on consumer behavior in the past under the traditional environment is divided into two processes (buying decisions and actual behavior). The big data environment, consumer behavior by other external influences is more and more significant. It changed the internet age lifestyles for consumers. The large consumer behavior data and C2C mode docking analysis and will get considerable value excavation, which was shown in Figure 4:

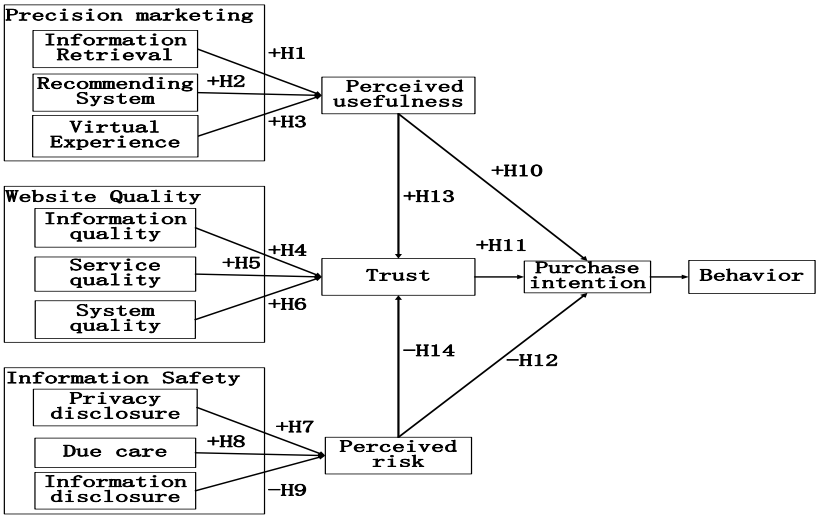

Figure 4. Big Data and Consumer Behavior Patterns C2C Process Diagram

Similarly, the role of big data on consumer behavior was during the docking of large data and behavior. Through data processing techniques, big data mining and consumer-generated process through the extracted C2C mode of data analysis based on tailored marketing, in order to take the results of analysis.

Copyright ⓒ 2016 SERSC

3.2 Consumer Behavior Supports Accurate Marketing Decisions

The first to collect and collate consumer information, excluding data does not comply with the rules and build consumer databases. Followed by mining data, classify the different consumers, different consumer analysis to identify the differences among consumers. Finally, the marketing campaign evaluation and feedback analysis the results of marketing activities. To learn more about the real needs of consumers, consumers predict future demand, the formation of a new marketing strategy. Consumer behavior model supported precision marketing system was shown in Figure 5.

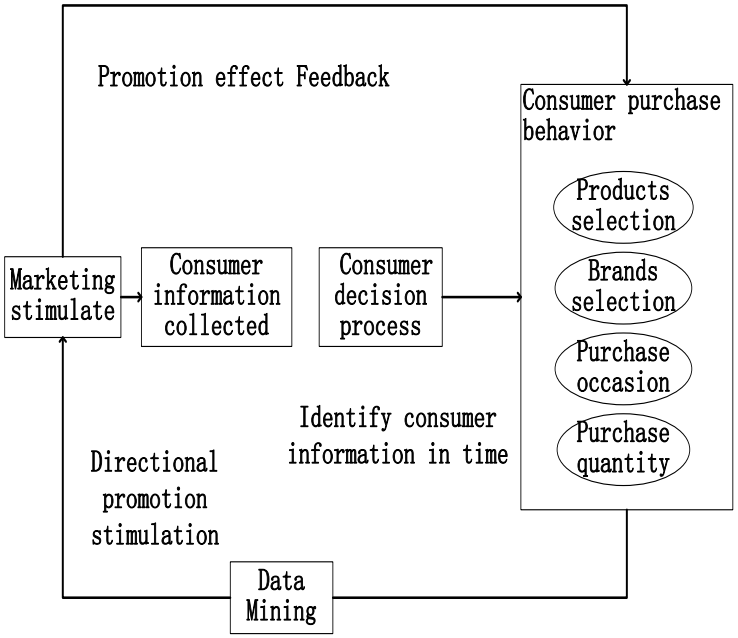

Figure 5. Model of Precision Marketing System Interact

Study of consumer decision-making behavior is focus on consumer. Many companies have spent a lot of effort in this regard. Consumers use shopping carts to collect consumer information, the establishment of consumer decision-making behavior model can study consumer behavior. Many shopping sites based browsing traces of consumers, the establishment of consumer behavior model suggested that it may buy products mining consumer behavior.

4. Experiment and Results

4.1 Data Acquisition and Pre-processing

Due to the different UGC and questionnaires can not accurately describe the property, and each comment can not contain all the attributes of the content, we comment on the digital processing is the most important part of the pre-treatment. We will feature user reviews of the property value range is defined as the product of 0-1. We determine the value of the characteristic properties of the user comments on the emotional bias characteristic attributes, reference statistical surveys way, if the user characteristics of a positive review (affirmative word), we believe that the value of the property feature full frontal: the contrary was entirely negative: for the UGC content not related to the characteristics of the properties, we believe that the user's emotional bias is neutral, defined as 0.5. Each sample in the form of digitized as follows:

$$Sample = (x_1, x_2, \ldots x_n)^T (n = 26, x_i = 0 / 0.5 / 1) \quad (5)$$

 Copyright ⓒ 2016 SERSC

Use normalization method statistical information into a value between 0 and 1. Referring to the digital standard mentioned above, we believe represents a full frontal; 0 represents totally negative; 0.5 is neutral. Normalization method can be stated as follows:

$$AttVal(x) = 0.5 + \frac{AttNum_P(x) - AttNum_n(x)}{TotalNum(x)} \quad (6)$$

In the above formula AttNmnp attributes representing the number of positive emotion, attribute indicates the number of negative emotions, TotalNmu appears showing the total number of properties.

Data cleansing is aimed at improving the quality of data, so that the data can be analyzed to achieve the standard. Cleaning data including data integrity and consistency checks on the selected subset of data cleaning, remove duplicate data, to estimate missing values.

Prosperity of the market is a key factor in increasing consumer demand for personalized. The pursuit of the basic properties of the goods has been unable to meet the growing consumer demand, has become a modern consumers for goods and services have a higher demand for mature consumers. For example, it has been satisfied with simple food tasty or not, while more concerned about health factors can provide food, taste requirements should also have the role of safety care; daily necessities for their products not only to meet the basic functions, but also requires natural environmentally friendly materials; the demand for electronic products can not only durable, but easy to use, beautiful appearance. Goods to be able to reflect the individual emotional factors and consumer self-expression, which will become an important factor affecting consumer purchase.

4.2 Large Amount of Test Data Training

Based on the same data multiple iterations of training, we can see that in the same system environment, and training time is directly proportional to the number of iterations basic, in order to reduce interference with other networks, and other factors, we use a single network node to obtain simulated training table 1 result.

Table 1. Model Single Point Training Time

	25M Data single point time (seconds)
Iteration 10 times	94
Iteration 100 times	913
Iteration 200 times	1475
Iteration 500 times	4301

Training time is proportional to the number of iterations, as shown in Figure 6 (ordinate is time-consuming exercise, in minutes):

Copyright ⓒ 2016 SERSC

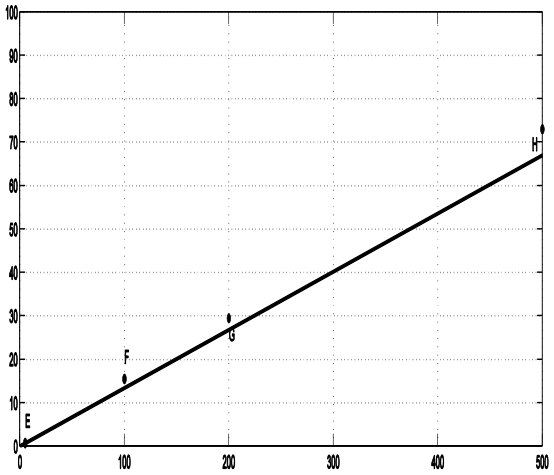

Figure 6. Training Time is Proportional to the Number of Iterations

Having reached the relationship between training time and the number of iterations, we use more data to test the parallel speedup algorithm to detect whether the algorithm has high scalability. Experimental use 100G data for testing. According to the results above, the training time and the number of iterations is linear, so the data in Table 1 describes only the first iteration when used.

Through this process, companies can collect the available real and reliable consumer information, including consumer brand purchase, the amount of consumption and propensity to buy. Data miners can be analyzed according to these data, consumer buying mining decision factors, the consumer budget, consumer attitudes and purchase order of priorities, and the consumer will be classified accordingly to develop targeted precision marketing strategy.

4.3 Reliability Analysis

In general, Cronbach is a coefficient between 0 and 1, if Cronbach's a coefficient larger then the internal consistency, reliability is possible. Cronbach's coefficient of 0.60-0.65 indicates the letter is not good; 0.65-0.70 is acceptable validity values; Cronbach's coefficient 0.70-0.80 represents it very good reliability; Cronbach's coefficient of 0.8 or more indicates the reliability is very good. Therefore, the research CITIC analysis preclude the use of internal consistency reliability test methods. The use of statistical software SPSS19.0 calculated Cronbach's a value. Reliability overall scale and subscales analysis results were shown in Table 2:

Table 2. The Cronbach's a Coefficients of Each Variables

Variables	Coefficients	Quiz numbers	Cronbach's a
Precision marketing	Information Retrieval	3	0.709
	Recommending System	3	0.789
	Virtual Experience	3	0.769
	Total Scale	9	0.708
Website Quality	Information quality	3	0.712
	Service quality	3	0.778
	System quality	3	0.709
	Total Scale	9	0.733

 Copyright ⓒ 2016 SERSC

Information Safety	Privacy disclosure	3	0.722
	Due care	3	0.700
	Information disclosure	3	0.831
	Total Scale	9	0.735
Purchase intention	Perceived usefulness	3	0.711
	Trust	3	0.709
	Perceived risk	3	0.744
Purchase intention	Purchase intention	3	0.775
Questionnaire overall scale		39	0.756

The data in Table 2 showed that the total amount of the Cronbachs coefficient is 0.756, indicating high reliability. Purchase intent Cronbach a coefficient was 0.775, more than 0.7, indicating that reliability is quite good. Big data screening site quality Cronbach a coefficient of 0.700, more than 0.7, indicating that reliability is quite good. Big data mining privacy Cronbach a coefficient of 0.750, more than 0.7, indicating reliability is quite good. The perceived usefulness Cronbach a coefficient of 0.736, more than 0.7, indicating that reliability is quite good. Trust a coefficient of 0.744, more than 0.7, indicating that reliability is quite good. Perceived risk Cronbach a coefficient was 0.806, more than 0.8, it indicates the reliability is very good.

5. Conclusions

Big data has changed people's way of life and mode of thinking, it is to generate a source of new inventions and creativity, and create a new era. Data are given a new value, no meaning except that the data itself contains. It developed a potential value of the data contained, and that it applies to the future, which beyond its original basic role as a powerful weapon in the future. This article is the study and implementation of a distributed architecture platform Hadoop big data applications (Web data mining and processing platform), the use of Hadoop massive data processing capacity and strong elastic computing capacity expansion, data processing and analysis will be raised to a new heights. For this type of paper at the same time focus on the distributed cluster data storage, processing capacity optimization and key technology research and analysis.

References

[1] R. Fontugne, P. Borgnat, P. Abry and K. Fukuda, “Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking”, Proceedings of the 6th International Conference. Co-NEXT’10 ACM, New York, NY, USA; doi:10.1145/1921168.1921179, **(2010)**, pp. 8-1812.
[2] W. S. Baek, D. M. Kim, F. Bashir and J. Y. Pyun, “Real life applicable fall detection system based on wireless body area network”, In Consumer Communications and Networking Conference (CCNC), IEEE, **(2013)**, pp. 62-67.
[3] P. Zhou, Z. Yang, L. Li and S. Qiu, “Application of Big Data Processing Technology in the Intelligent Network Management System”, In Web Technologies and Applications, Springer International Publishing, **(2015)**, pp. 26-34.
[4] R. Bhatti, R. LaSalle, R. Bird, T. Grance and E. Bertino, “Emerging trends around big data analytics and security: Panel”, In Proceedings of the 17th ACM Symposium on Access Control Models and Technologies. SACMAT’ 12. ACM, New York, NY, USA, doi:10.1145/2295136.2295148, **(2012)**, pp. 67-68.
[5] R. M. Perianu, M. M. Perianu, P. Havinga, S. Taylor, R. Begg, M. Palaniswami and D. Rouffet, “A performance analysis of a wireless body-area network monitoring system for professional cycling”, Personal and Ubiquitous Computing, vol. 17, no. 1, **(2013)**, pp. 197-209.
[6] R. Berger, “Big data–from buzzword to strategy Information Management Functional know how Expertise Roland Berger”, **(2015)**.
[7] M. M. Vernon, B. Ulicny and D. Bennett, “An Information Provider's Wish List for a Next Generation Big Data End-to-End Information System”, InCIDR, **(2015)**.
[8] T. Lupo, “A fuzzy ServQual based method for reliable measurements of education quality in Italian higher education area”, Expert systems with applications, vol. 40, no. 17, **(2013)**, pp. 7096-7110.

Copyright ⓒ 2016 SERSC

[9] H. Xu, L. Wang and W. Gan, "Application of Improved Decision Tree Method based on Rough Set in Building Smart Medical Analysis CRM System", **(2016)**.
[10] M. E. Mastrangelo, F. Weyland, S. H. Villarino, M. P. Barral, L. Nahuelhual and P. Laterra, "Concepts and methods for landscape multi-functionality and a unifying framework based on ecosystem services", Landscape ecology, vol. 29, no. 2, **(2014)**, pp. 345-358.
[11] O. Diallo, J. J. Rodrigues, M. Sene and J. Niu, "Real-time query processing optimization for cloud-based wireless body area networks", Information Sciences, vol. 284, **(2014)**, pp. 84-94.
[12] C. C. G. Hsieh, "Building a Cloud Computing and Big Data Infrastructure for Cybersecurity Research and Education (No. 64713-CS-REP. 2)", Norfolk State University Norfolk United States, **(2015)**.
[13] M. J. Kim and Y. S. Yu, "Development of Real-time Big Data Analysis System and a Case Study on the Application of Information in a Medical Institution", International Journal of Software Engineering and Its Applications, vol. 9, no. 7, **(2015)**, pp. 93-102.
[14] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Communications of the ACM, vol. 51, no. 1, **(2008)**, pp. 1-13.
[15] D. X. Ge and C. C. Gao, "Value Network Module of Creative Industries Cluster Based on Big Data Analysis", Revista de la Facultad de Ingeniería, vol. 31, no. 6, **(2016)**, pp. 43-52.
[16] A. Ayanso and D. Visser, "Analytics and Performance Measurement Frameworks for Social Customer Relationship Management. Social Media and Networking: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications", **(2015)**, pp. 252.

Author

Yang Liu. She received her bachelor's degree in computer science and technology specialty from Chongqing University of Posts and telecommunications in 2007. Currently, she is an experimentalist and work in Chongqing College of Electronic Engineering. Her current research interests include software engineering, electronic information technology.

 Copyright ⓒ 2016 SERSC