

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/335479833>

Extending the National Lake Database of Malaysia (MyLake) as a Central Data Exchange using Big Data Integration

Conference Paper · July 2019

DOI: 10.1145/3352411.3352417

CITATION

1

READS

267

4 authors:



Fara Yahya

Universiti Malaysia Sabah (UMS)

32 PUBLICATIONS 332 CITATIONS

[SEE PROFILE](#)



Bashirah Mohd Fazli

National Hydraulic Research Institute of Malaysia

13 PUBLICATIONS 25 CITATIONS

[SEE PROFILE](#)



Mohammad Fikry Abdullah

National Hydraulic Research Institute of Malaysia

15 PUBLICATIONS 73 CITATIONS

[SEE PROFILE](#)



Harlisa Zulkifli

National Water Research Institute of Malaysia (NAHRIM)

9 PUBLICATIONS 31 CITATIONS

[SEE PROFILE](#)

Extending the National Lake Database of Malaysia (MyLake) as a Central Data Exchange using Big Data Integration

Farashazillah Yahya[†]
Faculty of Computing and Informatics
Universiti Malaysia Sabah, Kota Kinabalu,
Sabah Malaysia
fara.yahya@ums.edu.my

Bashirah Mohd Fazli, Mohd Fikri Abdullah,
Harlisa Zulkifli
National Hydraulic Research Institute of Malaysia
(NAHRIM), Selangor, Malaysia
{bashirah; fikry; harlisa}@nahrim.gov.my

ABSTRACT

With the rise of heterogeneous large spatial and non-spatial data, systems are developed to manage these data sets. Big data emphasizes heterogeneity among systems leading to data integration issues due to the nature of big data which includes volume, variety and velocity. MyLake is a National Lake Database to manage information and knowledge sharing on lakes in Malaysia. At the moment, data are uploaded by each agency using MyLake as a platform. Nevertheless, this is carried out manually and require timely human effort. Each agency does one-to-one data integration (in-silo) where the integration is developed according to the agency-specific needs resulting from a possibility of integration issue across agencies. Therefore, this paper introduces a big data integration approach that extends the MyLake repository. The big data integration platform is a preliminary idea of how data can be shared, integrated, retrieved, and disseminated within a reliable and authenticated environment. The proposed centralized platform consisting of a set of standards, tools, repository and registry that enable multiple integrations between different agencies. The platform offers the potential to provide a reliable platform that acts as data retriever and disseminator.

*Article Title Footnote needs to be captured as Title Note

[†]Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WOODSTOCK '18, June, 2018, El Paso, Texas USA

© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...\$15.00

CCS CONCEPTS

• Insert CCS text here • Insert CCS text here • Insert CCS text here

(For how to generate CCS:

<https://www.acm.org/publications/class-2012>)

KEYWORDS

Big Data, Data Integration, Big Data Integration, Lake, National Lake, Database

ACM Reference format:

1 Introduction

The National Hydraulic Research Institute of Malaysia (NAHRIM) is a water research institution under the Ministry of National Resources and Environment of Malaysia (NRE). NAHRIM has taken the initiative of developing a National Lake Database of Malaysia back in 2011. The system is an improved and upgraded version of the 2007 Malaysia Lake Database, which unfortunately functions as a data storage and accumulating data based on the lake brief programme. The access to the system is limited to stakeholders (government agencies and departments) and lake managers. Through MyLake, dashboard on MyLake information is made available to the public community to access and get the information. The process of information dissemination and explanation should be conveyed from stakeholders and lake managers from time-to-time through awareness program to ensure the information is explicable for public understanding.

The aim of extending MyLake is to establish a coherent platform to collect, integrate, analyse and store various

sources data in a centralized database. The interactive and user-friendly system provides a platform with an ability to update and analyse data and information on the fly as the data and information deposited in the platform are reliable. This is vital to avoid data ambiguity, data redundancy, data manipulation and data corruption. Data presented through this centralised platform will ensure data security, data integrity and data validation for various purposed in lake management.

Multi-dimensional data integration among government agencies can provide agencies the source of insights and information to offer a seamless citizen-centric service. [1]. This is important to ensure interaction between the government and community can enhance a good decision-making process across various sectors. Previous researchers mentioned that digital services provided by the government can possibly increase the information needed by the government, policy-makers and the citizen with data governance and awareness activities carried out online and offline [2].

The government of Malaysia encourage engagements from the community in lake basin management. This is in line with the principle of the World Lake Basin principles which clearly stated that “citizens and stakeholders should be encouraged to participate meaningfully in identifying and resolving critical lake problems”. Therefore, establishing a national lake information centre is seen as a good initiative towards a sound and sustainable management of lakes and their basins [3]. A comprehensive, centralised and visualised dynamic database will further improve the current methods used in lake basin management, promoting transparency and empowerment of all stakeholders which is the fundamental core of the International Lake Environment Committee Foundation (ILBM) [4][5].

The data repository i.e., where data and information about lakes are generated, shared and disseminated among stakeholders must be maintained as the single source of truth; consistent and accurate source to academics, scientist, and lake managers. [6].

To overcome the challenge as a single source of truth, there is a requirement to affirm a suitable platform that ensures the data are reliable, accessible, comprehensive, collective, timely and available. The questions are:

a) How can big data integration be carried out among large data sets of various heterogeneous data (including spatial and non-spatial data etc.)?

c) What platform can be established as a central data exchange among lake basin stakeholder(s)?

The upgraded version of the national lake database will not only play a distinctive role as centralised information centre and support system for the policy makers, key stakeholders and lake managers. It will also act as a platform to encourage more active engagement on lake management and foster greater awareness of lakes in Malaysia by exposing data to the data broker available at registry database. These web services include services which retrieve data from repository database and web services which are exposed by providing agencies. Thus, this will generate a valid, verified and updated information to be referred via a big data integration platform.

The paper will be presented in several sections. The next section will discuss the main concepts emerging around this work. We will also present about MyLake and discuss the proposed big data integration platform. Our preliminary idea for big data integration consists of a set of standards, tools, repository and registry that enable multiple integrations between different agencies. Lastly, the final section draws conclusions and suggests future work.

2 Related Work

Big Data

Since 2011, the word Big Data has been a hype among researchers [7]. A researcher implied that the characteristics of big data in the context of data management generally include the three V's which are Volume, Variety, and Velocity [8]. The three V's has been used as the main components that construct the big data framework. [9][10]. A definition by Gartner, Inc. states that: “Big data is described as a high volume-velocity-variety of data that is economical, an innovative way to process and analyse data for improved insight and decision making.” [11].

Data Integration

Information and communication technologies (ICT) play an important role for data sharing and data integration in information systems. The development of modern ICT is the contributing factors to the development of efficient land

information systems in an organization. Due to the current practices in an organization; it shows that there is a silo-practice in the land department. Besides, lacking availability, not up-to-date, no standard data and information; and incomplete information and data will affect the process of decision-making for better land administration. Also, other issues are related to delay the process of managing and analysing data, as well as increase the cost of collecting the data that contributed to the redundancy of data and information.

Big Data Integration

The term big data integration (BDI) can be defined as fusing or connecting large heterogeneous data that can come from multiple sources. There are several distinct differences between big data integration and data integration itself. Among the differences are [12]:

- Single-domain data sources has increased (such as omics, bioinformatics, spatial etc.)
- The data is dynamic in nature
- Data are heterogeneous in structure
- The data sources have different and sometimes unique qualities in the domain, accuracy and timeliness of data provided.

The challenging areas of data integration in a big data includes (but not limited to) mapping data schemas, matching database records and data fusion [12] [13]. Supposedly, there are data managers from two different organisations, having different big data architectures with the idea of collaborating their process or activities and would want to operate centrally without changing their existing system. The main focus would be in terms of data integration and interoperability as both organisations have different data management and policies [12][14]. Data integration identifies the suitability and data readiness, either at the integration process, administrative tasks, and databases [15]. The challenges of big data integration involve data integration and the interoperability between data storage and data structure [16][17]. The nature of a heterogeneous the is emphasized specifically in the context of big data; big data integration supports data sharing between citizens, data users, and organisations. Once data has been shared, data are able to be exchanged through a platform that governs the data exchange process. Although it may sound effortless, the technicality, administration and interoperability are tricky issues to tackle. Undeniably, the big data integration provides an explosion of various sourced data. Nevertheless,

gathering data at such volume may promote data growth but linking these data to ensure it is always accessible is deemed as complex [18][19][20].

Following we present on MyLake as the main use case of the work.

3 Lake Database for Malaysia (MyLake)

In 2012, the Malaysian National Water Council Strategic Planning established a National Lake Centre through the development of the Lake Database for Malaysia (MyLake). MyLake was developed by the National Hydraulic Research Institute of Malaysia (NAHRIM), which is a prominent water research institution under the Ministry of National Resources and Environment of Malaysia (NRE). NAHRIM also manages data, information and knowledge about lakes in Malaysia.

Background

MyLake is a central data repository for lakes in Malaysia that benefits the citizens and research communities (stakeholders, academician, the public/private sector and consultants) in Malaysia. MyLake is a centralised database platform that stores data including but not limited to data and information about lakes in Malaysia. The information has been used to increase understanding about lakes and the associated variable surrounding it. Besides that, MyLake has been used as a platform to upload (store) and later shared (download) between respective agencies in the federal, state and local authorities. The central database is an improved and upgraded version of the 2007 Malaysia Lake Database, which unfortunately has been functioning as an information repository and data bank served to store and accumulate data based from the lake brief program. At the moment, access to the system is only limited to stakeholders (government agencies and departments) and lake managers.

Data Sources

The two main data sources in MyLake includes spatial and non-spatial data (Table 1). The spatial data includes data and information about demographic, geology and soil types, hydrology, infrastructure and urban landscape, land use map, and utilities are gathered from relevant government agencies. The main government agencies that has contributed as data providers are listed as below:

- Department of Drainage and Irrigation (DID)

<https://www.water.gov.my/>

- Department of Survey and Mapping Malaysia (JUPEM)
<https://www.jupem.gov.my/>
- Mineral and Geoscience Department of Malaysia (JMG)
<https://www.jmg.gov.my>
- Department of Agriculture (DOA)
<http://www.doa.gov.my/>
- Malaysian Meteorological Department (MET)
<http://www.met.gov.my>
- Department of Statistics Malaysia (DOSM)
<https://www.dosm.gov.my>
<https://www.jupem.gov.my/>
- Department of Town and Country Planning Peninsular Malaysia (JPBD)
<https://www.townplan.gov.my/>

Type	Data
Spatial Data	State Boundary
	District Boundary
	Sub-district / City/Town Boundary
	River Network
	River Basin
	Rainfall distribution
	Ex-Mining area
	Topography
	Road / Transportation
	Land Use / Land Cover
Non-Spatial Data	Mineral Sources
	Mining
	Lake
	Lake Profile
	Lake Water Quality
	Trophic State Index
	Flora
	Fauna
	Socio Economy
	Lake Management
	Heavy Metals
	Hydrological Data
	Point Source Inventory

Table 1: List of Spatial and Non-Spatial Data

MyLake also stores non-spatial data which consist of other related information on lake management and the ecosystem.

This also include the data uploaded by the lake managers and owners.

Owner of MyLake

MyLake is a framework that consist of an information system, a repository, servers hosting the system. MyLake is copyrighted and protected by the Malaysian Intellectual Property. Currently, the custodian of MyLake is NAHRIM and data are uploaded by the lake managers/owners, universities or research institutes. At the moment, data and information residing in MyLake belongs to each agency that provided the data to MyLake. The data owners are also responsible of maintaining the data. This means uploading the latest, clean and accurate data every time. IN a nutshell, MyLake is a symbol of collaboration between NAHRIM and lake managers in Malaysia.

The information accessible from MyLake includes the background of lake, management info, the contact details, the water quality and trophic state index, the concentration of heavy metals, and the ecosystem (flora and fauna) of the lake. To add, MyLake has a user interface that can show navigation of the map (like google earth system).

4 How can Multiple Data Sources be integrated in MyLake?

The big data integration platform is a preliminary idea of how data can be shared, integrated, retrieved, and disseminated within a reliable and authenticated environment. The big data integration platform is introduced as a centralized platform that enables multiple integrations between different agencies.

The existence of data silos (when integrations are undertaken in silos) at many agencies can result in data inconsistency. With the integration platform, information shared is at the latest version, accurate and consistently retrieved from a single source of truth. The big data integration platform enables the requested data to be redistributed to the user upon approval from the data owner. Government agencies can interact digitally, where the process of data sharing may no longer require written correspondence, discs, USB and others but instead uses the platform.

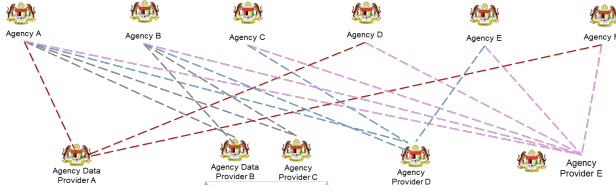


Figure 1. In- Silo Data Integration (“As-Is”)

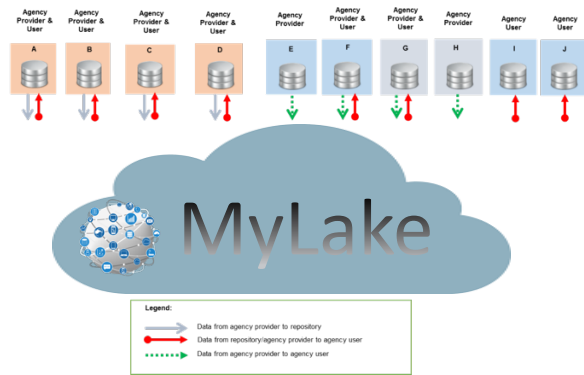


Figure 2. Conceptual Diagram of MyLake Big Data Integration (“To-Be”)

Currently, data are uploaded by each agency using MyLake as a platform. Each agency does one-to-one data integration (in-silo) where the integration is developed according to the agency-specific needs resulting from a possibility of integration issue across agencies (Figure 1). Therefore, the platform offers a centralized integration platform consisting of a set of standards, tools, components, repository and registry that enable multiple integrations between different agencies as in Figure 2. The big data integration will establish as a reliable platform that acts as data retriever and disseminator.

5 Proposed MyLake Big Data Integration Approach

Spatial data integration promises potential prospect for multidisciplinary area [21]. One of the common big data integration models for spatial and non-spatial data is the middleware integration model [16][21]. The middleware provides interaction between the data and application layer. The middleware is built as the MyLake Integration layer that acts as a central data exchange offering a unified data access interface. The data access interface interacts with other layers for example between the MyLake Portal and MyLake

Integration layer in Figure 3 i.e. the management and also the database layer.

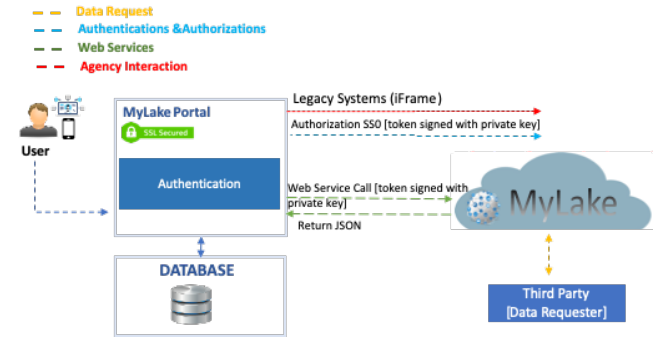


Figure 3. Proposed interaction between MyLake Portal and MyLake Integration layer

The middleware also protects the heterogeneity of the database and improves the interaction between the integration layer and the data providers (agencies) through web service call as shown in Figure 4. Therefore, in this paper, we propose ad middleware integration model.

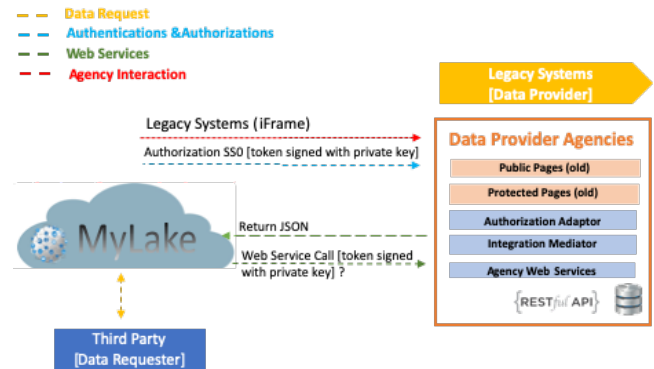


Figure 4. Proposed interaction between MyLake Integration layer and other Legacy Systems (Data Providers)

6 Conclusion & Future Work

MyLake allows continuous data and information provided and shared by lake managers and stakeholders. The main objective of extending Mylake is to introduce a centralized integration platform that enables multiple integrations between different agencies. The existence of data silos (when integration is undertaken in silos) at many agencies can result in data inconsistency. With MyLake information shared is at the latest version, accurate and consistently retrieved from a single source of truth. Subsequently, MyLake catalyzes a seamless integration to reduce cost and the duplication of operation and maintenance work. MyLake

standardizes data through the platform that serves as an information sharing hub. The platform sends and receives standardized data from the provider agency through the platform. Through MyLake initiative, the public sector service delivery system has transformed into an integrated government digital service. MyLake users can use the centralized platform to simplify their data processes. This simplification allows agencies to focus more on their data analytics and insights, rather than on the arduous IT and integration development. Finally, MyLake delivers this convenience through a single platform that processes and integrates data for streamlined user efficiency. This has also fostered and strengthens the collaboration among participating agencies.

The platform provides potential use for stakeholders' strategic planning in water resources management process and avoiding in-silo integration efforts that may lead to interoperability and integration issues. By referring to a centralised big data integration platform that supports various formats of data and information, stakeholders and lake managers have a wider option to effective analytics on lakes including their catchment management.

Future work involves investigating an appropriate a big data integration method with subject matter experts and look into the algorithms and methods suitable for the domain. Active data integration between agencies is attainable through an appropriate platform. At the moment, data are uploaded manually through MyLake. Thus, extending MyLake as a central data exchange among agencies would be a start of transformation tool in reducing the wide gap on information sharing and data access where the more active contribution for lake basin management. Furthermore, MyLake would be able to bridge policymakers, lake managers, planners, NGOs, water resource authorities to work together with the public in general and lake community in particular to safeguard one of the main water resources in Malaysia.

REFERENCES

- [1] Cegarra-Navarro, J. G., Pachón, J. R. C., & Cegarra, J. L. M. (2012). E-government and citizen's engagement with local affairs through e-websites: The case of Spanish municipalities. *International Journal of Information Management*, 32(5), 469-478.
- [2] Mossberger, K., & Jimenez, B. (2009). Can e-government promote civic engagement? A study of local government websites in Illinois and the U.S. Institute for Policy and Civic Engagement, University of Illinois at Chicago. Retrieved December 12, 2018, from <http://www.uic.edu/cuppa/ipce/egovtreportfinal.pdf>
- [3] NAHRIM 2009. *Technical Report: Gedung*. Seri Kembangan, Selangor: National Hydraulic Research Institute of Malaysia.
- [4] International Lake Environment Committee Foundation (ILEC), 2005. Managing Lakes and their Basins for Sustainable Use: A Report for Lake Basin Managers and Stakeholders. Technical Report, 67-74.
- [5] Nakamura, M., & Rast, W. 2008. Key Challenges to Lake Governance. Japan: Shiga University.
- [6] N. Zulkifli, M.T.S. Mohammad and A.H. Hamzah (2015). Suggested Information Strategies for Malaysian Land Administration. *Global Journal of Business and Social Science Review (GJBSSR)*, Vol. 4(1), 380-389
- [7] A. Gandomi, M. Haider (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35, 137-144
- [8] D. Laney (2001). 3-D data management: Controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc. Retrieved from <http://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Variety.pdf>
- [9] H. Chen, R.H.L. Chiang, V.C. Storey (2012). Business intelligence and analytics: From big data to big impact *MIS Quarterly*, 36 (4), pp. 1165-1188
- [10] O. Kwon, N. Lee, B. Shin (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, 34 (3), pp. 387-394
- [11] Gartner IT Glossary (n.d.). Retrieved from <http://www.gartner.com/it-glossary/bigdata/>
- [12] H. Abbes and F. Gargouri (2016). Big Data Integration: A MongoDB Database and Modular Ontologies based Approach. *Procedia Computer Science* 96, 446 – 455
- [13] X. L. Dong and D. Srivastava (2013). Big data integration. *IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 1245-1248. doi: 10.1109/ICDE.2013.6544914
- [14] H.J., Oulidi and A. Moumen (2015). Towards a Spatial Data Infrastructure and an Integrated Management of Groundwater Resources. *Journal of Geographic Information System*, 7, 667-676. <http://dx.doi.org/10.4236/jgis.2015.76054>
- [15] G. Kang, J.Z. Gao and G. Xie (2017). Data-Driven water quality analysis and prediction: A survey. 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), 224-232.
- [16] J. Fan, J. Yan, Y. Ma and L. Wang (2018). Big Data Integration in Remote Sensing across a Distributed Metadata-Based Spatial Infrastructure. *Remote Sensing*. 10(1):7
- [17] M. E. Shafiee, Z. Barker and A. Rasekh (2018). Enhancing water system models by integrating big data. *Sustainable Cities and Society*, Volume 37, 485-491
- [18] A. Oussous, FZ. Benjelloun, AA. Lahcen and S. Belfkih (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*. Volume 30, Issue 4, 431-448
- [19] Y.F Huang, S.Y Ang, K.M Lee and T.S Lee (2015) Quality of water resources in Malaysia. *Research and Practices in Water Quality*
- [20] B. Arputhamary and L. Arockiam (2015) Data Integration in Big Data Environment
- [21] C. Yang, Q. Huang, Z. Li, K. Liu and F. Hu (2016): Big Data and cloud computing: innovation opportunities and challenges, *International Journal of Digital Earth*, DOI: 10.1080/17538947.2016.1239771