Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2018.22581

122 (2018) 36–41
August

Key technologies of a distributed and unstructured water resources big data system

Yuan Dong[a,b,c,d], D. Xiao[e,*], BaoQing Hu[f], ShiLun Zhang[c], JiaHai Liang[a], GuoCai Nong[a], ZhiXian Liu[a], RongYang Zhao[a], MeiXing Liu[b], ZhenHua Xu[c], Jin Tao[c], Kai Deng[c], Li Zhou[c], Xin Han[c]

[a]*Key Laboratory for Advanced Technology to Internet of Things, Qinzhou University, Qinzhou, 535011, China*
[b]*Guangxi Colleges and Universities Key Laboratory of Complex System Optimization and Big Data Processing, Yulin Normal University, Yulin 537000, China*
[c]*College of Resources and Environment, Qinzhou University, Qinzhou, 535000, China*
[d]*Faculty of Information Engineering, China University of Geosciences, Wuhan, 430074, China*
[e]*Imperial College London, London, UK, email: dh.xiao@imperial.ac.uk*
[f]*Key Laboratory of Environment Evolution and Resources Utilization in Beibu Bay (Guangxi Teachers Education University), Ministry of Education, 530001, Guangxi, China*

Received 24 February 2018; Accepted 04 June 2018

ABSTRACT

In recent years, with the impact of climate change and human activities, the spatial and temporal distribution of water resources change significantly. Data storage, computing and analysis of water resources data have become increasingly difficult. The traditional data storage and management technology are very difficult to handle current water resources data. In this case, the big data technology is used to tackle this challenging water resources management issue in the social and economic development. This paper, for the first time, addresses three key issues of designing and implementing distributed water resources big data management system. The performance and advantages of water resources big data system are evident in terms of data access, computing and analysis speed.

Keywords: Water resources; Distributed; Big data; Spatial

1. Introduction

Water resources are one of the most important resources to human beings and necessary for a wide range of areas such as agricultural, industrial, household, recreational and environmental activities. All living creatures require water to grow and reproduce. Therefore, having a good management of water resources is vital. However, the traditional water resources management is not able to manage such a huge amount of data efficiently since the data is high volume, high velocity, and high variety (3Vs) [1]. The data volume refers to the huge amount of data, velocity relates to the speed of data accessing and processing, and variety refers to the various types of data. In this case, we need a more powerful tool to handle such huge amount of heterogeneous complex dataset efficiently. In this case, big data technology is presented to tackle this challenging issue recently and received great attention [2].

This is an era of big data and the big data have been penetrating into our everyday life [3]. It has been applied successfully into several application areas such as biology [4], business management [5], etc. Water resources data involve a lot of temporal and spatial data, and also has a long time-scale

* Corresponding author.

Presented at the 3rd International Conference on Recent Advancements in Chemical, Environmental and Energy Engineering, 15–16 February, Chennai, India, 2018.

making it difficult to manage. Current development of big data technology provides a feasible way of managing it efficiently. In this paper, we apply the big data technology into water resources management system to handle the data efficiently. In particular, we give details of the water resources data, storage design and distributed computing.

The structure of the paper is as follows: Section 2 describes water resources management data. Section 3 describes the key technologies of designing big data water resource system, which involves storage design, distributed computing and structures. In Section 4, the performance and advantages of this water resources big data system are given. The conclusion is given in Section 5.

2. Water resources management data

2.1. Basic data

Water resources management aims to achieve the reasonable development of water resources, comprehensive management, optimization, comprehensive conservation and effective protection. The key factor is timely and accurate grasp of water resources and water resources development, utilization and protection. Accurately grasping the characteristics and rules of the changes require strong basic data supports [6]. The source of data, functions and timeliness of the fundamental data is limited.

The data mainly include:

(1) Monitoring data such as: water resources, water environment, water ecology and rainfall forecasting data;
(2) Total water usage data, including surface and groundwater water usage, water allocation, water intake licensing, water resources scheduling, water users and water rights trading data;
(3) Water usage efficiency data, including water saving indicators, water usage planning indicators, water quota, water usage efficiency, unconventional water and other data;
(4) Water active areas data, including water active zones, water area acceptance capacity assessment, water quality monitoring and evaluation of water level, investigation and monitoring of river outfall, drinking water source protection, etc.;
(5) Water resources economic accounting data, including total water consumption and sewage discharge accounting data, water resources fee accounting, water supply fee accounting, water rights transfer transaction accounting, ecological compensation standard quantitative accounting and other data.

2.2. Association and spatial data

Different water conservation and environmental protection sectors obtain and store their data (only covering their own territories), respectively. The sources of data are various: satellite remote sensing data, global positioning system data, geographic information system data, wireless sensor network data and other modern measurement technology data [7]. At present, the mechanism of data sharing mechanism between the various sectors has not formed. In order to adopt the data and make it beneficial to our society and economy, a united platform of water resources management information system for various departments is needed [8]. This platform can be used to manage the integration of water data, water quality, water supply, water entry–exit scale, groundwater overdraft, dumping sewage into the river mouth, real-time monitoring and data sharing.

In addition, some data (including structured, unstructured and semi-structured data) are scattered at different institutions, enterprises, factories and irrigation management sectors, residential and industrial parks. For data of those type, we use mobile internet, social networking and cloud computing tools to integrate them into a flexible and open high-performance platform [9].

At present, industrial sectors use traditional database system to manage the data, which has a unified structure. This unified data are structured data. However, with the development of the internet, various data with different data formats come into our life such as: graphics, images, voice video, spreadsheets and other multimedia information [10]. This kind of information cannot be expressed in a digital or unified structure, and it is another data: unstructured data. The water resources data also include structured and unstructured data such as geography information spatial data.

2.3. Data features

The water resources spatial data mainly include 26 object types. The vector data include eight themes including rivers and lakes, water conservancy projects, economic and social water usage, river and lake development and conservation, water and soil conservation, water conservancy industry capacity, irrigation areas, and groundwater wells. Also, it includes the three-level water resources regionalization and watershed regionalization elements, which in total has 44 regionalization elements. The data integration model is shown in Fig. 1; the overall planning model hierarchy is shown in Fig. 2.

3. Key technologies of water resources big data system

The basic idea of constructing a distributed spatial unstructured water resources big data system has three key aspects: columnar storage, distributed storage and distributed computing. The data can be broadly divided into attribute data and spatial data. The storage of spatial data is achieved using ArcGIS system and the attribute data can be stored in database systems.

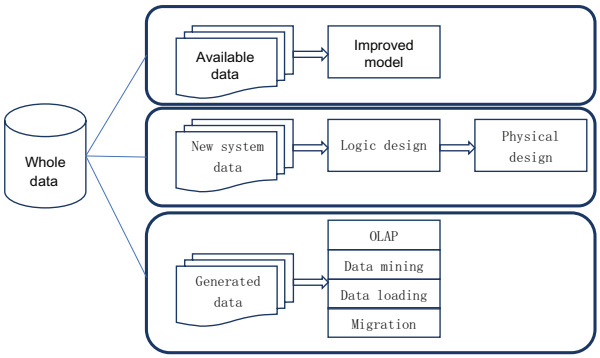

Fig. 1. Data integration model.

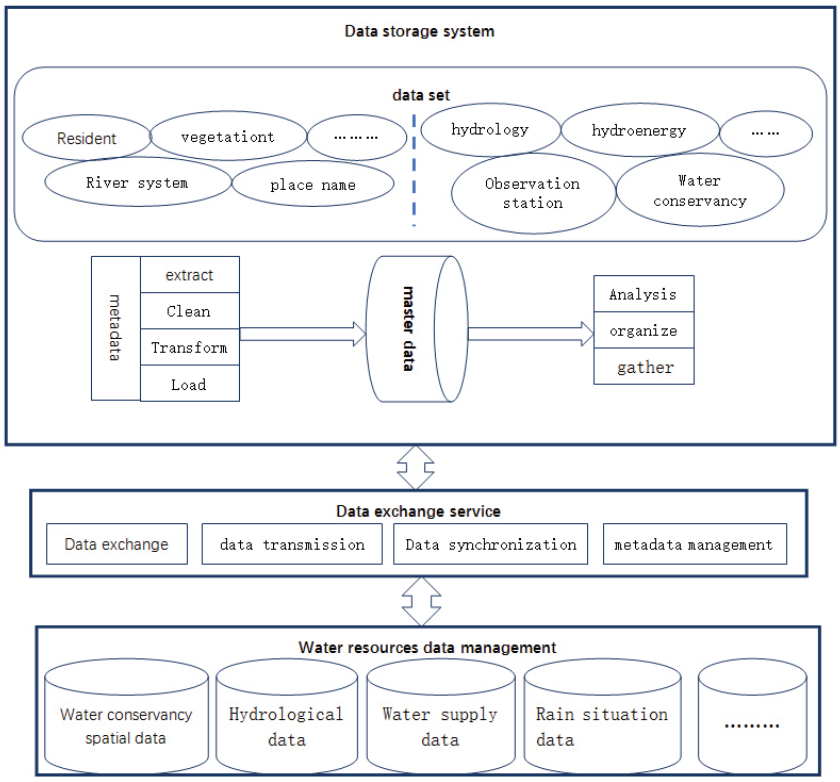

Fig. 2. Hierarchical data model.

3.1. Columnar storage

The distributed spatial unstructured water resources big data system adopts columnar storage to store the basis data mentioned in Section 2. The columnar storage facilitates the accessing speed of the water resources basic data.

The columnar storage has quicker reading speed than row storage. Unlike row storage, the columnar storage takes fields as clustering units and reads the corresponding columns directly, and therefore, there is no redundant column. Additionally, water resources data in each column are homogeneous, and there is no two-semantic issue. Compared with row storage, the columnar storage has lower storage efficiency since the disk head is required to move to the corresponding positions when writing multiple rows. However, the columnar storage still has great advantages. For example, the columnar storage makes the water resources big data system easier to parse the water resources basic data. Also, it is easier to design a better compression/decompression algorithm for the water resources basic data [11]. Row and column storages are shown in Fig. 3.

In order to improve the storage efficiency of columnar storage, a number of methods have been used. First, multiple hard disks with multi-threaded parallel I/O is used. The multi-threaded parallel data reading and writing technology reduces disk reading and writing contentions effectively. In addition, a rollback mechanism which is similar to relational database is considered. Using the rollback mechanism, the previously written data all fail when a column write fails. The hash code is also considered to further ensure the integrity of water resources basic data.

3.2. Distributed storage

Various water resources data are distributed in different sectors at different locations; therefore, an efficient data integration system has to be designed to manage data at different sectors and locations. In this case, distributed storage is used to manage the water resources basic data.

The distributed file system is able to operate a large number of servers of different water resources sectors to manage data. We do not feel storing and reading the water resources data on different servers (located at different water resources sectors) when we use the distributed file management technology. Essentially, the distributed file system manages a thing called a server cluster. In this server cluster, the water resources data are stored in the cluster node (the server in the cluster), but the file system shields the differences between the cluster nodes [12]. Therefore, we feel that managing our water resources data in a same server. Actually, the data are distributed at different servers at different water resources management sectors.

In the distributed storage system, the data dispersed in different nodes may belong to the same file. In order to organize a large number of files, the files can be put into different folders, and the folders can be included in the first level. We call this mechanism “namespace”. The namespace manages all the files in the entire server cluster. Obviously, the responsibilities of namespaces are different from other nodes [13]. The node responsible for the namespace is called the master node, and the node responsible for storing the real data is called the slave node. The

master node is responsible for managing the file structure of the file system, and the slave node is responsible for storing the real data, which is known as the master–slave architecture. When operating, we first deal with the master node, query which nodes are stored on the slave nodes, and then read from the nodes, as shown in Fig. 3. In the master node, in order to accelerate the speed of data accessing, the entire namespace information will be placed in the RAM memory. When the large size of files is stored, the master node needs larger amount of RAM memory storage space. When storing data from a node, some raw data files may be large, some may be very small, and the size of the file is not easy to manage, then you can abstract an independent storage file unit, called block. Data stored in the cluster may be due to network reasons or server hardware causes access failure, it is best to use replica mechanism, the data backup to multiple servers at the same time, so that the data are safe and data loss or access failure probability is small, as shown in Fig. 4.

In the master–slave architecture, the master node contains the directory and structure information of the whole file system, as thus, it is very important to the distributed storage system. In addition, since the master node operates the namespace information in the RAM memory, therefore, the larger size of RAM memory is required when the file size is huge.

3.3. Distributed computing

After storing the water resources data properly, a reasonable data analysis and processing method should also be designed. For example, an efficient computing system to query and predict rainfall, total water usage, water supply fee and so on in real-time. Also, data virtualization, preparation, stream analytics, extracting, cleansing and loading the messy data need powerful computing. The distributed computing technology, therefore, is used to do computing for the water resources data.

The computer code is distributed at slave nodes and executed concurrently in parallel. This greatly shortens the execution time of the program. The computer code is moved to slave nodes to do computing. This method of moving computer code to the slave nodes is known as mobile computing [14].

Since computing is distributed at different water resources sectors, which located at different locations, a piece of code is required to summarize the intermediate computing results. Distributed computing completes this in two stages. The first stage reads and processes the raw data in each slave node. And then the processing results are aggregated to generate the final results, as shown in Fig. 5.

The computer code is handled by the master node, and the master nodes assign the computer code to different slave nodes to do efficient computing.

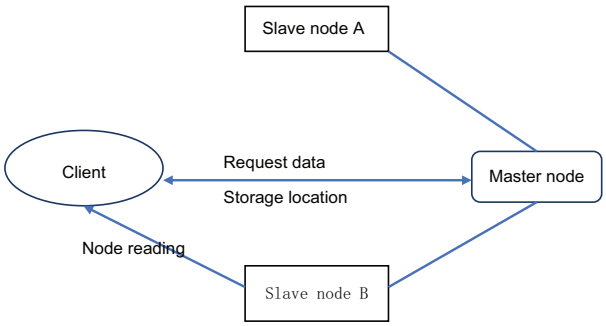

Fig. 4. Distributed storage system.

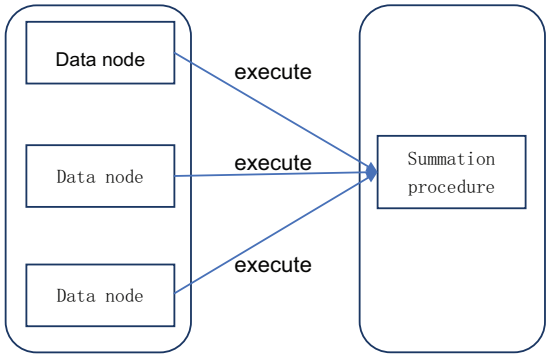

Fig. 5. Distributed computing.

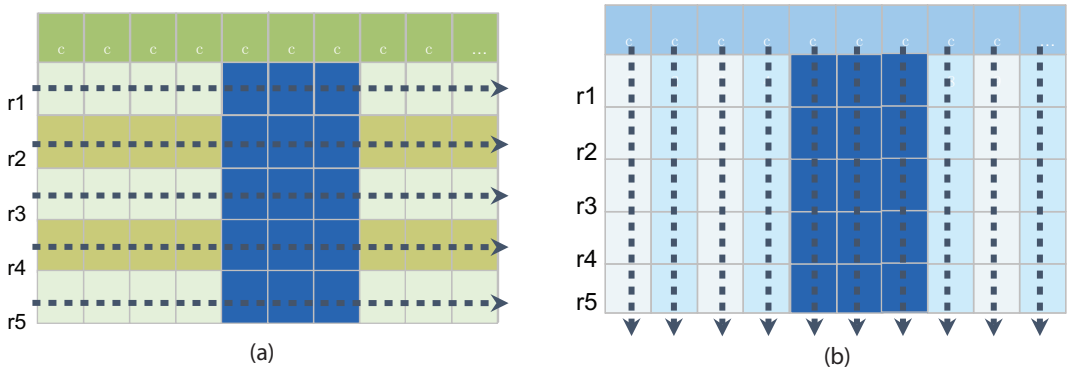

Fig. 3. (a) Row and (b) column data storage.

4. Design and implementation of water resources big data system

4.1. Physical structure of the water resources big data system

The physical structure, as shown in Fig. 6, includes the following: separation of storage network and computing network; storage devices and storage servers are distributed over the network; hierarchical storage based on disk tape; using distributed storage management software to manage storage space.

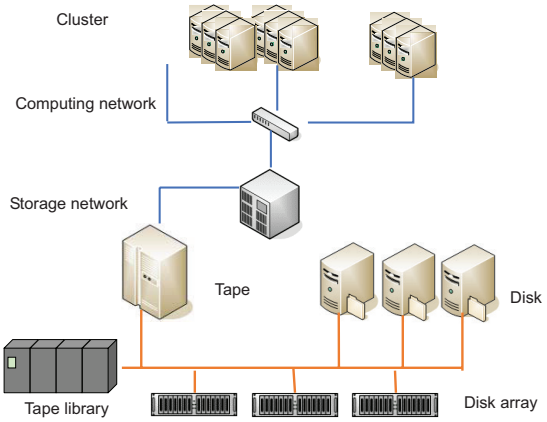

Fig. 6. Physical structure of the system.

4.2. Logical structure

Logically, the water resources big data system is designed in terms of users' requirements flow: the water resources survey, development, demand analysis, utilization, conservation, supply and protections. The users' requirements flow, as mentioned, has several sections, and each section has a number of applications. The water resources big data system prepares the spatial and associate data for each section in the users' requirements flow gradually. The "top-down" structure analysis method is used here aiming to provide necessary and sufficient data for each section and application [15]. The logical structure is shown in Fig. 7.

4.3. Performance and advantages

(1) After building the water resources big data system, the high-speed tool such as Hadoop can be used to identify easily new sources of water resources basic data and analyse the data in real-time and further help us make decisions quickly. Distributed spatial analysis service uses distributed computing method to decompose an analysis task into multiple subtasks, and finally obtains the results quickly through distributed computing. This big data system can access to various data sources and formats, and perform data analysis, support HDFS directory and enterprise spatial database. Multiple terminals in big data system can be quickly invoked and visually demonstrate

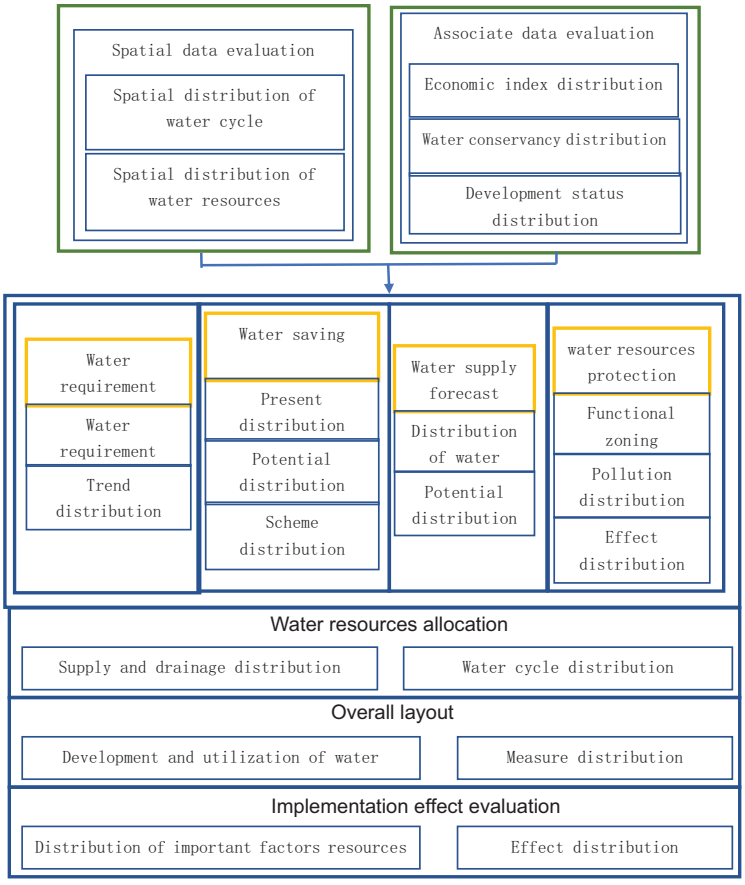

Fig. 7. Logical structure of the system.

the execution process and show results of the big data analytics task. The scalability of the water resources water big data system is very high, and analysis of the horizontal expansion of the node is also very convenient. In order to increase the efficiency, we can simply do this by adding processing nodes and registering the data.

(2) *Data analysis*: Water resources big data analysis is based on the application subjects. Big data analysis and processing can be applied into data mining, machine learning and statistical analysis in parallel and cloud computing. This solves the challenging large-scale computing problem facing data preparing and analysis [16]. The advantage is evident especially in real-time stream processing, the results are not a balance between real-time performance and accuracy, but rather accurate results under big data conditions.

(3) *Data applications*: Like the commonly used data analysis tools, the big data analysis also needs a vivid and effective data visualization. This helps users understand and analyse the results more easily. Since the results of water resources big data analysis are often complicated spatial-temporal data, the application of multi-dimensional visualization, tag cloud, historical stream, spatial information stream, etc., based on GIS is required [17]. According to the characteristics of water resources management applications, it can allow users to dynamically participate in and put in some a priori knowledge. This would make the big data system more intelligent and smart to deal with water resources applications.

5. Conclusion

This paper presents the design of water resources big data system. In particular, the paper provides the details of data source, storage design, distributed computing, physical and logical structure. The water resources big data system is able to identify more effectively and evaluate the quality of water resources data, analyse and predict water resources data. In addition, it provides strong technical support for various public events monitoring, early warning, and decision making. It also provides a way of accessing data in real-time, and performs data analysis (data mining, online analytical processing, deep learning, etc.) more effectively. It is innovative over the traditional data management methods. It supports distributed and unstructured processing. The main difference from the traditional water resources data centre lies in the ability to store data and process data.

Acknowledgments

This research is supported by a project "Research on Key Technologies of distributed spatial unstructured Big Data processing" (No. 2016CSOBDP0303) from Guangxi Colleges and Universities' Key Laboratory of Complex System Optimization and Big Data Processing. The project promotes the ability of young teachers in colleges and universities of Guangxi. The authors would like to acknowledge the projects: "The 3D terrain of the northern Gulf seabed Based On multi beam and remote sensing" (No. 2017KY0786) and "Research on Key Technologies of spatial unstructured large scale data processing" (No. IOT2017D01) from the Key Laboratory for Advanced Technology To Internet of Things, Qinzhou University, and the National Science Foundation of China grants "the Beibu Bay transitional ecological environment evolution mechanism and simulation study" (No. 41361022) and NSFC grant 11502241. The authors also like to appreciate the help from the teachers of Guangxi Colleges and Universities Key Laboratory of Complex System Optimization and Big Data Processing, Pei Wang and anonymous reviewers for their detailed comments and corrections.

References

[1] Y. Shi, Research on the application of the culture resource management based on big data technology, J. Appl. Sci. Eng. Innov., 4 (2017) 64–68.
[2] R. Wang, C. Yang, K. Fang, Removing the residual cellulase by graphene oxide to recycle the bio-polishing effluent for dyeing cotton fabrics, J. Environ. Manage., 207 (2018) 423–431.
[3] M.G. Armentano, D. Godoy, M. Campo, NLP-based faceted search: experience in the development of a science and technology search engine, Expert Syst. Appl., 41 (2014) 2886–2896.
[4] C. Vitolo, Y. Elkhatib, D. Reusser, Web technologies for environmental Big Data, Environ. Model. Software, 63 (2015) 185–198.
[5] J.L. Toole, S. Colak, B. Sturt, L.P. Alexander, A. Evsukoff, M.C. Gonzalez, The path most traveled: travel demand estimation using big data resources, Transp. Res. Part C: Emerg. Technol., 58 (2015) 162–177.
[6] A. McGovern, D. John Gagne, N. Troutman, Nathaniel, R.A. Brown, J. Basara, J.K. Williams, Using spatiotemporal relational random forests to improve our understanding of severe weather processes, Stat. Anal. Data Min. ASA Data Sci. J., 4 (2011) 407–429.
[7] X. Xu, F. Xie, X. Zhou, Research on spatial and temporal characteristics of drought based on GIS using Remote Sensing Big Data, Cluster Comput., 19 (2016) 757–767.
[8] X. He, N.W. Chaney, M. Schleiss, Marc, J. Sheffield, Spatial downscaling of precipitation using adaptable random forests, Water Resour. Res., 52 (2016) 8217–8237.
[9] Y. Kim, N. Kang, J. Jung, H.S. Kim, A review on the management of water resources information based on big data and cloud computing, J. Wetlands Res., 18 (2016) 100-112.
[10] R. Chalh, Z. Bakkoury, D. Ouazar, M.D. Hasnaoui, Big Data Open Platform for Water Resources Management, Cloud Technologies and Applications (CloudTech), 2015 International Conference on IEEE, 2015, pp. 1–8.
[11] L. Hao, R. Wang, K. Fang, Y. Cai, The modification of cotton substrate using chitosan for improving its dyeability towards anionic microencapsulated nano-pigment particles, Ind. Crops Prod., 95 (2017) 348–356.
[12] J. Yang, Research on improve of bat algorithm in the cloud computing resources, J. Appl. Sci. Eng. Innov., 4 (2017) 31–35.
[13] S. Adamala, An overview of big data applications in water resources engineering, Mach. Learn. Res., 2 (2017) 10–18.
[14] S.J. Walker, Big data: a revolution that will transform how we live, work, and think, Int. J. Adv., 33 (2014) 181–183.
[15] M. Swan, The quantified self: fundamental disruption in big data science and biological discovery, Big Data, 1 (2013) 85–99.
[16] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, Big Data: the Next Frontier for Innovation, Competition, and Productivity, Report, McKinsey Global Institute, 2011. Available at: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
[17] H. Chen, R.H.L. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, MIS Quarterly, JSTOR, 2012.