Big Data Platform Based on Hadoop and Application to Weight Estimation of FPSO Topside

Seong-Hoon Kim¹, Myung-Il Roh¹²*, Ki-Su Kim¹, and Min-Jae Oh²

¹Department of Naval Architecture and Ocean Engineering, Seoul National University, Seoul, 08826, Korea
²Research Institute of Marine Systems Engineering, Seoul National University, Seoul 08826, Korea

(Manuscript Received December 21 2016; Revised January 8, 2017; Accepted February 24, 2017)

Abstract

Recently, the amount of data to be processed and the complexity thereof have been increasing due to the development of information and communication technology, and industry’s interest in such big data is increasing day by day. In the shipbuilding and offshore industry also, there is growing interest in the effective utilization of data, since various and vast amounts of data are being generated in the process of design, production, and operation. In order to effectively utilize big data in the shipbuilding and offshore industry, it is necessary to store and process large amounts of data. In this study, it was considered efficient to apply Hadoop and R, which are mostly used in big data related research. Hadoop is a framework for storing and processing big data. It provides the Hadoop Distributed File System (HDFS) for storing big data, and the MapReduce function for processing. Meanwhile, R provides various data analysis techniques through the language and environment for statistical calculation and graphics. While Hadoop makes it easy to handle big data, it is difficult to finely process data; and although R has advanced analysis capability, it is difficult to use to process large data.

This study proposes a big data platform based on Hadoop for applications in the shipbuilding and offshore industry. The proposed platform includes the existing data of the shipyard, and makes it possible to manage and process the data. To check the applicability of the platform, it is applied to estimate the weights of offshore structure topsides. In this study, we store data of existing FPSOs in Hadoop-based Hortonworks Data Platform (HDP), and perform regression analysis using RHadoop. We evaluate the effectiveness of large data processing by RHadoop by comparing the results of regression analysis and the processing time, with the results of using the conventional weight estimation program.

Keywords: Big data, Hadoop, Weight estimation, FPSO

1. Introduction

1.1. Research Background

Big data is not simply a large amount of data, but a concept that encompasses all processes from collection to the processing of big data. Gartner, the well-known information technology research and advisory firm in the United States, has defined big data as an information asset with 3V characteristics (volume, velocity, and variety) that is used for better decision making. Here, the ‘volume’ refers to processable data larger than 100 TB (Tera Byte). The ‘velocity’ refers to high-speed real-time processing of data, while the ‘variety’ refers to...
the types of processable data, such as structured and unstructured data. In addition to 3V, big data is recently referred to as having the characteristics of 5V, which extends 3V to include veracity and value. The ‘veracity’ refers to high quality data selected from the entire data, while the ‘value’ refers to useful information deduced by big data analysis.

Recently, as Information Technology (IT) is constantly developing, the amount and complexity of data have been increasing. Therefore, industry’s interest in big data technology is increasing. In industry, big data is mainly used in the distribution and service sectors, but efforts are being made to utilize it in the manufacturing sector. For example, Volvo, a world-renowned automotive company, applies sensors to every vehicle, and collects big data to detect and deal with defects while moving. General Electric is making profits using process automation and optimization through big data analysis. In addition, big data is being used in various manufacturing fields, and its utilization method is being studied. According to Lee (2013), big data will be more important in the manufacturing sector, due to the explosive increase in the amount of production data through sensors, and the convergence of future industry with big data.

As in other manufacturing industries, the shipbuilding and offshore industry is generating considerable data in its design, production, and operation processes, and is interested in gaining market competitiveness using these data. If this big data can be effectively utilized in the shipbuilding and offshore industry, it will be possible to increase productivity and efficiency in various tasks, such as new ship design and development, process optimization, substitution of model test, and weight estimation. Therefore, this study proposes a platform using Hadoop, which is a representative big data framework, and applies it to the weight estimation of offshore structure topsides, to examine the applicability of big data in the shipbuilding and offshore industry.

1.2. Related Works

The importance of big data in the shipbuilding and offshore industry is growing, but research on big data is still in its beginning stages. Therefore, we have broadened the scope and analyzed researches related to big data in related industries, as well as the shipbuilding and offshore industry. Some of the analysis results are summarized below.

Many of the researches about big data merely suggest that big data needs to be applied in some way, as its importance grows. Kim and Kang (2014) summarized the trends of cloud computing technology, big data, and big data solution, and claimed that the related outcome should be accomplished by providing big data as a cloud service. They also summarized current big data analysis techniques, and discussed the importance of future technology prediction and big data analysis techniques.

In the field of shipbuilding and offshore engineering, there have been researches on design, production, and operation (shipping); and researches on the importance and application of big data in each field. In the field of design, Kim et al. (2013) identified the technologies and requirements to support ship design automation, and studied big data technology and analysis techniques. In the field of production, Lee (2014) researched the processing of data of the entire production cycle to derive meaningful information, and improve production management capabilities. In the field of operation, Kim (2014) emphasized the necessity of developing technology for applying big data collected during the operation of ships to the field of optimal operation and hull design. Perera (2017) proposed a big data handling framework for harnessing ship performance and navigation monitoring data. Wang et al. (2015) proposed a new framework integrating big data analytic for offshore support vessels based on a high performance computing platform. In addition, Kim et al. (2013) proposed a method to evaluate the influence of external forces using information obtained during operation.

As mentioned above, some researches have been carried out to apply big data to the shipbuilding and offshore industry, but there are few cases that used big data technology, and applied it practically. Therefore, this study proposes a platform to effectively utilize big data, and evaluate its applicability and utility in the shipbuilding and offshore industry, through application to an example. Table 1 compares the related studies mentioned above to this study.
Table 1. Summary of related studies and comparison with this study

| Studies            | Application field     | Application target               | Big data technology |
|--------------------|-----------------------|----------------------------------|---------------------|
| Kim et al. (2014)  | Computer Science      | Big data as cloud service        | N/A                 |
| Kim et al. (2014)  | Computer Science      | Big data analysis target         | N/A                 |
| Kim et al. (2013)  | Ship design           | Ship design Automation           | Hadoop              |
| Lee (2014)         | Ship production       | Ship production support          | N/A                 |
| Kim (2014)         | Ship operation        | Optimal operation and hull design| N/A                 |
| Perera (2017)      | Ship operation        | Ship performance and navigation monitoring | N/A |
| Wang et al. (2015) | Ship operation        | Offshore support vessel operation| N/A                 |
| Kim et al. (2013)  | Ship operation        | External forces on ship          | N/A                 |
| **This study**     | Offshore structure design | FPSO topside weight estimation    | Hadoop              |

2. Big data platform for the shipbuilding and offshore industry

This chapter describes Hadoop and Hortonworks Data Platform (HDP), which is a big data platform based on Hadoop for applying big data in the shipbuilding and offshore industry.

2.1. Hadoop

Hadoop is a representative big data framework that uses a simple programming model to efficiently distribute large amounts of data to many computer clusters. Hadoop is designed to scale from a single server to thousands of other devices that can be connected to local storage, and can be configured to detect and tune the failure of each device, to create a robust big data operation environment.

The main components of Hadoop are the Hadoop Distributed File System (HDFS), and MapReduce. HDFS enables fast and efficient data distribution storage and access, while MapReduce enables data distribution processing.

In addition, Hadoop offers a variety of sub-projects to make it more efficient for businesses. As these sub-projects became commercially available, Hadoop ecosystem was constructed, and Fig. 1 shows its configuration. In the Hadoop ecosystem, HDFS and MapReduce correspond to the Hadoop core project, and all other projects comprise the Hadoop sub-project.

2.2. The Hortonworks Data Platform (HDP)

Since Hadoop requires a complicated configuration for its operation, it is quite difficult for non-experts to use it directly. Therefore, some vendors provide platforms that implement Hadoop’s most important components, HDFS and MapReduce. Examples of these platforms include HDInsight from Microsoft, Cloudera platform from Cloudera, and HDP from Hortonworks.

Among these, HDP is free to use as a public program. And it can be based on the Windows OS (Operating System), which has the largest number of users. HDP is a platform that can connect storages based on Hadoop, and apply commercial technology that makes it possible to apply it to various fields. Also, it is very easy to connect with other programs, and apply new technology. Therefore, in this study, the application of HDP to the shipbuilding and offshore industry is examined.

![Fig. 1. The Hadoop ecosystem](image)
2.3. R and RHadoop

With the support of MapReduce, Hadoop can process big data quickly, but it is difficult to handle big data elaborately for a specific purpose.

On the other hand, R includes a scripting language for statistical computation and graphics, and its development environment. R provides various statistical techniques, and the installation of additional packages makes additional functions easily available. R has advanced analytical capabilities such as statistical processing, but has difficulty in handling big data.

Therefore, RHadoop package has emerged, which can offset the weaknesses of Hadoop and R programs, and takes advantage of them, making it possible to use R based on Hadoop. To use RHadoop package, a map and reduce function needs to be created to access the data stored in the HDFS, and to use MapReduce.

2.4. Big Data Platform Based on Hadoop for the Shipbuilding and Offshore Industry

In the shipbuilding and offshore industry, wide and various data are being produced through the procedures of design, production, and operation. Therefore, there is a constant need for industrial companies to effectively manage and utilize such big data. At present, there is no standardized and integrated system for integrating, storing, managing, and properly handling these big data (Kim et al., 2012).

Therefore, this study proposes a big data platform for application in the shipbuilding and offshore industry. Fig. 2 shows its configuration. In order to implement the big data platform, big data generated in the shipbuilding and offshore industry should be stored in servers and distributed storage, and distributed processing based on Hadoop should be possible. Depending on how the data is processed in the shipyard, it should be possible to add Hadoop ecosystem technology. Also, the ability to store or use processed data is needed. This can be achieved by installing HDP in the database, and interconnecting the data.

3. Offshore Structure Topside Weight Estimation Using the Big Data Platform

This chapter examines the applicability of the Hadoop-based big data platform to the shipbuilding and offshore industry. Items that can utilize big data in the shipbuilding and offshore industry include new ship design and development, process optimization, model test substitution, and weight estimation.

In this study, we applied the big data platform to the weight estimation of offshore structure topside, which has recently been attracting much attention.
3.1. Concept of Offshore Structure Topside Weight Estimation

Offshore structures, such as the Floating, Production, Storage, and Offloading Unit (FPSO), are composed of topsides and hulls. In particular, it is very important to accurately estimate topside weight at the initial design stage, as the most important and expensive equipment for crude oil and gas production is located on topside. Most of the weight estimation of offshore structure is based on past records (Um et al., 2014; Ha et al., 2016; Ha et al., 2017). Meanwhile, although large shipyards have carried out many offshore projects, taking into consideration the price and duration of construction, the number of projects is not very significant. In other words, the past records are insufficient to be treated as big data. However, if the construction results continue to accumulate in the future, it is expected that utilization as big data will further increase.

In this chapter, we performed the weight estimation of FPSO topside using the Hadoop-based big data platform introduced in Chap. 2, and used the HDP and RHadoop package for practical implementation.

The weight of the FPSO can be divided into the upper weight and the lower structure, including the accommodation area. In some cases, the lower turret may be separated. We focus on estimating the topside weight of FPSO, and Table 2 shows the parameters that affect it, where SC is the storage capacity, OP is the oil production, GP is the gas production, WP is the water processing, and WD is the water depth.

In this study, statistical correlation analysis and regression analysis were performed using 37 existing FPSO data, and as a result, weight estimation formulas of FPSO topside were derived. We apply linear and nonlinear methods to regression analysis, and use R language for this process.

3.2. Weight Estimation Method Using Correlation Analysis and Linear Regression Analysis

Correlation analysis is a statistical technique that analyzes how close the linear relationship between two variables is. In this method, the correlation coefficient represents the associated intensity of the two variables. In other words, the larger the correlation coefficient, the more closely the two variables are related. Correlation coefficients can be obtained by the Pearson method, which is a parametric correlation coefficient, and the Spearman method, which is a nonparametric correlation coefficient, depending on whether or not variables follow certain distributions (normal distribution, t-distribution, etc.). The Pearson method was used in this study, assuming a correlation coefficient of 0.5 or more, and a significance level of 0.15 or less.

Table 3 shows that the independent variables of B, D, T, DWT, SC, OP, GP, CREW, and WD satisfy the criteria. Therefore, these independent variables were selected as candidates that can be included in the weight estimation formula.
Table 4. Result of the FPSO topside weight estimation using linear regression analysis

| FPSO     | Actual weight (A) | Estimated weight (B) | Difference ton (A-B) | ((A-B)/A) |
|----------|-------------------|----------------------|----------------------|-----------|
| Skarv    | 16,000            | 23,800               | -7,800               | -0.48     |
| OSX 1    | 12,000            | 10,700               | 1,300                | 0.11      |
| Glas Dowr| 4,500             | 7,110                | -2,610               | -0.58     |
| Mean     |                   |                      | 3,903                | 0.39      |

Table 5. Result of the FPSO topside weight estimation using nonlinear regression analysis

| FPSO     | Actual weight (A) | Estimated weight (B) | Difference ton (A-B) | ((A-B)/A) |
|----------|-------------------|----------------------|----------------------|-----------|
| Skarv    | 16,000            | 20,300               | -4,300               | -0.27     |
| OSX 1    | 12,000            | 8,090                | 3,910                | 0.33      |
| Glas Dowr| 4,500             | 5,030                | -530                 | -0.12     |
| Mean     |                   |                      | 2,913                | 0.24      |

Assuming that the FPSO topside weight estimation formula can be represented by a linear combination of the parameters derived from the correlation analysis, a simple linear regression equation can be generated, as in Eq. (1):

\[
LWT_T = x_0 + x_1 \cdot B + x_2 \cdot D + x_3 \cdot T + x_4 \cdot DWT + x_5 \cdot SC \\
+ x_6 \cdot OP + x_7 \cdot GP + x_8 \cdot CREW + x_9 \cdot WD
\]  (1)

Then, Eq. (2) can be obtained by performing linear regression analysis on the big data platform using the existing FPSO data and Eq. (1). The results are consistent with the results of regression analysis using existing programs.

\[
LWT_T = -23,136.3 + 42.8 \cdot B + 1805.6 \cdot D - 1441.5 \cdot T - 0.0283 \cdot DWT \\
- 5920.1 \cdot SC - 3242.2 \cdot OP + 9.7 \cdot GP + 85.2 \cdot CREW + 6.0 \cdot WD
\]  (2)

3.3. Weight Estimation Method Using Correlation Analysis and Nonlinear Regression Analysis

Assuming that the FPSO topside weight estimation formula is nonlinear, better results can be obtained. Ha et al., (2016, 2017) performed a nonlinear regression analysis for the variables in Table 2, and as a result, the
Table 6. Computer specification

| Item            | Specification               |
|-----------------|-----------------------------|
| Operation system| Linux (Redhat 64-bit)       |
| CPU             | Intel® Core™ i5-6400 @ 2.70 GHz |
| Memory (RAM)    | 11 GB                       |

FPSo topside weight estimation formula can be expressed as a nonlinear \( f \) for \( D^3 \), \( SC^3 \), GP, and CREW\(^3\) as Eq. (3):

\[
LWT_T = x_0 + x_1 \cdot D^3 + x_2 \cdot SC^3 + x_3 \cdot GP + x_4 \cdot CREW^3 \quad (3)
\]

Then, Eq. (4) can be obtained by performing nonlinear regression analysis using the existing FPSO data and Eq. (3). The results are consistent with the results of regression analysis using existing programs.

\[
LWT_T = -558.2 + 0.3409 \cdot D^3 + 858.9 \cdot SC^3 + 14.54 \cdot GP + 0.001024 \cdot CREW^3 \quad (4)
\]

### 3.4 Result of FPSO Topside Weight Estimation Using Big Data Platform

This section verifies the accuracy of the FPSO weight estimation formulas obtained through correlation analysis and linear/nonlinear regression analysis by comparing the FPSO topside weight with the results of weight estimation.

Tables 4 and 5 and Figs. 3 and 4 show that the weight estimated from the linear and nonlinear regression equations has many errors compared with the actual weight, and the nonlinear case is slightly better than the linear one. This large error is also related to the number of data used in regression analysis. Although we used 37 existing FPSO data to perform this regression analysis, it is insufficient for regression analysis, and not enough for big data.

In order to increase the effectiveness of the big data platform proposed in this study, it is also important to secure big data for its application. In fact, offshore structures like FPSO are expensive and long-term delivery structures, so it will take some time to build a significant number of data. However, if large amounts of data are accumulated in the future, their utility as big data will increase.

### 3.5 Processing speed of FPSO topside weight estimation by data size

In this section, we examine the processing speed of FPSO topside weight estimation to check the effectiveness of the big data platform. Linear regression analysis on the big data platform is carried out by data size from 100 MB to 7.2 GB, and the processing time is measured. Table 6 shows the computer specifications for this.

![Fig. 5. Processing time of FPSO topside weight estimation](image-url)
Table 7. Processing time of FPSO topside weight estimation

| Data Size | R built-in function (A) | biglm package (B) | RHadoop package (C) | Differences (A-C) |
|-----------|--------------------------|-------------------|---------------------|-------------------|
| 100 MB    | 14 s                     | 12 s              | 92 s                | -78 s             |
| 200 MB    | 30 s                     | 26 s              | 107 s               | -81 s             |
| 400 MB    | 73 s                     | 53 s              | 112 s               | -71 s             |
| 800 MB    | 124 s                    | 122 s             | 165 s               | -43 s             |
| 1.6 GB    | -                        | 238 s             | 269 s               | -31 s             |
| 2.4 GB    | -                        | 350 s             | 377 s               | -27 s             |
| 3.2 GB    | -                        | 488 s             | 479 s               | 9 s               |
| 4.8 GB    | -                        | 763 s             | 691 s               | 72 s              |
| 6.4 GB    | -                        | 987 s             | 875 s               | 112 s             |
| 7.2 GB    | -                        | 1,111 s           | 958 s               | 153 s             |

To create various sizes of data from 100 MB to 7.2 GB, the 37 FPSO data were copied and pasted. Then, using these data, FPSO topside weight estimation was carried out with the R built-in function, biglm package, and RHadoop package to compare the processing speed.

Table 7 and Fig. 5 show that the R built-in function was not available from 1.6 GB due to insufficient memory. The biglm package and RHadoop comparisons show that the RHadoop package needs more time to estimate FPSO topside weight in data size smaller than 3 GB, but needs smaller time in data size larger than 3 GB.

In other words, in the case of low capacity data, the method without big data technology is faster; and in the case of large data, when applying big data technology, the data processing speed is fast. Therefore, it is clear that the bigger the data size, the greater the effect of applying the big data technology.

4. Conclusions and Future Studies

Recently, interest in big data has increased in the shipbuilding and offshore industry, and much research has been done to apply it. However, effective research for actual application is very rare. Therefore, this study suggests the big data platform for practical use of big data in the shipbuilding and offshore industry, and verifies its applicability by applying it to FPSO topside weight estimation. The effectiveness of large data processing by RHadoop is evaluated by comparing the results of regression analysis and the processing time with the results of the conventional weight estimation program.

The big data platform proposed in this study is to improve the existing large-capacity data for Hadoop-based distributed storage (HDFS), to process data for the application purpose (MapReduce), and to store the processed data. We use RHadoop on HDP to implement the proposed platform practically, and verify that HDFS and MapReduce can be used effectively.

In the future, we plan to apply the platform proposed in this study by deriving more diverse and vast data application targets in the shipbuilding and offshore industry.

Acknowledgements

This work was partially supported by (a) the MSIP (Ministry of Science, ICT and Future Planning), Korea, under “Development of Ship Design Standard PLM Platform based on Big Data” (NIPA-2016-S1106-16-1025) supervised by the NIPA (National IT Industry Promotion Agency), (b) the BK21 Plus Program (Education and Research Center for Creative Offshore Plant Engineers), funded by the Ministry of Education, Korea, (c) the Research Institute of Marine Systems Engineering, of Seoul National University, Korea, and (d) the Program of University Specialized for Offshore Plant Engineering, funded by the Ministry of Trade, Industry and Energy, Republic of Korea.
References

Bae, D.M., Park, H. S. and Oh, K. H., 2013, Big Data Trend and Policy Implication, Information and Communication Policy, 25(10) (2013) pp.37-74.

Ha, S., Seo, S.H., Roh, M.I. and Shin, H.K., Simplified Nonlinear Model for the Weight Estimation of FPSO Plant Topside Using the Statistical Method, Ships and Offshore Structures, 11(6) (2016) pp.603-619.

Ha, S., Um, T.S., Roh, M.I. and Shin, H.K., A Structural Weight Estimation Model of FPSO Topsides using an Improved Genetic Programming Method, Ships and Offshore Structure, 12(1) (2017) pp.43-55.

Lee, D.H., Analysis of Production Process in Shipbuilding Industry using Process Mining, Ph.D. thesis, Pusan National University (2014), Korea.

Lee, H.H., Application of Big Data for Strengthen of Manufacturing Business, Seoul, Korea: Korea Institute for Industrial Economics and Trade (KIIET), (2013).

Kim, K.I., Jung, J.S. and Park, K.K., Assessment of External Force Acting on Ship using Big Data in Maritime Traffic, Journal of Korea Intelligent Information System Society, 23(5) (2013) pp.379-384.

Kim, S.R. and Kang, M.M., The Trends and Prospects in Cloud-Based Bigdata Technology, Journal of Korean Institute of Information Scientists and Engineers, 32(2) (2014) pp.22-31.

Kim, S.R. and Kang, M.M., Today and Tomorrow of Big Data Analysis Technology, Journal of Institute of Information Scientists and Engineers, 32(1) (2014) pp.8-17.

Kim, U.G., Ship & Offshore industry and Big Data collected during Operation of Ships as ICT Convergence Model, Journal of Mechanical Science and Technology, 54(12) (2014) pp.49-52.

Kim, Y.J., Park, J.K., Lee, J.H., Yang, H.Y. and Jung, M.A., A Study on the Bigdata Technology and Analysis Technique for Vessel Design Automation, Journal of Korea Institute of Communication Sciences, 2013(6) (2013) pp.213-215.

Kim, W.K., Park, M.K. and Han, M.K., Design of a Framework for Support System of Ship Design Engineering, Journal of Korea Institute of Information and Communication Engineering, 16(10) (2012) pp.2316-2322.

Perera, L.P., Handling Big Data in Ship Performance & Navigation Monitoring, Smart Ship Technology, London, UK, January 24-25, (2017)

Um, T.S., Roh, M.I., Shin, H.K. and Ha, S., Simplified Model for the Weight Estimation of Floating Offshore Structure Using the Genetic Programming Method, Transactions of the Society of CAD/CAM Engineers, 19(1) (2014) pp.1-11.

Wang, H., Osen, O.L., Li, G., Li, W., Dai, HN. and Zeng, W., Big Data and Industrial Internet of Things for the Maritime Industry in Northwestern Norway, IEEE Region 10 Conference, Macao, China, November 1-4, (2015), doi: 10.1109/TENCON.2015.7372918