The Design and Implementation of Combining the Standard of Data and Integrated Water Resource Data in Distributed Cloud Computing Environment

Feng-Cheng Lin¹, Ting-Wu Ho², Chen-Yu Hao¹, Che-Hui Lin¹

¹Geographic Information System Research Center, Feng Chia University, Taiwan
²Department of Management Information Systems, National Chengchi University, Taiwan

Abstract Water Resource Agency (WRA) has integrated cloud computing technologies based on cloud service framework for setting up easy-to-use data exchanging portal. Due to climate change and its effects on global society, the sustainability of water resource became a significant and highly attentional issue. In Taiwan, the demands of water resource information have increased either government agency, private institution, non-profit organization or domestic. Therefore, WRA followed the trend of open data, expanded nodes of current cloud environment and upgraded the version of a management tool for keeping the high performance, large capacity and utility of the cloud environment. Cloud computing is designed to provide the platform for data storage and user-friendly application procedures and interfaces in WRA. Data security of personnel is seen as well in this paper. In addition, for proceeding with the correction and stability of Taiwan water exchanging data standard, several procedures and systems are development as counterparts that help the affairs of data examination.

Keywords Cloud Computing, Hadoop, Open Data, Taiwan Water Exchanging Data Standard

1. Introduction

Due to climate change and the consensus of public participant, government has changed attitude from passive website with individual information query to actively provide open data (the term "open data" is spoken of frequently) with standard based format. To utilize and share large number of water or environmental management data, it should be collaborated with several organizations, integrating the power of central government as well as the city and county governments. The water resource management is a complicated work involving in many different units; therefore, Water Resource Agency (WRA) of Taiwan provides Water Resource Information Service Platform (WRISP - https://data.wra.gov.tw/), that already integrated cloud technology and established a distributed environment for the requirements of real-time and full-time. Data management is one of the most significant challenges in the field of water resources and environmental engineering. Our previous works and systems can be found in [1-2], and the general and preliminary study on the topics related to information and communication technologies (ICTs) in Smart water management in [3-4]. Nerveless, the application of data exchanging between organizations and open data still have many issues need to be solved in order to improve the performance of data exchanging and stability while multi-requests.

WRISP is a hub or broker that performs exchanging the information or data of rivers between different agencies in Taiwan. WRISP was built with the concept of Service Oriented Architecture (SOA) that based on the acquiring the instant information of different rivers, WRISP provides one-stop service to get the information or data of rivers [1-2]. Any system in WRA can obtain the information or data from another via the service bus (WRISP) instead of traditional data base linked. For users, portal entrance of WRISP platform provides the services or applications list, user can register their services or applications and apply the service request according to the application. This paper expanded WRISP for dealing with privacy protection policy and environmental data exchanging tasks from EPA (Environmental Protection Administration) of Taiwan. Cloud computing, data encryption, and data interfacing technologies were implemented in this paper for improving the functions, interfaces and procedures of WRISP. Data processing will be also a focus in this paper: data is information that are derived from WRISP, data are computed by proposed computing templates (user-defined, data mining, statistics, and text processing templates) to help answer queries, the queries you want answered access the information stored in the HDFS of Hadoop. Hadoop is an open source software solution for computation and...
storage in big data processing area, and Hadoop consists of three primary resources: the Hadoop Distributed File System (HDFS), the MapReduce programing platform, and the Hadoop ecosystem. Built for data-intensive applications, the HDFS is designed to run on a group of inexpensive commodity servers [5].

For further and various types of applications on hydrological disciplines, the quality examination process of Taiwan water exchanging data standard was established in this paper in order to provide correct information to other government agencies and open data to the public in Fig. 1.

2. The Trends of Cloud Applications

The concept of cloud computing was purposed by Google for the first time, and it is being used as a function for searching websites. This terminology was defined as a service model by National Institute of Standards and Technology, NIST, i.e., computing sources that can be accessed and shared through the Internet (such as servers, storage, and applications) with lower management, communication cost, and faster resources distributing and applying. Cloud computing is a great framework to be seen as an extension of distributed computing [12]. It breaks huge process down into several sub-processes, processes these sub-processes in parallel on servers, and then returns the results back to the user. Cloud computing abstracts the details of system implementation from users and developers. There are two essential concepts about cloud: Applications run on physical systems that aren’t specified, data is stored in locations that are unknown, administration of systems is outsourced to others, and access by users is ubiquitous. Cloud computing virtualizes systems by pooling and sharing resources. [13].

Big data becomes the cutting-edge technology recently. According to the research of Industrial Economics & Knowledge Center (IEK) of the Industrial Technology Research Institute (ITRI) of Taiwan, R.O.C, there are increasing demands on the integration between big data and current information technologies for enterprises. Those enterprises will need to have enough backgrounds on Hadoop[6-7], R[8-9] or MapReduce[10] in order to set a cloud infrastructure up for their big data analysis, which is not every enterprise can afford the cost of fundamental infrastructure. Several commercial solutions have been developed for the demands on IT market, include Elastic MapReduce by AWS[11], Hadoop as a service by Cloudera. Since commodity servers are used in Hadoop, it is easy and economical to increase the number of machines in case of extra load or increasing the number of users. In addition to cloud supporting services, the customized service on the device-tier is also a crucial process for setting up the cloud environment successfully. Therefore, the way to understand and solve problems for the users is the solution to create benefit from cloud and big data. In other words, improve the design of an information system and simplified the procedure that clients will face are the most important action for providing a successful service.

On the side of government, the concept of precision marketing can be applied for improving the current service systems. Mobile applications have been wildly used in the modern society and human behaviors; moreover, environmental information can be transferred or regenerated for commercial service to the public, which includes the applications of smart city, smart transportation, smart hydrology and smart home etc.

Figure 1. The architecture of water resource integrated cloud.
3. The Application of Cloud Storage

Cloud based service has been a mature product in the information technology market. One of the most popular products is software as a service, SaaS, for instance, Google drive and Dropbox. However, public cloud has such conveniences, lower cost on infrastructure and operation, it still has some shortages, which include security and capacity limitation which makes public cloud cannot satisfy water management affairs. Therefore, Water Resource Agency developed the WRACloudBox. This software is not only operating upon private cloud environment, but integrating public cloud service, Google Drive and Dropbox. In addition, mobile application for portable cloud service has been upgraded as well.

WRA cloud storage service provided an interface that allows the user to merge with WRACloudBox, Google Drive and DropBox into a single operating interface. Users can create, delete, download, transfer data format to ODF and online preview through this single interface. The files can be transferred from one to the other easily in Fig. 2.

![Figure 2. The portal of hybrid data cloud (Chinese version).](image1)

WRA cloud storage service provided an interface that allows the user to merge with WRACloudBox, Google Drive and DropBox into a single operating interface. Users can create, delete, download, transfer data format to ODF and online preview through this single interface. The files can be transferred from one to the other easily in Fig. 2.

![Figure 3. Mobile personal cloud storage space (Chinese version).](image2)
Mobile device becomes necessary tool in our daily life. Water resource agency developed mobile application service that integrated mobile single sign-on service. User can access to WRACloudBox through personal mobile device by ID and password (left of Fig. 3) and list the folders and files (right of Fig. 3) after a successful login.

4. Taiwan Water Exchanging Data Standard

Taiwan water exchanging data (TWED) standard is positioned as the format of data exchange of water resource agency. This standard was formulated based on the framework of ‘data format standard of water resources’ and consulted with the users and managers of internal systems. In 2014, the standard considered the requirements of applications, created detail categories with different levels, and merged ‘WRA geospatial GML standard’ which increased the consistency of different standards in Table 1.

| Main Category     | Sub-category                        |
|-------------------|-------------------------------------|
| River             | Water Level                         |
|                   | Water Flow                          |
|                   | River Management                    |
| River Cadaster    |                                     |
| Groundwater       | Groundwater Level                   |
|                   | Groundwater Quality                 |
| Land Subsidence   |                                     |
| Related Agency    |                                     |
| Drain Facilities  |                                     |
| Sensitive Environment | Water Source Quality and Volume Protection Area |
|                   | Ecology                             |
| Reservoir         |                                     |
| Flood Protection  |                                     |
| Water Rights      | Water Usage Registration            |
| Inshore           | Tidal                               |
|                   | Wave                                |
| Administration    | Construction                        |
|                   | Project                             |
|                   | Investigation                       |
|                   | Knowledge                           |
|                   | Administration                      |
| Statistic of Water Resources | Statistic of Water Usage |
|                   | Statistic of Water Rights           |

TWED standard and WRA geospatial GML standard are both standards that are described and designed base on eXtensible Markup Language (XML). However, WRA geospatial GML standard took geographical feature as its fundamental element which adapted the framework of international standard of geographic markup language of OGC. This means it has constructively different with Taiwan water exchanging data standard that has characters as multiple types and highly customized. Therefore, we compare two projects based on the name and its contents, use annotation, and import name space for referencing. We imported the complete GML standard into the framework of Taylor water exchanging data standard.

We adapted rolling management procedure to manage the content of TWED standard. The items and contents can be added, modified, and deleted based on the requirements from each execution unit and management affairs. In the current process, the information of 121 of 148 datasets of applications has been collected in data storage. Due to various data sources with complex and wild fields’ affairs and data formats, so the data accuracy became a significant issue for assuring the quality and correction of data exchanging. For different purposes, we initialized two management teams: standard evaluation team and standard examination team. The latter is responsible for administrative work, communication, data processing, and providing standard evaluation and examination.

5. The Implementation of Computing Template

The greatest advantage of cloud computing is that developer can process big data by personal computer. In order to bring out the cloud computing capability of “Water Resource Integration of Information Cloud” for rapidly using computing resource by each section of the Water Resources Agency system. We apply JAVA to develop computing template and enclose the front-end interface of computing template by Web Services to enhance the resilience of utilizing computing template in Fig. 4.

![Figure 4](image_url)

Figure 4. The architecture of Computing Template.

It’s hard to proceed with the data mining from boundless big data. Thus, Water Resources Agency developed several kinds of model for data mining and set up raw data format to use mining template easily. The data mining template was developed include decision tree, clustering, and association rules. Decision tree needs to have training data with CSV file format, and we can obtain the same output through the completed calculation of decision tree.
Clustering can group similar or relative elements; therefore, computation of K-Means will acquire different category result through the input data in CSV format based on attribute of keywords as data set. Association rules can provide the resolution output of relative elements, can be applied on relatively official document recommendation and purchasing products recommendation, etc.

This paper reused “electronic official document system of Water Resources Agency” for the case of SaaS application to practically verify feasibility of computing template in Fig. 5. In this case, we analyzed the information of people, affair, time, place and substance by compiling and importing official document to database systems in the cloud; therefore, we established electronic official document relationship diagram to query specific affair from official document context, offer official document recommendation, and anticipate amount of official document analysis model application.

To verify the case of activating official document, we created four cases as examples: searching official document by keyword, anticipating the process time of the official document, anticipating the amount of official document, and clustering electronic official document. The implementation of data processing module was to read an official document in a fixed time then convert data into the data format of R, Solr and Mahout in disparate templates to develop and set up those aforementioned functions of subsystems in Fig. 6.

Figure 5. The framework of official document system.

Figure 6. The flow of official document processing.
In this case, main task is to process the strings of Chinese; therefore, we used mmseg4j published by Mr. Chih-Hao Tsai [14] in 2000 as our preprocessing program. The mmseg4j contains two algorithms and four fuzzy resolution rules; therefore, the accuracy of Chinese word segmentation can achieve 98.41%. The default vocabulary deducted no-meaning words and incorporated with vocabularies of a Chinese word (140 thousands vocabularies totally). Default of mmseg4j used mmseg4j vocabulary that can adopt the utf-8 code to be customizable setting. Based on this case, we have already increased 4219 vocabularies on Water Resources Agency exclusive dictionary.

Searching function included full text search, noun classification, association computing template; moreover, user not only can hasten query pace, prompt automatically, and hit indicator function to search official document more conveniently but also search similar key words to overcome query restriction through relative official document searching function in Fig. 7. This new function can hit more key words by finding relative official document of high relevant, and it can be widely promoted the efficiency of official document query. Full text search function provides official document basic query, official document terms category query, and relative official document query. The result after querying provides relative official documentary inventory, and this approach used user-based collaborative filtering technology to compile relative official document of user’s information. This algorithm applied the shortest neighborhood search in keywords and computed the similarity of already evaluation and the items that are waiting for examination. For instance, as proceeding a similarity between the item A and item B, you need to find out the combination which gives A and B score at the same time, and make these combinations proceed a similarity computing, then generate a recognizable result.

Anticipating the amount of official document will use time series analysis computing template, and user can input an ending time from now. System will anticipate the possible amount of the official document in this interval. The advantage of this function is to enhance efficiency on allocating manpower.

Clustering function extracts the amount of official documents that need to be analyzed, and then Chinese parsing/segmentation system will segment short words of each subject to obtain non-repeated keywords list for K-Means clustering algorithm. The coding way is 0 or 1 means there is existed keywords in this document. Clustering dismember a group of data to several groups based on the distance measure, and the concept of this approach is that the items in one group with the nearest distance from centroid, but the relations between different groups with far distance from each centroid of each group.

6. Monitoring Platform

Water resources integrated cloud platform was in charge of data exchanging, putting data into the database by agent, single sign-on system, web site, node computing, etc. Managers need to know each crucial indicator to understand it’s efficiently and loading conditions during the operation.

Besides crucial services, system also needs to monitor the case of hardware like CPU, hard drive, memory, Internet, and process in Fig. 8.

Our agent plays an important role to put data into data database, so it can be monitored by socket transmission between agent server and agent client to check whether TCP/IP connection is working normally or not. According
to data scrutiny principles of Water Resources Agency, there are two indicators: correctness and completeness for putting data into the database. Correctness is to examine the consistency of content and type of standard (if we set the standard as date, it will detect automatically unusual when entering the type of words). Completeness is column numbers of dataset are conformed, and vacancy of content cannot be greater than 30%; moreover, the required field information should not be vacant.

7. Conclusion and Future Work

The greatest advantage of cloud computing is to apply the parallel processing in dealing with big data in distributed architecture to speed up the efficiency of data analysis and processing. This paper developed word processing template, statistic template, and data mining template; moreover, we took official document for sample to verify the real time characteristic of cloud computing that can be used in different sections of Water Resources Agency and enables IT to more rapidly adjust resources to meet requirement. Data activation has successfully accomplished 148 items of dataset by content checking and establishing, and 49 exchanged standard datasets (18 datasets are open data) has been collected in a warehouse. Via data exchanging and opening operation, it can increase the efficiency of execution in organization, we also promoted the transparency of WRA administration to meet the direction of government policy.

Open government data platform is a trend (transparency in the data) for the government to encourage residents join public affairs. If government can establish the mechanism for internal data providing and setup reward system for opening data, data owner (authority) can provide data initiatively, it will make more complete for open data inventory of Water Resources Agency and Environmental Protection Administration to promote the usage of data where cross-administration data can be more circulating and fulfill the need of people. From now, open data of WRA combined water resources data standard by data gathering and exchanging flow, it already meets four star under Tim's 5 star open data plan, and this paper advises the data can be five star where data can be connected by others systems. Then it can also links another open data dataset to become newly open data. This concept just like the new framework of the World Wide Web allows people to understand the data venation and explore more relevant information through data linking.

![Figure 8. The report of real-time monitoring (Chinese version).](image)
The Design and Implementation of Combining the Standard of Data and Integrated Water Resource Data in Distributed Cloud Computing Environment

Acknowledgements

This work is supported by the Ministry of Science and Technology of Taiwan, under grant 104-2627-M-035-005.

REFERENCES

[1] Cheng-Te Yang, Geui-Lin Huang, Fu-Hsun Huang, Ting-Wu Ho, PiHui Huang, and Jiann-Min Yang. SOA-based platform for water resource information exchanging. In Geoinformatics, 2009 17th International Conference on, pages 1-4, Aug., 2009.

[2] Lin Che-Hui and Ho Ting-Wu, SOA-Based Water Resources Information Exchanging In Taiwan, ISPRS workshop on Dynamic and Multi-dimensional GIS, Oct., 2011.

[3] Pagani, G.A. & Aiello, M., An ICT overview of water management. [Online] Retrieved from: http://www.cs.rug.nl/~andrea/publications/ictWMgt.pdf, 2012.

[4] Partnering for solutions: ICTs in Smart Water Management, International Telecommunication Union (ITU), United Nations Educational, Scientific and Cultural Organization (UNESCO), [Online] Retrieved from: http://www.itu.int/dms_pub/itu-t/oth/0b/11/T0B11000253301PDFE.pdf, March 2014.

[5] Kevin Sitto and Marshall Presser, Field Guide to Hadoop, O’Reilly Media, Inc., March 2015.

[6] Tom White, Hadoop: The Definitive Guide, Fourth Edition, O’Reilly Media, Inc., 2015.

[7] Alex Holmes, Hadoop in Practice, Second Edition, Manning Publications Co., 2015.

[8] Viswa Viswanathan and Shanthi Viswanathan, R Data Analysis Cookbook: Over 80 recipes to help you breeze through your data analysis projects using R, Packt Publishing, 2015.

[9] Robert I. Kabacoff, R in Action: Data analysis and graphics with R, Manning Publications Co., 2015.

[10] Thilina Gunarathne, Hadoop MapReduce v2 Cookbook, Second Edition, Packt Publishing, 2015.

[11] Amarkant Singh and Vijay Rayapati, Learning Big Data with Amazon Elastic MapReduce, Packt Publishing, 2015.

[12] Mahmood, Z. and Hill, R., Cloud Computing for Enterprise Architectures, Springer-Verlag London Limited, 2011.

[13] Sosinsky, B., Cloud Computing Bible, Wiley Publishing, Inc., 2011.

[14] Tsai, Chih-Hao, http://technology.chtsai.org/mmseg/, 2012.