EPiC Series in Engineering
Volume 3, 2018, Pages 2383–2392
HIC 2018. 13th International Conference on Hydroinformatics

Research and Design of Hydrological Big-data Sharing Platform

Jian Xu [1] and Hua Chen [1]
[1] The State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, China
xujianwhu@ gmail.com, chua@whu.edu.cn

Abstract

In this paper, the methodology of hydrological big-data standardization is discussed upon analyzing on the characteristics of hydrological data. Three main standards of hydrological data in China are considered in the standardization, which are respectively "Structure and identifier for real-time hydrological information database", "standard for structure and identifier in fundamental hydrological database" and "Structure and identifier for water quality database". Solutions on data pre-processing, data indexing and highly efficient data reading and writing are also introduced. The mass storage capacity and high speed computing capability of Hadoop are utilized for designing and implementing hydrological big-data sharing platform. Then a prototype of the hydrological big-data sharing platform is introduced in this article. Accordingly, the platform can be the technical support for information sharing and space integration between water conservancy industries and other industries, as well as the interdisciplinary sustainable development.

1 Introduction

As a significant driving force of scientific and technological innovation, theories and methods of data sharing has been widely concerned by scientists and the public. In the United States, the Distributed Active Archive Centers had been established by NASA in 1990s. The DAACs were composed of 9 centers, and the Global Hydrology Resource Center (GHRC) DAAC included. In the Europe, the European Environment Agency (EEA) is in charge of researching data sharing of water science data, and the European Topic Center on Water (ETC) is responsible for assisting EEA's daily work and releasing related water information [1]. In China, the Meteorological Data Sharing System was launched in 2001, in order to improve the ability of science data sharing [2]. In 2002, another four data sharing centers were established by the Ministry of Science and Technology. the Surveying and Mapping Data Sharing Service Center, the Hydrological Information Sharing Service Center, the

Seismological Data Sharing Center, the Forestry Science Data Center and the Agricultural Science Data Center.

As the development of information technology, especially the maturity of Internet technology, there has been a lot of research on hydrological data sharing in recent years. Kochilakis et al. proposed a web tool for the management of floods and wildfires in urban areas [3,4]. The Hydroshare project put forward by The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) that enables users to share and publish data and models in a variety of flexible formats [5-9]. While data collected by hydrological stations are growing at a high rate every day, which considered by few papers. Yang et al. put up a big-data-based urban flood defense decision support system [10]. Ai et al. presented a framework for processing water resources big data and application [11]. Chalh et al. proposed a big data open platform for water resources management [12]. Although some progress has been made in hydrological scientific data sharing, there are still many problems, which can be summarized as follows.

- As the basis of hydrological scientific data sharing, mechanisms and standards are not enough, especially in the age of data explosion. There are a few published standards in how data organized in relational databases. However, no one is about data organization on big-data.
- Data storage is scattered in one institution. For instance, the same data of a reservoir's water level could exist both in real-time hydrological information database and fundamental hydrological database, which has an impact on data consistency and also causes storage redundancy.
- The same data may exist in different institutions, which also has the impact on data consistency and causes storage redundancy.
- Data storage capacity is limited, while the growth of hydrological data is unlimited. Each relational database management system has the maximum size for databases. Even if the hardware performance is improving, it will not keep up with the demand for data growth. How to realize the distributed sharing of resources and computing power and how to cope with the rapid growth of the current data is also an urgent problem to be solved in the field of hydrological data management and data processing.

Based on the characteristics of hydrological data, this paper discusses the method of hydrological data standardization and key technology of hydrological big-data sharing platform. The implementation of a prototype of hydrological big-data sharing platform is also mentioned in this paper.

2 Methods

2.1 Characteristics of Hydrological Data

Hydrological data refers to data about hydrology in yearbook, statistics, atlas and survey data, etc. There are three kinds of main databases in China, which are the real-time hydrological information database [13], the fundamental hydrological database [14] and the water quality database [15], that cover the most hydrological data in China. Hydrological data has the characteristics of large amount, variety, fast growth and high value due to its characteristics.

- The characteristic in data amounts: hydrological datasets are collected by stations all over the country. Each station collects the hydrological data 4 times a day at least. The total amount of hydrological data is big.
- The characteristic in data varieties: the real-time hydrological information database contains 13 kinds of data, such as precipitation, evaporation, river water level, reservoir water level, gate dam, pumping station, tide, sand, ice, groundwater, soil moisture, special water regime and hydrological forecast; the fundamental hydrological database contains 10 kinds of data, such as precipitation, evaporation, water level and flow, sediment, water temperature, ice, tides and so on; the water quality database includes water quality monitoring information in all kinds of water, such as meteoric water, surface water, and groundwater. Besides, geographic information data, remote sensing data and socioeconomic data should be also taken into account.
- The characteristic in data increasing: early in 2007, there are tens of thousands of stations across the country, and huge amounts of hydrological data are produced one day in each station. Hence, the amount of data is increasing quickly.
- The characteristic in data value: hydrology is an important branch of geography. And hydrology data is the important basis for the study of regional or global water cycle processes, which means hydrological data is of great value.

2.2 Classification for Hydrological Data

Taking above-mentioned three kinds of databases in China, the hydrological data could be classified in several types, such as, precipitation, evaporation, run-off (stage, flow rate and flow velocity), sediment content, water quality and so on.

As the platform is researched and designed for water conservancy industries in China, data stored and transferred in this water conservancy industries is considered. Three main standards of hydrological data in China are considered in the standardization, which are respectively “Structure and identifier for real-time hydrological information database”, “Standard for structure and identifier in fundamental hydrological database” and “Structure and identifier for water quality database”. Many tables are the same in this three standards, while others are unique. There are 9 primary categories and 121 secondary categories in total, and an identification code consisting of two letters is used for identifying data categories. The classification for hydrological data is presented as follows.

- The real-time data category, represented by letter ‘R’, includes real-time information, such as precipitation(R1), snowfall(R2), river run-off(R5), reservoir run-off(R6), and so forth.
- The forecast data category, represented by letter ‘S’, is the collection of forecasting information for precipitation, snowfall, run-off and other aspects of hydrology.
- The historical excerpts data category, represented by the letter ‘B’, is the excerpts of real-time data.
- The daily historical data category, represented by letter ‘C’, is the archive of daily historical data for precipitation, snowfall, run-off and so on.
- The ten-day historical data category, represented by letter ‘D’, is analogous to the daily historical data category. The difference is that data in ten-day historical data category is the average of daily data in ten days.
- The monthly historical data category, represented by letter ‘E’, is the average of daily data in a month.
- The yearly historical data category, represented by letter ‘F’, is the average of daily data in a year. Besides, the maximum figures are also recorded in this category.

- The survey data category, represented by letter ‘G’, includes the survey data, such as the survey data for cross section of river(G1), the research results of the investigation of flood and dry water(G6) and so forth.
- The water quality data category, represented by letter ‘Q’, is the set of data for recording water quality. Non-metallic inorganic data(Q2), metal inorganic data(Q3), phenols data(Q4), statistics on sewage discharge in river(QD) and so on.

2.3 Infinite Hydrological Datasets

In other aspect, hydrological data could be classified in finite datasets and infinite datasets. Data of low updating frequency could be the finite dataset, such as, the list of stations. Data that needs to be constantly updated can be classified as the infinite datasets, such as, the flow rate data produced by stations. As the low frequency of updating of finite datasets, relational databases is perfect for its storage, while another solution should be put forward for managing infinite datasets.

Hadoop Distribute File System (HDFS) is a high performance distributed file system for web-scale applications such as, storing log data, Map/Reduce data etc., which is designed for running on commodity hardware, and providing high-throughput data access. HDFS is an open-source implementation of Google File System (GFS), which was first proposed by Ghemawat, Gobioff and Shun-Tak Leung in 2003 [16]. HBase is an open-source, non-relational, distributed database modeled after Google's Bigtable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing Bigtable-like capabilities for Hadoop. HDFS and HBase are ideal systems for infinite datasets storage.

The infinite datasets could be classified in structured infinite datasets and unstructured infinite datasets. Defined by the definite properties, the structured infinite datasets are stored in relational databases currently and will be migrated to HBase. Accordingly, a data storage procedure will be put forward in the paper. The brief specification for structured infinite datasets is shown in Figure 1.

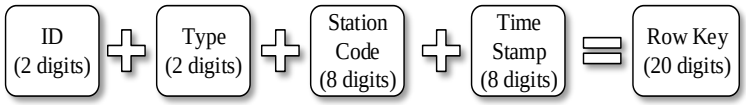

Figure 1: Structured infinite datasets storage specification

The ID stands for a sequential code, in order to differentiate the different editions of the same data in same time and same station. The type stands for the classification of data, which will be specifically introduced in 2.2 Classification for Hydrological Data. Then the station code follows, which is an 8-digits number specified by the Ministry of Water Resources. The last 8-digits number is the time stamp, stands for the data acquisition time.The unstructured infinite datasets usually refer to the unstructured hydrological data, such as pictures, audios and videos. This kind of data will be managed by HDFS.

2.4 Two-layer Index for Structured Infinite Datasets Retrieval

The retrieval time is short if the key word is type, station code and time stamp. But it takes long to retrieval data by any other key word. A two-layer index for structured infinite datasets retrieval is proposed for this issue. The diagram of the two-layer index is shown in Figure 2.

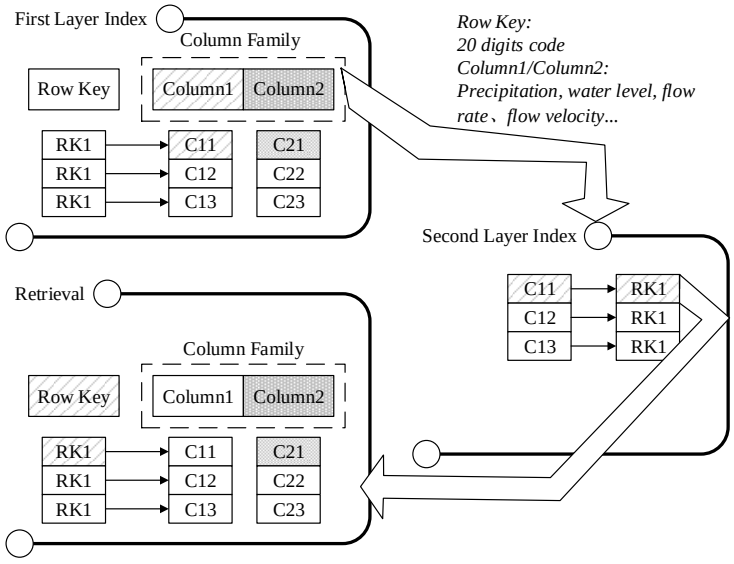

Figure 2: Diagram of Two-layer Index for Structured Infinite Datasets Retrieval

A corresponding index table is created for each data table, then the index table region corresponds to the data table region, and the pair of corresponded table is stored on the same regional sever. Eventually, every index table are the local index for corresponded data table region. When retrieving data, the index table region is retrieved before the retrieving on data table. Then all the eligible data table row keys are screened, the objective data is sought out.

2.5 Software architecture

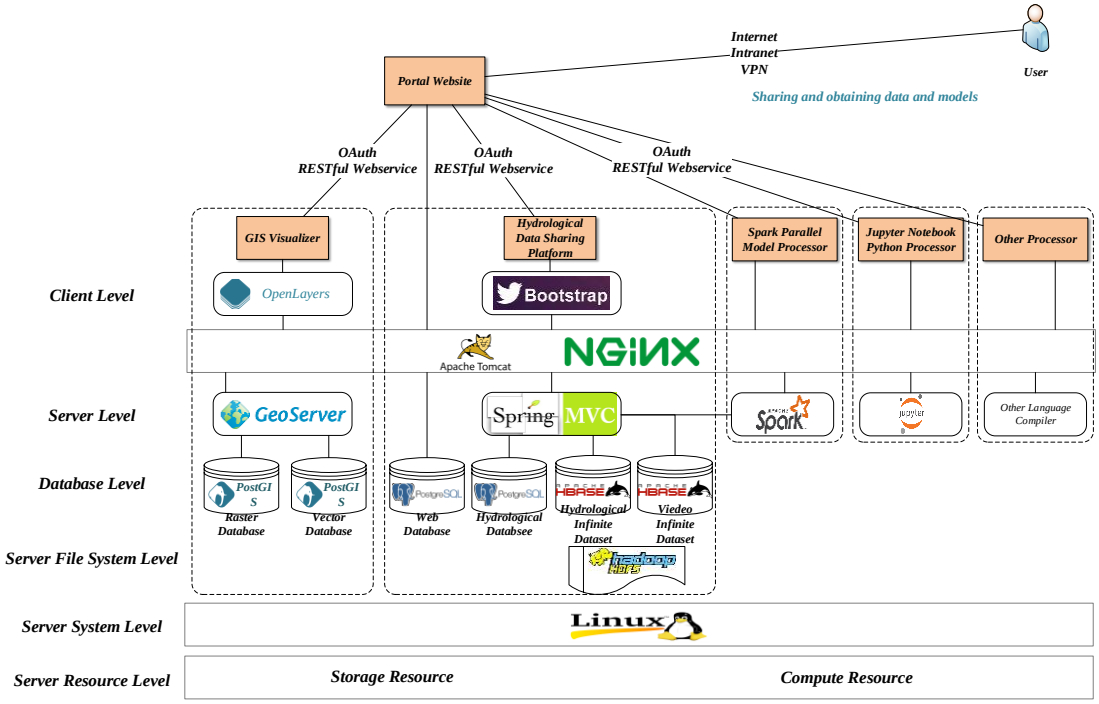

Figure 3: Hierarchical diagram of hydrological big-data platform

The hierarchical diagram of hydrological big-data platform is presented in Figure 3. It is a six-layer bottom-to-top hierarchical structure, which are respectively server resource, server system, server file system, database, server and client. The server resource includes storage resource and compute resource, which means users in client only need a visualization equipment without powerful performance, i.e. computer, laptop, tablet computer, and mobile phone. On the server system level, the hydrological big-data platform is deployed on the cluster equipped with CentOS, which is a Linux operation system. The top four levels are server file system level, database level, server level and client level, can be respectively divided in several subsystems, which is GIS Visualizer, Hydrological Big-data Sharing Platform, Spark Parallel Model Processor, Jupyter Notebook Python Processor and other Processors from left to right in Figure 3. As for GIS Visualizer, the PostGIS, a spatial database extender for PostgreSQL object-relational database, is used as the Spatial Database for data storage, then GeoServer is deployed as the spatial data server, while OpenLayers JavaScript API is utilized as the client interpreter. As for Spark Parallel Model Processor, Jupyter Notebook Python Processor and other Processors, code of hydrological process written in Java, Python and other program languages can be uploaded, compiled and run on server. Data and results for hydrological process can be respectively acquired on Hydrological Data Sharing Platform and demonstrated on GIS Visualizer. The Processors will be introduced in the future.

As for Hydrological Big-data Sharing Platform, HDFS is deployed for storage unstructured infinite datasets, and the underlying environment of HBase. PostgreSQL, a relational database, is used as web database and hydrological database. Spring MVC, like many other web frameworks, is designed around the front controller pattern where a central Servlet, the Dispatcher Servlet, provides a

shared algorithm for request processing while actual work is performed by configurable, delegate components, which is used in the server level. In Client, Bootstrap, a popular front-end component library, is the underlying environment for developing front webpage at the client end. Apache Tomcat is used for publishing website. And consider the situation that many people visit the website at the same time, Nginx, a web server which can also be used as a reverse proxy, load balancer and HTTP cache, is deployed between server end and client end.

All the subsystems are gathered by the portal website through OAuth and RESTful Web Service. The OAuth is an open standard for access delegation, commonly used as a way for Internet users to grant websites or applications access to their information on other websites but without giving them the passwords. Users, roles, institutions, privileges can be shared through OAuth between subsystems. Representational State Transfer (REST) is an architectural style that specifies constraints, such as the uniform interface, that if applied to a web service induce desirable properties, such as performance, scalability, and modifiability, that enable services to work best on the Web. The RESTful web service exposes a set of resources that identify the targets of the interaction with its clients. The applications communicated by RESTful web service are simple, lightweight, and fast. On hydrological big-data platform, users are able to share their data and models through Internet, Intranet, and VPN. All the support libraries and software are designed and maintained by open source software projects, and used under the license of GPL, BSD and Apache License, which are completely open to scientific research.

3 Software Implementation and Case Studies

A prototype of Hydrological Big-data Sharing Platform has been developed. There are 8 servers in the data server cluster, which are respectively named master, slaver01 to slaver07.The allocation of data server cluster is shown in Table 1. NameNode of HDFS, Hmaster of HBase, Driver of Spark, JobTracker of MapReduce, ResourceManager of YARN and Server of Zookeeper, are all deployed on server master. DataNode of HDFS, RegionServer of HBase, Worker of Spark, TaskTracker, are all installed on master and salver01 to slaver07. While Node Manager of YARN, Client of Zookeeper, are deployed on slaver01 to slaver07. Secondary NameNode of HDFS is installed on slaver07, which is a specially dedicated node whose main function is to take checkpoints of the file system metadata present on NameNode.

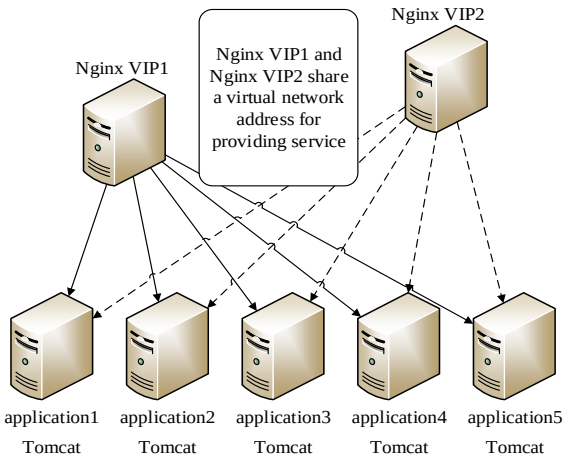

Figure 4: The deployment of web server cluster

The deployment of web server cluster is shown in Figure 4. There are 5 application servers with Apache Tomcat installed in web server cluster and 2 servers with Nginx installed. Two Nginx servers share a virtual network address for providing service.

Table 1: The allocation of data server cluster

	master	slaver07	slaver01-slaver06
HDFS nodes allocation	NameNode DataNode	SNameNode DataNode	DataNode
HBase nodes allocation	Hmaster RegionServer	RegionServer	RegionServer
Spark nodes allocation	Driver Worker	Worker	Worker
MapReduce nodes allocation	JobTracker TaskTracker	TaskTracker	TaskTracker
YARN nodes allocation	ResourceManager	NodeManager	NodeManager
Zookeeper nodes allocation	Server	Client	Client

A case study shows how the Hydrological big-data platform performed in sharing data collected from hydrological stations in Yangtze River. There are 32442 hydrological stations in total in Yangtze River. There are 102856280 rows for real-time rainfall data(R1), 28675695 rows for real-time reservoir run-off data(R6), and 52055316 rows for real-time river run-off data(R5). Guests can retrieve station information on a map without any authorization, but with no privilege for exploring figures provided by hydrological stations. After registered, users can explore data stored in HBase in order of stations, but still with no privilege for downloading data. Specific users can obtain downloading privilege by applying for correspond privilege and providing specific certificate, then after the administrators confirm the validity of the certificate document, users can obtain the privilege for downloading specific data. The downloading user interface is shown in Figure 5.

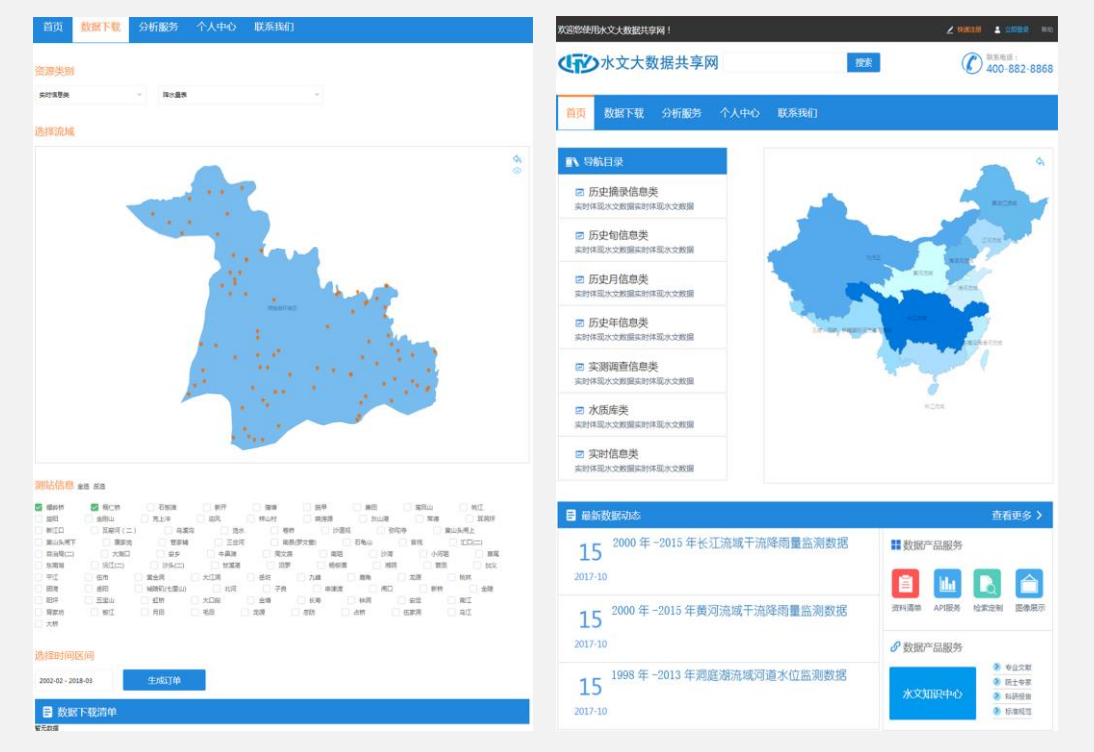

Figure 5: The user interface of hydrological big-data platform

4 Conclusions

The prototype of hydrological big-data sharing platform has been running for a period of time in good condition. In this paper, the methodology of hydrological big-data standardization is discussed upon analyzing on the characteristics of hydrological data. Solutions on data pre-processing, data indexing and highly efficient data reading and writing are also introduced. The mass storage capacity and high speed computing capability of Hadoop are utilized for designing and implementing hydrological big-data sharing platform. Then a prototype of the hydrological big-data sharing platform is introduced in this article. As the continuous research in hydrological big-data, more problems need to be solved, which can be summarized as follows.

- Hydrological big-data sharing in actual working environment need to be practiced.
- Research on service-based hydrological big-data sharing should be developed.
- Research on multisource heterogeneous big-data fusion could be carried on.
- Theories of machine learning could be utilized for data reorganizing, mining and summarizing, for the sake of exploration of hydrologic process.

Reference

[1] CAI Jianan, GENG Qingzhai. Study on concept and data collection system of scientific data sharing of water resources, Journal of China Institute of Water Resources and Hydropower Research 4 (2006) 31-35.

[2] GUO Yaxi. Meteorological Science Data Sharing System Construction and Service in China, China Science & Technology Resources Review 40 (2008) 14-18.

[3] Kochilakis, G., et al., FLIRE DSS: A web tool for the management of floods and wildfires in urban and periurban areas, Open Geosciences 8 (2016) 711-727.

[4] Kochilakis, G., et al., A web based DSS for the management of floods and wildfires (FLIRE) in urban and periurban areas, Environmental Modelling & Software 86 (2016) 111-115.

[5] Horsburgh, J.S., et al., Hydroshare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain, JAWRA Journal of the American Water Resources Association 52 (2016) 73-889.

[6] Kadlec, J., A.W. Miller, and D.P. Ames, Extracting snow cover time series data from open access web mapping tile services, JAWRA Journal of the American Water Resources Association 52 (2016) 916-932.

[7] Swain, N.R., et al., A new open source platform for lowering the barrier for environmental web app development, Environmental Modelling & Software 85 (2016) 11-26.

[8] Morsy, M.M., et al., Design of a metadata framework for environmental models with an example hydrologic application in HydroShare, Environmental Modelling & Software 93 (2017) 13-28.

[9] Rajib, M.A., et al., SWATShare – A web platform for collaborative research and education through online sharing, simulation and visualization of SWAT models, Environmental Modelling & Software 75 (2016) 498-512.

[10] Yang, T., G. Chen, and X. Sun, A Big-Data-Based Urban Flood Defense Decision Support System, International Journal of Smart Home 9 (2015) 81-90.

[11] Ai, P. and Z.X. Yue. A framework for processing water resources big data and application, Applied Mechanics and Materials 519 (2014) 3-8.

[12] Chalh, R., et al. Big data open platform for water resources management, in: Cloud Technologies and Applications (CloudTech), Morocco, 2015, pp. 1-8.

[13] Structure and identifier for real-time hydrological information database, The ministry of Water Resources of the People's Republic of China, 2011.

[14] Standard for structure and identifier in fundamental hydrological database, The ministry of Water Resources of the People's Republic of China, 2013.

[15] Structure and identifier for water quality database, The ministry of Water Resources of the People's Republic of China, 2014.

[16] Ghemawat, S., Gobioff, H., & Leung, S. T., The Google file system, ACM SIGOPS operating systems review, 37 (2003) 29-43.