A Proposal on the Development Methodology and Data Governance for the Implement and Operation of Big Data System in Water Resources (based on K-water case)

By

SHIN, Yongwon

CAPSTONE PROJECT

Submitted to

KDI School of Public Policy and Management

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF PUBLIC MANAGEMENT

2020

A Proposal on the Development Methodology and Data Governance for the Implement and Operation of Big Data System in Water Resources (based on K-water case)

By

SHIN, Yongwon

CAPSTONE PROJECT

Submitted to

KDI School of Public Policy and Management

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF PUBLIC MANAGEMENT

2020

Professor Lee, Junesoo

A Proposal on the Development Methodology and Data Governance for the Implement and Operation of Big Data System in Water Resources (based on K-water case)

By

SHIN, Yongwon

CAPSTONE PROJECT

Submitted to

KDI School of Public Policy and Management

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF PUBLIC MANAGEMENT

Committee in charge:

Professor Lee, Junesoo, Supervisor Junesoo Lee

Professor Hwang, Phyll Sun

Approval as of August, 2020

1. Introduction

Big data means massive generated data in a digital environment that is vast in size, has a short cycle of generation, and includes not only numerical data, but also text and image data. In the big data environment, the amount of data has increased more dramatically than in the past, and the types of data are much various, so that it can analyze and predict not only people's behavior but also their thoughts and opinions through GPS and SNS (Jung, 2013).

Traditionally, the big data feature was the 3V (Volume, Velocity, Variety) mentioned above (O'Reilly Radar Team, 2012). More recently, value or complexity is added so 4V+1C. Such diverse and vast amounts of data has been attracting attention because of it can be an important and fundamental resource that determines the future competitive edge. In the past, there were attempts and efforts to analyze large amounts of data and find meaningful information. The current big data environment, however, represents a paradigm shift in terms of quantity and quality as well as diversity compared to the former. In this regard, big data is regarded as a key factor of innovation, competitiveness and productivity improvement in IT and smart revolution era, just like coal in the industrial revolution era (McKinsey, 2011). The Davos Forum in 2016 dealt with the fourth industrial revolution as a main agenda, which links, analyzes and uses a large number of data. Many developed countries have been innovating in all areas of society by using data. In 2011, McKinsey predicted that the big data generated $330 billion worth of value in the US healthcare sector, and saved €250 billion in the European public sector annually. It also expected big data also created additional demand of 140,000~190,000 data analysis professionals and 1.5 million data-based manager in the U.S. alone by 2018. On August 31, 2018, at the field visit for regulatory

innovation to revitalize the data economy, President Moon Jae-in said, "I want to change from a country that handles the internet best to a country that handles the data best and most securely". The data economy has become an inevitable trend worldwide, and Korea also makes and implements strategic plans. The IT industry shifts from hardware-focused industries in 2000s to software-focused industries in 2010s and to data-driven industries in 2020s. The Web sector is changing similarly. This is change from the stage #1 of simple delivery of information to the stage #2 of collective intelligence based on participation and communication to the stage #3 of creating convergence knowledge based on data analysis. It is unimaginable for us to live without data. In fact, big data is being used closely in our lives. Some banks in the U.S. let borrowers write the loan reasons, and analysis the words in the article to estimate whether the person will pay well or not. Semiconductor companies also find out where and what shape the defect occurred by analyzing semiconductor wafer photos.

These changes drive the adoption of big data in the public sector as well as in global companies. In the next computing paradigm which is focused on data, many countries are increasingly interested in using public data. The use of public information is more effective when integrated analysis with privacy information as well as public information itself. By finding meaningful patterns through monitoring social and natural phenomena, effective customized policies can be established. Table #1 shows a variety of available examples for the public sector. It can represent its potential value in a variety of areas, including social issues analysis, customized civil service, crime prevention & response, environment monitoring, and traffic management & optimization (Heo et al., 2011). For a specific example, "Owl Bus" in Seoul, which operates late at night, determines the route by analyzing the location information of telecommunication company's

customers. In fact, big data is being used closely in our lives. Therefore, managing and using data effectively is a core competence of both countries and companies and is directly related to future value and survival (McKinsey, 2011).

[Table #1] Potential big data application field in public sector.

Use Case	Contents
VOC	(data) VOC log, civil complaints, SNS, etc. (analysis) Discover policy tasks by understanding citizens' voices on specific topics.
Social issue	(data) Newspaper, Search word in Portal, SNS, VOC log, civil complaints, etc. (analysis) Automatic detection of regional social issues and trends of the related subjects, the establishment of a regional tailored strategy.
Medical and Welfare Service	(data) Medical & welfare expenditure, income, health insurance tax, etc. (analysis) Optimizing health insurance costs, monitoring unfair aid claims.
National R&D	(data) National R&D project proposals, reports, budget, patents, etc. (analysis) Optimizing National R&D, Future Technology and Gap.
Education Policy	(data) Education budget, reports, learning performance, SNS, etc. (analysis) Improving education environment, analyzing educational performance, and student concerns, and optimizing budget.
Crime Prevention and response	(data) Crime rates, police work records, reports, news, SNS, etc. (analysis) Crime pattern analysis by region, time and type, and Prevention Strategy.
Financial Supervisory, Tax Collection	(data) Tax data, financial transaction data, asset status, income and expenditure data (analysis) Anomalies in financial transactions, tax avoidance and evasion detection, and financial soundness analysis.
Environmental monitoring and response,	(data) Measuring instrument acquisition data, SNS, GPS, reports, etc. (analysis) Monitoring and responding to environmental pollution and establishing environmental policies.
Traffic situation mgt. and response	(data) Road sensor acquisition data, CCTV, accident record, weather, event, etc. (analysis) Traffic flow modeling, prediction and Optimization, and traffic signal system improvement policy.
Urban Control and Disaster Response	(data) Sensor data, CCTV, SNS, etc. (analysis) Monitoring of problems in cities, disaster detection, emergency assistance
Defense and National Security	(data) Defense reports, internal news, SNS and key facilities data, etc. (analysis) Monitoring in/out national security issues and planning defense policies

(Source: An easy-to-understand analysis guide for big data in the public sector (2012, NIA))

2. Big data in water resources sector

In accordance with the development of big data management and analysis technology, it is also actively introducing water management sector. K-water pushes forward smart water management by combining water data and big data technology. Smart water management refers to a next-generation water management system that implements informationization and intelligentization throughout the water sector, including management, production and transport of water resources, and treatment and reuse of sewage (Lee, 2015). This is based on data, and the convergence of distributed diverse data enables to create new value. Big data on water management in Korea is public data collected and managed by the government or public organizations for common purposes. It is individually utilized by many organizations for various purposes such as quantity management, water conservation, disaster and agricultural water (Lee, Kim, & Kim, 2016). Public data is more valuable than general data and can produce efficient policies by analyzing various information contained in public data (Lee, 2012). Although public water service organizations try to create new values by utilizing various water information for state-led smart water management, data-based analysis and application gradually reach the limit due to insufficient integration of quality data or management system. To this end, it is important to systematically manage and continuously optimize big data of water management. However, researches and investments in the one-time has been concentrated without establishing a whole strategy for the essential big data governance and management. This makes it difficult to realize sustainable big data. Big data governance refers to a management system that evaluates, directs, and monitors the entire process from collecting, analyzing, and using big data (Choi, Cho, & Lee, 2018). In the data collection and refining stages from various source, which account for more than 80% of big data analysis time,

quality are low due to the absence of standardization and temporal and spatial consistence. It cannot meet the needs of decision making. In order to solve these problems, someone must identify and improve them when continuously managing data quality and standardization. Ultimately, organizations and regulations need to be restructured as part of big data governance and support them.

The volume, velocity and variety of water data doesn't matter. It is time to think about how to improve our life quality by utilizing vast and complex big data in water resources sector as public data. To that end, the report researches how to integrate big data and establish an operating system in water management sector focused on K-water case.

3. Big data Policy and Technology Trends

3.1. International Policy Trends

According to the 2019 Data Industry White Paper published by the Korea Data Agency, the Obama administration's open data policy is strengthening its stance in the Trump administration as the OPEN Governance Data Act, which emphasizes machine-readability of public data, was enacted in 2019 following the implementation of the Data Act of 2014 to enhance transparency of government financial data. Recently, the government is developing innovative ways of using data collected by strengthening security.

In Europe, in order to secure the competitiveness of the data industry, it has been investing in the big data industry since 2014. Together with the Big Data Value Association, which covers businesses, research institutes and schools, the government will provide €2.5 billion in total in the form of public-private joint investment to energy, manufacturing and health by 2020. It also

focuses on increasing data utilization and developing business models for creating new values by enhancing free data access and data analytics capabilities among EU member states. In recent years, it has been legally ensuring the free use of personal data by strengthening protection and accountability related to data processing.

China is fostering big data as an emerging industry of national strategy. The 13th Five-Year Regulations published in 2016 specified the national big data strategy and emphasized the importance of developing the big data industry. Major policies include 1. Expanding the level of data opening, 2. Supporting innovation such as platform and open source technologies, 3. Improving the SW level of applications specializing in big data, 4. Upgrading the government support system and 5. Training of professionals.

3.2. Domestic Policy Trends

In Korea, the Ministry of Science and Technology plans to spend a total of ₩1.15 trillion over three years from 2019 on the establishment of a big data 100 centers where diverse data can be collected and provided by public and private sectors such as finance, communications and transportation, and a big data 10 platforms where quality data can be combined and distributed and new services can be created. Its purpose is to support the creation of an ecosystem that generates value based on data such as analyzing, distributing data collected from big data centers and discovering and spreading innovative services on platforms through cooperation between the public and the private sector. K-water is a performance agency in the environmental field among the 10 platforms selected this year, and will provide customized water information services and air quality outdoor activities by converging data such as water, weather and climate, fine dust, geology and disaster, ecology and resource, chemistry and material, and environment SNS. It also

works to establish a national research data platform, support data vouchers and create a transaction base. Other ministries including the Ministry of Public Administration and Security, the Ministry of Trade, Industry and Energy, the Ministry of Health and Welfare, the Ministry of Land, Infrastructure and Transport, the Ministry of National Defense and Statistics are also supporting the establishment and revitalization of big data. On January 9, 2020, the National Assembly passed revision bills such as the Personal Information Protection Act, the Information Network Act and the Credit Information Act. The key point is that if privacy information is processed safely and anonymously, data can be used for research, service or technology development without the consent of the person who provided the information. Another is to unify the distributed privacy protection system under the current law. A legal basis has been set up for many companies to use anonymous and anonymous information. It will be possible to find innovative services by converging data.

[Table #2] Key contents of three data acts

Act	competent department	Deregulation key contents
The Personal Information Protection Act	Ministry of the Interior and Safety	-Pseudonymous data can be used for commercial purpose. -Unify personal information management into a personal information protection committee
Credit Information Act	Financial Services Commission	-Allow big data analysis in the financial sector by using pseudonymous data -Pseudonymous data can be used and provided without the consent of the owner.
Information and Communications Network Act	-Ministry of Science, Technology, Information and Communication -Korea Communications Commission	Change to the personal information protection committee, which is authorized to supervise personal information online

As such, big data is the foundation of global innovation. Managing data well and fusing with each other to discover new values is a core competence of the state and enterprise, and it is directly related to future competitiveness.

3.3. Technology Trends

According to the 2019 data industry white paper published by the Korea Data Agency, the development of data collection and distribution technology has commercialized hub, tool and agent for integration between DBs in the entire process from generation to analysis, distribution and reflux. Therefore, it is possible to collect and analyze large amounts of data in real time and share or trade them. Trends in data analysis are integration and automation. In addition to the analysis itself, a technologies are being developed to support and automate data preprocessing or post-processing work comprehensively. Languages are analyzed using R and Python. Use cloud computing for large-scale analytics in open-source environments. Common cloud service providers such as Amazon Web Services, Microsoft Assert, Google Cloud and others support their own analytics software along with open source software. Machine learning is actively adopted for predictive analysis. As the development of deep learning technology that requires a lot of complex and difficult resources, 75 percent of major companies will use the technology in 2022. Accordingly, many machine learning framework technologies have been developed to support these quickly and easily. In addition, as an alternative to Black-box type AI, which only shows the result without explaining the process, White-box type AI, which can explain how to make the results, is developed actively. It makes easy to analysis the causes of phenomena. So far, there have been difficulties in collecting and analyzing vast amounts of data, but recent technological developments seem to have solved some of these difficulties.

4. Big data in water resources

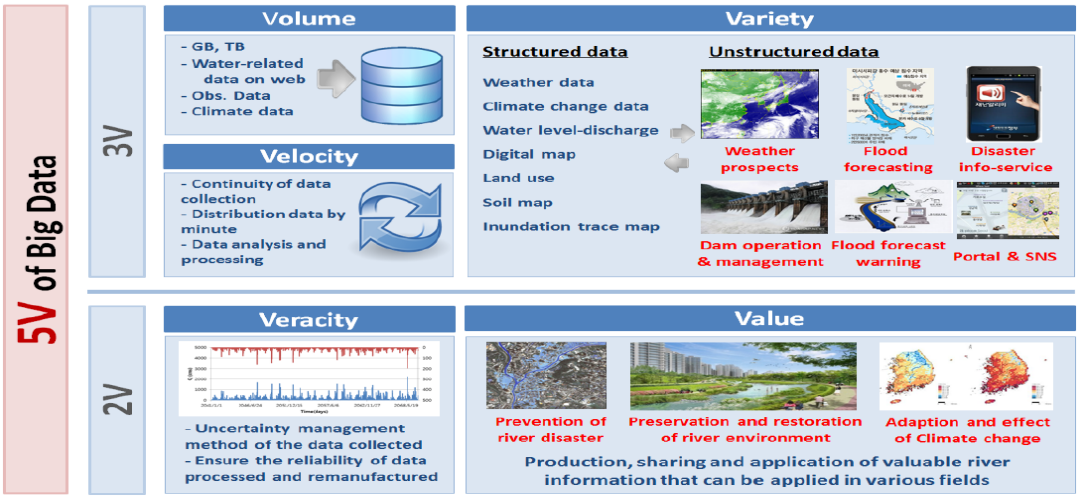
Nowadays, it is essential to introduce big data to get more accurate and valuable information. It has continued to be argued in previous studies. Big data solves problems in the field by analyzing customers' data generated from various channels. It also supports scientific and rational decision making by using non-intuitive data for important decision making. Therefore, Big Data leverages new technologies to store and analyze data, provide insights to support decision making, support processes and optimize work (Sirisha, 2017). In the future, a strategy is needed to recognize big data as a key factor, to build infrastructure to extract necessary information, and to secure external big data resources. It is very important to systemize data management and ensure data reliability (Jung, 2012).

4.1. The Necessity

The need for big data has also been studied a lot in the water resources sector. Kim, Kang, Jung & Kim (2016) studied that big data and cloud computing-based water resource information management was needed to solve water management problems. Water resource information satisfies all 5V (Volume, Velocity, Variety, Veracity, Value) attributes of big data. The effect of water resources information management is as follows by introducing big data technology due to limitations in the management and utilization of the existing method. 1. It is possible to derive valuable new information by adding unstructured data to existing structured data and to make various and wise situation judgments in water resources planning and decision making. 2. It enables rapid status sharing and decision making through real-time collection and analysis. 3. Not only water quantity and quality but also weather, disasters, ecology and climate change can be

used in a wide variety of fields, and many data can be created to extract new values through data convergence.

[Figure #1] Big data 5V properties in Water resources



(Source: A Review on the Management of Water Resources Information based on Big Data and Cloud Computing (Kim et al. 2016))

Lee & others (2017) studied that water management technology such as water pipe spatial analysis, precision leak detection and pipeline performance evaluation should be secured by using big data analysis based data acquisition and collection in order to strengthen core competitiveness of water pipe maintenance. In addition, Jung, Jang & Kim (2019) introduced various international and domestic cases to utilize big data in the hydraulic model experiment. Furthermore, he argued that the use of various field observations data such as flow rate, sediment discharge and water level as big data could contribute greatly not only to academic development in hydro engineering field but also mankind.

4.2. The Difficulty and Governance

There have also been studies on the current problems of data management in Korea and the need for governance to use big data. Kim, Kang, Jung & Kim (2016) said that in 2010, Kim found that

considerable investment and effort are needed to accumulate long-term data with high quality levels in the water resources sector, but due to lack of awareness and effort, there are many difficulties in various hydrologic-related research and design. In 2016, Lee, Kim and Kim found that Korea's water resources information service systems which are related to flood, weather, water quality and quantity is currently developed by each ministry in a way to build and operate individual systems and further specialize them. It has contributed to the establishment of infrastructure for water resource information service. However, as the importance of water resource management increases due to climate change, traditional approaches have reached the limits. Integrated management between related organizations is essential for efficient water management. Big data technology can solve it.

[Table #3] Current Status and Problems of Water Resources Information

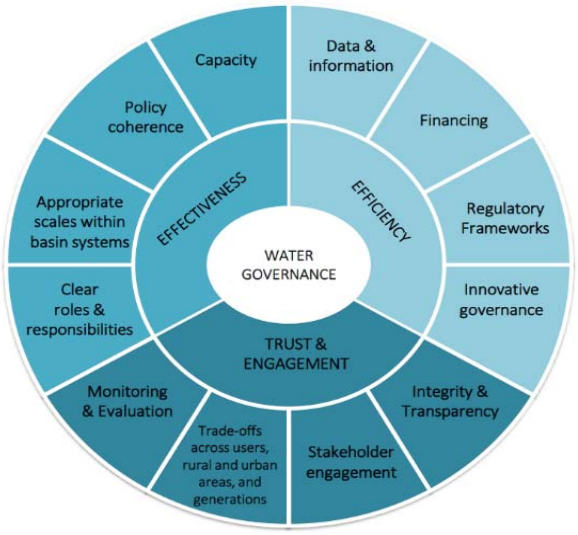
Aspect	Explanation
Data Collection & Storage	Each organization collects data through agreement with other organizations and stores it in its own database. However, data redundancy and data quality management are difficult
Data Disclosure & Common utilization	Multilateral data linkage is only data linkage for some functions, so master plan for data link and information sharing is urgently needed at the national level.
Data prediction & Operation	Without securing the accuracy of the data received from other organizations, it is difficult to make predictions that fit the organization's role.

(Source: Hydroinformatiics (2016, Lee, Kim & Kim, Water for future))

Heo et al. (2013) said "Public institutions used to create, collect and manage data at a high cost for their original public functions, and when that purpose is achieved, these are discarded or neglected, but recently the new technologies such as internet, open API and mash-up have opened the door for new uses of public data". However, Kim (2019) stated "Generally, the reasons for the low usage of public data include policy factors that do not allow the data required by the private

sector to be opened in a timely manner, low data quality, and technical elements for data integration". Jung (2012) argued "In the future, strategies are needed to recognize big data as a core resource, to expand resources to extract necessary information and to secure external big data resources. Data quality has a significant impact on its utilization results, so it is very important to systemize data management and ensure data reliability". In other words, data quality control is important as big data increases the utilization of public data. This can be managed systematically through big data governance. Choi, Cho, & Lee (2018) suggested a big data governance model for smart water management in a three-dimensional cube format. The X-axis has four types of big data needed for water management, the Y-axis has six types in the water industry, and Z-axis presented data as eight elements for management. They emphasized the need for governance to improve the quality and operational performance of big data. OECD (2015) also recognized the necessity of water-related governance and established principles for the promotion of water management policies. The Water Governance Principle seeks to complementary public policy in three ways. 1. Effectiveness is to define clear and sustainable water policy objectives at all levels, to implement policies and to achieve them. 2. Efficiency is to maximize the benefits of sustainable water management and welfare at minimal cost. 3. Trust and participation are to build public trust and ensure stakeholder inclusiveness through legitimacy and fairness in society as a whole.

[Figure #2] Overview of OECD Principles on Water Governance (Source: OECD (2015))



5. Changes of the Data Usage in the Big Data Era

Ann Winblad, the legendary investor and senior partner at Hummer-Winblad, said "data is the new oil." in a CNBC talk show, 2012. On May 6, 2017, the Economist published an article titled "The world's most valuable resource is no longer oil, but data. This is the age of software and data. Cleary, software is useless without data. For example, AI learns data and creates algorithms. Companies with more data gain bigger markets, and countries that produce and use data well develop economically. For this reason, companies have begun to take over and merge to secure more data, and countries have begun enforcing regulations to protect their own data. The data economy is already in full swing.

5.1. Data usage trend

Data usage is different depending on the needs of the times. Let's take a look at the changes in data usage flow from the initial simple data processing to the age of data connection. In the data-processing era, computer programming languages made it possible to process large amounts of data quickly and accurately. Companies computerized existing handwriting tasks and applied them to tasks such as payroll and accounting. The data did not provide new value for business processes.

In the age of data integration, previously accumulated data lacked consistency because it was generated from each business perspective. To solve this problem, data was integrated through business process reengineering and DB modeling. Since then, it has been used in some areas such as data inquiry, report making, and cause analysis.

In the age of data analysis, organizations that find insights from data and use them for their

business began to performance better. AI technology has also been developed to learn data and make accurate decisions faster than experts.

In the age of data connectivity, an important feature is super connections. Many companies open their services and data with open APIs. Even the government and public organizations are providing open APIs. Currently, providing open APIs depends on the discretion of the companies concerned, but it is gradually expanding.

5.2. Changes in data usage patterns

Up to the age of data analysis, the data life cycle is summarized as the process of data → information → knowledge → wisdom transformation. On the other hand, in the big data era (the age of data connectivity), the data life cycle is summarized as the process of data → insight → execution (Park, & Kim, 2016). These differences in transformation lead to differences in data resource management that support the data lifecycle. In the past, data became less effective after a certain period of time. Therefore, this data was not transferred separately in the next generation system project and DB infrastructure improvement project. Many of them only kept the originals to be viewed when needed or only to view the originals with the existing system. But as we moved into the big data era, data has always been used. There is no expiration date for the effectiveness of the data.

6. In the Big Data Era, the limitations of traditional systems development and management.

Since the development of a centralized system based on data in the 1970s, there have been significant changes in data and system development processes. Through this, a lot of data have

systematically been accumulated that can be used in big data era. In addition, systems have evolved from simple task support and data storage to innovations in business processes and analysis for decision making. As a result, system building methods and data management process have been improved.

6.1. CBD (Component Based Development) Methodology

As the scale and complexity of software development increased, the theory of development methodology emerged with the need for systematic development such as communication, efficiency of development process, and quality securing. It evolved with technology development and the trend of the times. The theories on structural methodologies (1970s), information engineering methodology (1980s), object-oriented methodology (1990s), and CBD methodology (2000s) emerged and field works applied these theories. The government also distributed CBD SW Development Standard Output Management Guide (2011.12, NIA). The standard for development methodology and the essential outputs for development were selected, and the association and system between outputs were established to establish their role as a guide for analysis and design. Let's learn how to develop CBD, it is a development method for creating another component or software by creating and assembling each component that constitutes a system or software, such as a Lego block (Jeong, 2002). A component is a software block and an independent software package to create another component or an application. This is a way to reduce software development time, lower development costs, and ease maintenance by ensuring reusability and interoperability of component software modules. K-water's development standards are similarly operated by referring to them.

6.2. Data Operation & Management focused on task process

The IT governance has been established and operated for the efficient introduction and operation of IT. IT governance is the structure and process of relationships for developing, directing and controlling IT resources so that IT can achieve business goals by adding value to the organization (Kakabadse, & Kakabadse, 2001). In today's IT governance, data is handled by system functions in accordance with each task process and registered at each business stage. These processed data are stacked and used for reporting and statistics. For example, in a financial system, a budget is allocated for a purchase and a statement with details is entered and recorded for each purchase. This makes it easy to manage budget execution, asset acquisition, and expenditure history. Data management is also managed around consistency and redundancy at each stage.

6.3. Limitations of System Development and Operations

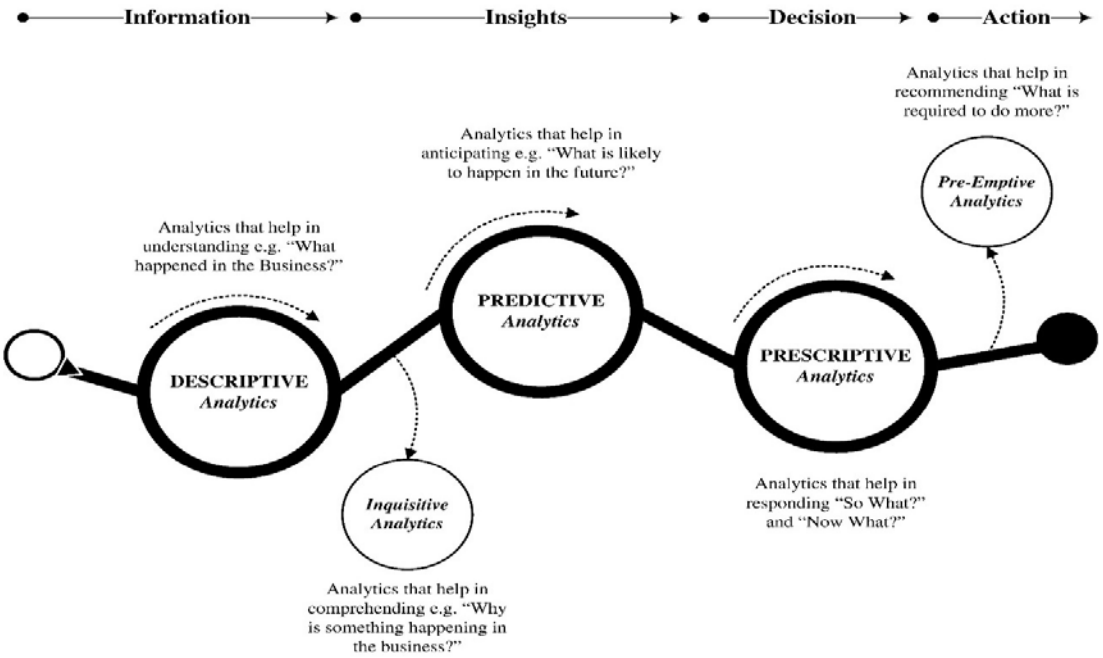
The system development perspective so far is function-oriented. This is not suitable for the process of data collection, storage and analysis of big data.

[Table #4] The Differences between Traditional System and Big Data System

	Traditional System	Big Data System
Purpose	Data accumulation, Improvement of business process	New insights
Data Type	Structured	Structured, Semi structured, Unstructured
Data Acquisition	Key-In, Upload	Crawling, Batch, Real-time linkage
DB Management	Relation DBMS	NoSQL, Hadoop
Feature	Menu arrangement and data CRUD (create, read, update, delete) based on business process.	Menu arrangement based on data and technique (analysis, visualization)

The possibility of using big data in all areas of society, such as national safety and security, healthcare, education, welfare, and the environment, is increasing and the actual demand is increasing. But, practical methodologies to drive it remain in the past. Guide lines are needed specifically for procedures and methods for effective application to big data operations. Sivarajah, Kamal, Irani, & Weerakkoday (2017) classified big data objectives as figure #3. As you can see here, big data system's aim is to discover new values, regardless of the business process. In order to achieve this goal, the appropriate system development and data governance are required.

[Figure #3] Classification of types of big data analytical methods



(Source: Sivarajah, Kamal, Irani, & Weerakkoday (2017))

7. Development Method and Governance in Big Data System

So far, the perspective of system development and data management is function-oriented. This is not suitable for big data. Its process is data collection, storage and analysis. This chapter suggest

system development and governance methodology needed to implement and operate big data system.

7.1. Development Method of big data system

The development methodology must be changed because there is a difference in the purpose of use from the traditional system that innovates the business process and enable staff to work more easily and quickly.

1. Analysis – Define User Requirements

Understand and Define issues that are intended to be solved through big data analysis and utilization clearly. Based on a basic understanding of the business (domain), identify requirements through meetings with the business and IT staff. Collect and summarize related data to establish the scope and goals for the analysis subject.

2. Analysis - Analyze Current Systems

Grasp legacy systems such as meta-data management system, big data portal in relation to requirements and each configuration by using big data.

3. Data Preprocessing – Define analysis data

Specify the data lists obtained in the requirements analysis phase to summarize the data collection targets within the analysis scope, and investigate the authority institution of data and the linkage methods.

4. Data Preprocessing – Configure data Gathering environment

Set up system environment to collect internal and external data using existing big data infrastructure with IT staffs

5. Data Preprocessing – Collect data

Ask for data related IT staffs and configure system environment to collect internal and external data using the own big data infrastructure. If the data cannot be obtained through system linkage, request data from owners.

6. Data Preprocessing – Verify collected data

Check for data gathered from various channels to determine if there are errors such as false reading, missing value or outliers. Data is refined by this verification.

7. Data Preprocessing – Structure and save data

Standardize the data structure and define the primary and reference key to match. Store data and manage its quality. If data acquisition is performed repeatedly, recommend to implement automation process to collect, verify, and store data.

8. Modeling – Explore data

Explore the entire data to identify patterns, characteristics, and points of interest based on an understanding of the domain in the field. This helps to reduce the amount of data to manageable size and focus efforts to optimize analysis.

9. Modeling - Modeling

Analyze data in multiple perspectives and methods for the purpose of analysis and explain

interrelationships between data logically. Design analysis models according to the purposes of four analysis types (descriptive, diagnostic, predictive, and prescriptive). Implement analysis model after improving degree of completion through various verification.

10. Build– Design Algorithms

Review whether the analysis techniques used in modeling are applicable and implement algorithms so that the modeling results can be executed on computer actually.

11. Build – Model Implement

Determine which modules to apply the algorithms and implement the program accordingly. Recommend to use existing packages or develop new packages depending on repetition and unit job. This implements a new analysis model.

12. Build - Visualization

Express the analysis results in charts, diagrams or pictures so that anyone can easily understand them. Design and development visualizations to consider results from various perspectives and discover patterns. Visualization results should be available in reports easily.

13. Build - Data sharing

Share refined and high-quality big data internally or externally, if necessary. In this case, comply with data sharing standards to use simple. It uses Open API the most.

14. Test, Delivery

It is similar to test step of traditional system development. So the details would not be mentioned in this study.

[Table #5] The Differences between Traditional System and Big Data System Building Process

	Traditional system	Big Data System
Analysis	Define user requirements → Analyze current systems → Process/Entity Modeling	Define User Requirement → Analyze Current Systems
Design	Technical Architecture → UI / DB Design → Function Specification	(Data Preprocessing) Define Analysis Data → Configure Data Gathering Environment → Collect data → Verify collected data → Structure and Save data (Modeling) Explore data → Modeling
Build	Secure Development Environment → DB creation → Programming Implement → Function Verification	Design Algorithms → Model Implement → Visualization → Data Sharing
Test	Test Scenario → Unit Test → Integration Test	Same as the left
Delivery	Deployment → Beta Test → User Training / Manual → Launching	Same as the left

7.2. Big data Governance

Building big data governance is necessary for efficient big data operation and management. It is different from previous simple data management. Data management focused on quality and maintenance of data itself. Governance, on the other hand, is a holistic approach for managing the people, processes, and technologies that operate big data, including data management, and establishes rule, roles, policies, and cultures. It contributes to the achievement of new values and enterprise strategies.

1. People: Adapting to R&R, culture of data era. It need technology education and awareness improvement of big data to create organization culture for new data age from CEO to new employee. The core stakeholders, such as data architect, manager, modeler, owner for main data, provide training on R&R in addition to new technologies.

2. Process: Adapting to new method and step for using big data. It manages policies of data governance operation, R&R for each position, and the task process. The governance committee makes and changes key policies and processes fairly.

3. Technology: Adapting to new approach and function for data analysis. It manages the current state of big data-related technologies those the institution has implemented and used. It also reviews new technology trends and best practices and discovery the new sector or task for using them. This establishes a system in which new technologies can be continuously utilized within the institution.

4. Data Management: Adapting to management method for securing data quality and architecture. It is about direct operation of big data and consists of 6 elements.

- Master Data: It manages the enterprise data architecture to monitor whether the actual DB operates according to enterprise DB standard and to review and handle requests for adding new DB table and field.

- Meta-data: It manages and updates the detail descriptions of each data so that users can easily find what data exists, what it means and where it is.

- Quality/Quantity: It manages the quality of input data from the perspective of completeness, validity, consistency and completeness.

[Table #6] Characteristics of data quality elements

Element	Criteria	Inspection
Completeness	individual completeness	Check if required column values are missing
Effectiveness	Range Validity	Determine if column value is within given range
Consistency	format standardization	Verify if the format of values such as date and time is standardized
Complementarity	error throughput rate	Check if correcting error and missing data

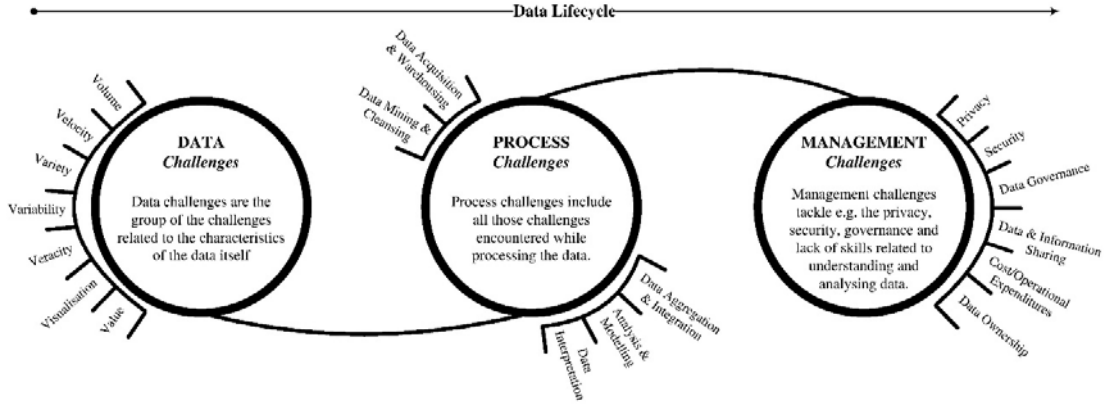
It manages the quantity of input data. Big data has very big size and variety. Thus, the institution introduce distributed DB, such as Hadoop, rather than traditional relational DB for data analysis, and GPU for supporting parallel processing. It manage them to be optimized through monitoring.

- Integration: It manages that data collected from inside and outside can be integrated for data analysis.
- Security: This manages compliance with data security policies, such as leakage prevention and user consent, in the entire process from data acquisition to utilization.
- Lifecycle: This manages the data usage process that leads to creation, storage, utilization and disposal. Policies focus on the discarding cycle for each data.

7.3. Correlation between Challenges and Development method & Big Data Governance

Sivarajah, Kamal, Irani, & Weerakkoday (2017) defined challenges in the data life cycle of the big data era. There are challenges in all areas, from data to management. The development method and governance in this study can easily address these challenges. As shown in the table below, all 18 challenges can be solved. This indirectly confirms the excellence of this study's result.

[Figure #4] Conceptual classification of Big Data challenges.



(Source: Sivarajah, Kamal, Irani, & Weerakkoday (2017))

[Table #7] Matching Solution for Big Data Challenges

	Challenges	Solution	
		development method	Big Data governance
data	volume	data Preprocessing	Quality/Quantity
	velocity	data Preprocessing	Quality/Quantity
	variety	data Preprocessing	Quality/Quantity
	variability	data Preprocessing	Quality/Quantity
	veracity	data Preprocessing	Quality/Quantity
	visualization	Build	Technology
	value	Analysis	People
process	data acquisition & warehousing	data Preprocessing	Master Data
	data mining & cleaning	data Preprocessing	Lifecycle
	data aggregation & integration	data Preprocessing	Integration
	analysis & modelling	Build	Technology
	data interpretation	Modeling	People
management	privacy	-	Security
	security	-	Security
	data governance	-	process
	data & information sharing	-	Meta-data
	cost/operational expenditures	-	committee
	data ownership	-	process

8. Big Data Operation Case in K-water

8.1 Big Data Governance in K-water

K-water has recognized the data importance from old times and systematically managed the operation data of dams and water facilities through ICT. It launched My Water, a public water information service portal, in 2016 and prepared big data ear. Since then, the company has operates big data governance by reorganizing and improving various policies and systems. Let's take a look at how K-water operates each component from perspective of big data governance.

1. People: K-water has operated many training courses for big data since 2016 by level (basic, intermediate, advanced and expert), and type (on-line, off-line, reading, hybrid). Especially in the expert course, students learn big data analysis for a year and find model to solve actual business problems through individual long term-project. The excellent students have a benchmarking opportunity of global best practice. In addition, it has held in-house Big Data contest to directly perform required analyses in the field, share analysis results, and expand the outstanding model to the entire company since 2016. These activities also improve employee awareness and skills and make new culture.

2. Process: K-water establishes and operates an internal regulation for big data operation & management. Data management regulation include data management system, planning, quality control, standard management, data sharing & delivery, and operation of data management committee. K-water also seeks to promote systematic and efficient management, provision and utilization of stored, utilized and shared data, those are made from in-house business process. Since 2018, a dedicated department has been established

to accelerate data-driven working culture. In addition, the Data Management Committee deliberates and adjusts the necessary matters for the efficient management of data life-cycle.

3. Technology: K-water has had many efforts to internalize big data technology. PhDs in each field of research institute conduct data analysis directly through big data learning. It make in-depth data analysis possible based on the domain knowledge. To predict future phenomena for scientific water management, customized PDDs (Physical Driven Model) have been used, but recently, DDMs (Data Driven Model) have been developed by using big data analysis. So two models are used to increase the efficiency of water management. Through skill training, K-water actively support employees' growth so that they can analyze field problems using various tools, including Excel, R, and Python. This fosters domain-based analysis experts. Last year, it implemented a big data analysis platform to make modeling easy and convenient for staffs. Now, every employee can make analysis models and visualize them with simple click and drag even without knowing the R and Python program coding. In addition, it continuously adopts new technologies based on medium and long term technology strategy.

4. Data Management: As the big data's importance grows, K-water, which has lots of measurement data, has recently made great efforts to secure the raw data's reliability.

    - Master Data: Like other companies, K-water did not have enterprise data architecture in 1990's. Since then, it was difficult to manage the data of the increased systems and even after establishing enterprise data architecture, it was impossible to change legacy system according to it. However, the new data architecture was established through the introduction of the enterprise ERP system from 2017 and the DB of all K-water

systems was changed accordingly. In addition, Master Data Management System is developed and operated in 2019 for efficient architecture management.

- Integration: As ERP was introduced, the DB is made according to data architecture and data mart is operated through this. In 2019, K-water develops big data platform, which makes easy to use data by integrating external data as well as internal data. It also manages meta-data through the Master Data Management System so that it can easily find data and data meanings by reflecting users' needs.

- Security: The leakage of personal information and corporate core data has become a major issue in determining maintenance or abolition of enterprises. K-water has operated integrated control and monitoring through the Cyber Control Center since 2008, and doubled its security capability through joint control with the central government. It is under constant management through periodic inspections by the NIS and the Ministry of the Interior and Safety. It is evaluated as an excellent inspection institution every year.

- Quality/quantity: The SCADA (Supervisory Control and Data Acquisition) system has been in operation since late 1990 by introducing ICT to dams and water facilities. As the data’s importance grows continuously, the amount of measurement equipment and collected data has increased and quality has become more interesting. Data is accumulated in the headquarters in real time and a calibration system is established to check missing or false reading data by operator. Recently, it developed calibration technology that uses AI technology. Until now, only the total search function was supported for unstructured data (e-approval documents, KMS, company regulations /

notice board etc.), but this year K-water will develop analysis functions to infer the meaning from unstructured data. In the future, it plan to expand to collect and manage non-electronic documents and voice data.

- Lifecycle: Data generated for business processing is preserved permanently in principle. Even if the data architecture changes, such as the introduction of the ERP system, the data is transferred and managed through migration. However, in the case of measurement data, if the purpose of monitoring was completed and it was meaningless for analysis or statistics, K-water deletes them periodically according to its standards.

8.2 Big Data System in K-water

[Figure #5] Simple relationship of Data Application Flow in K-water

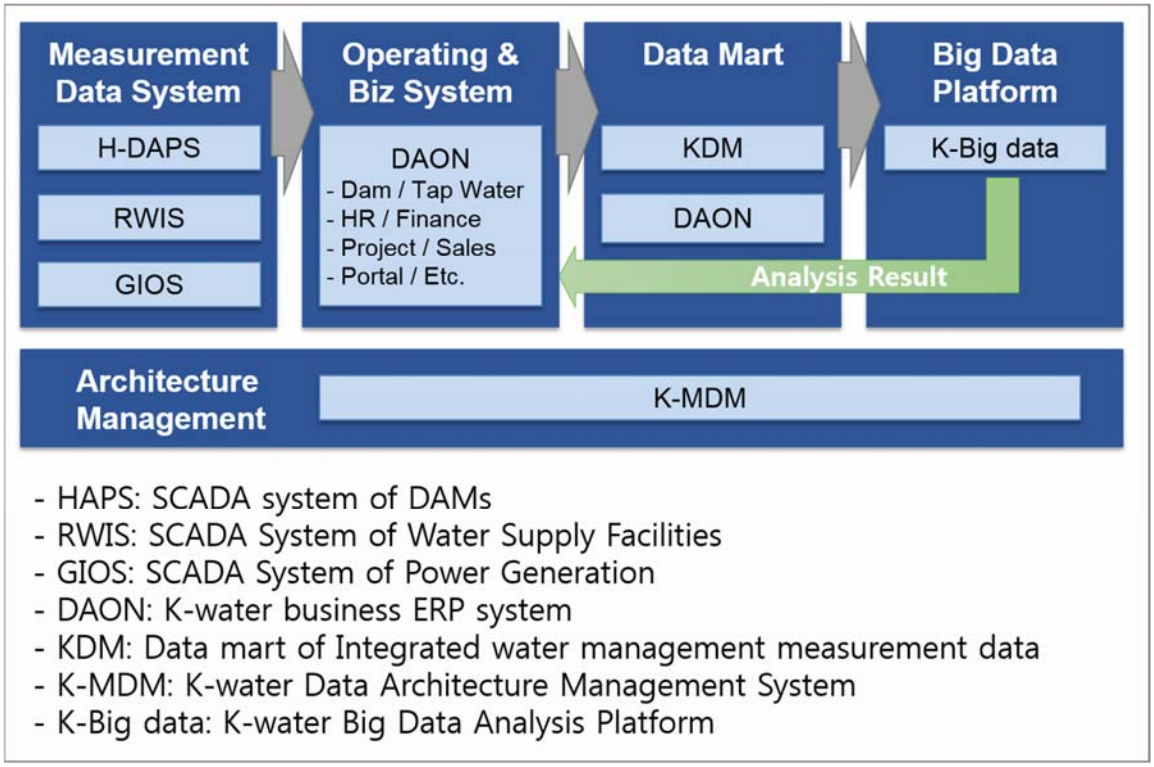
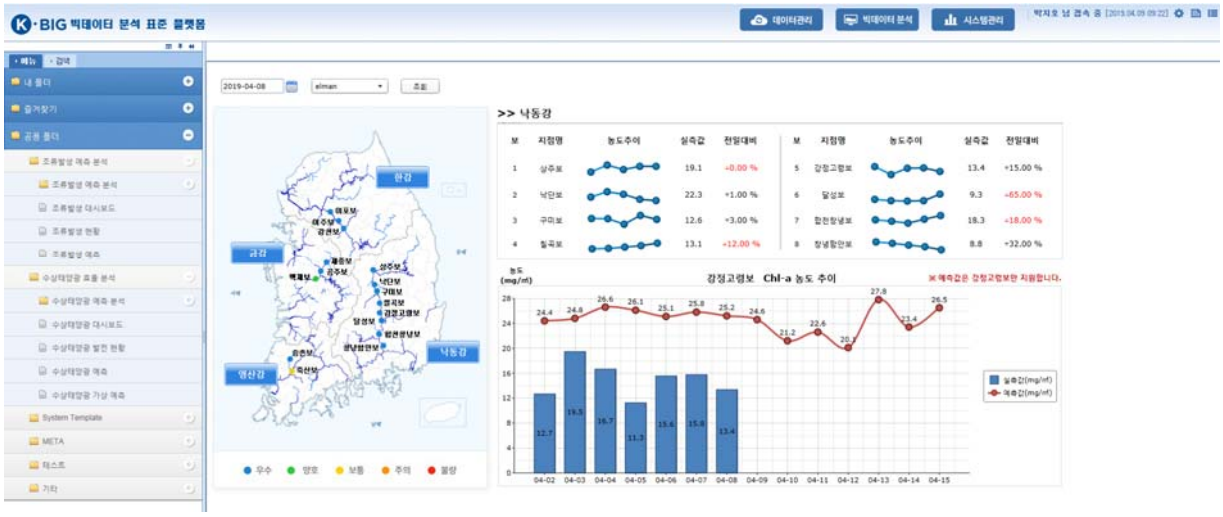


Figure #6 shows a brief relationship with K-water major systems. Measurement data system is FA part. It is acquired data for monitoring control of major facilities. This data is used as operating data for OA part. In addition, another data is accumulated according to business processes in various fields such as HR and finance. Next is data marts. KDM provides measurement data of water management. DAON as ERP also provides own data as Data Mart. K-water Big data Platform uses this data and external data if necessary for analysis and provides analysis results back to the operating system. K-MDM manages the foundation of all data for enterprise architecture management and quality. Figure #7 is Big data Platform's snap shot. It shows the model's analysis result. There are various analytical models on the left menu. Each time the model is clicked, the analysis results are visualized in various forms (bar-chart, pie-chart, histogram, etc.) so that the user can easily understand them.

By 2023, K-water plan to make more complex and in-depth analysis model by adding various analysis functions such as unstructured data analysis, voice recognition and image recognition, etc.

[Figure #7] Snap Shot of K-water Big data Platform



Reference

Jung Y.C (2013). Big data. Communication Books.

O'Reilly Radar Team (2012). Planning for Big Data. O'Reilly.

M. James, C. Michael, B. Brad, B. Jacques, D. Richard, R. Charles, & H. B. Angela (2011). Big data: The next frontier for innovation, competition, and productivity. Mckinsey.

Heo G.M., Park W.J., & Cho G.S. (2013). Domestic Market and Economic Impact of the Re-use of PSI (Public Sector Information). Electronics and Telecommunications Trends.

An easy-to-understand analysis guide for big data in the public sector (2012). NIA.

Lee S.H. (2015). Global Competition, Smart Water Services. Journal of Water Policy and Economy, 24, 5-14.

Lee J.H., Kim T.H., & Kim J.W. (2016). Hydroinformatiics. Water for Future, 49(11), 31-39.

Lee M.J. (2012). Big Data Analytics and Utilization of Public Data, The Korean Institute of Information Scientists and Engineers, Communications of the Korean Institute of Information Scientists and Engineers, 30(6), 33-39

Choi Y.H., Cho W.S, Lee K.H. (2018). Big Data Governance Model for Smart Water Management. The Korean Journal of BIGDATA, 3(2), 1-10.

Data Industry White Paper (2019). Korea Data Agency.

Sirisha Adamala (2017). An Overview of Big Data Applications in Water Resources Engineering. Machine Learning Research, 2(1), 10-18.

Jung J.S. (2012). Three Elements for Successful Big Data: Resources, Technology, and Manpower. IT & Future Strategy, 3.

Kim Y.S., Kang N.R., Jung J.W., & Kim H.S. (2016). A Review on the Management of Water Resources Information based on Big Data and Cloud Computing. Journal of Wetlands Research, 18(1), 100-112.

Lee H.D., Choi C.H., Gwak P.J., Knag J.M., Bang S.H., Kong M.S. (2017). Technology plan to improve water flow rate in water supply pipeline using big data. KICT.

Jung J.S., Jang J.R., Kim H.T. (2019). Directions for utilizing big data in hydraulic model test data. Water for Future, 52(1), 52-61.

Kim H.L. (2019). Data of AI era: the present and future of public data quality. NIA, Future2030, 7.

OECD (2015), OECD Principles on Water Governance, OECD.

Is Data The New Oil? Retrieved Apr 2, 2012, from https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#5f254dcf7db3

Park J.S., & Kim I.H. (2016). A Study on Data Resource Management Comparing. The Korean Journal of BIGDATA, 1(2), 91-102.

CBD SW Development Standard Output Management Guide. (2011). NIA.

Jeong U.S. (2002). Next Generation Software Development Methodology, Digital Content, Digital Contents, 12, 94-97.

Kakabadse, N. and Kakabadse, A. (2001), IS/IT governance: need for an integrated model, The International Journal of Effective Board Performance, 1(4), 9-11.

U. Sivarajah, M. Mustafa Kamal, Z. Irani, & V. Weerakkod (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286.