A Cloud-based System Framework for Storage and Analysis on Big Data of Massive BIMs

Hung-Ming Chen [a], and Kai-Chuan Chang [a]

[a] Department of Civil and Construction Engineering, National Taiwan University of Science and Technology, Taiwan (R.O.C.)
E-mail: hungming@mail.ntust.edu.tw

ABSTRACT

In this study, a system which utilizes Cloud technology to establish a data center to store and manage multiple BIMs simultaneously has been developed. This BIM data center can not only handle the big data of massive BIMs using multiple servers in a distributed manner, but can also be accessed by using various online devices anywhere, anytime for information sharing and visualization. Traditional BIM includes only static information, such as the geometric parameters, physical properties, and spatial relations, for modeling of the physical space. In this study, BIM was extended to dynamic BIM which also includes dynamic data, such as historical records from the monitoring of the facility environment and usage, as well as the user experience parameters from continuous observation of the interaction between users and the space. Due to such an extension, dynamic BIM became a parametric model which can be used to simulate user behaviors. Regarding the applications of dynamic BIM big data, this study proposes a Cloud-based system framework to effectively retrieve required information for various applications by big data analysis based on parallel processing of large data sets.

Keywords -
BIM; Cloud Computing; Big Data Analysis; MapReduce

1 Introduction

Building Information Model/Modeling (BIM) is the addition of storage and integrated management on the three-dimensional visualization models and digitized design information of a building project, such that in addition to the geometric data, interdisciplinary design information becomes the property of a three-dimensional visualization model. Based on this model, effective integration of design data from different disciplines can be achieved and passed on for the subsequent task use. Currently, BIM technology is developing rapidly. There are also many commercial BIM software products on the market. However, since all commercial BIM software products are based on the project management mode of specific file formats, all their files must be opened through a particular BIM software. Only then can a user view the three-dimensional visualization model and design property data of various disciplines. Although a common BIM file format such as the Industry Foundation Classes (IFC) has been proposed, the logic and definitions of BIM among commercial software vary endlessly. Therefore, it is very difficult to maintain consistency on the exported information format and content for the IFC exported by different commercial BIM software. This may even lead to loss of information. Even with the same commercial BIM software, different software versions are also likely to experience the problems previously mentioned. Moreover, today's commercial BIM software all operate using file-based management mode, meaning that a file corresponds to a BIM project. If one wants to carry out classification on a BIM collection set, search for the association rules between different BIM within the collection, or perform other analysis, it is possible that one can only execute the analysis when there is only a small number of BIM within the collection. If there is a large number of BIM within the collection, there is certainly a considerable degree of difficulty.

To solve the needs of statistical analysis on a large number of BIM, this study breaks through the restrictions of the file-based management mode of commercial BIM software by means of cloud computing technology. The so-called cloud refers to a network. As long as a network is used to communicate the computing tasks of multiple computers or to obtain services provided by a remote host via a network connection, it can be described as cloud computing. Moreover, cloud computing can be subdivided into two categories, namely cloud computing services and cloud computing technologies. Cloud computing services focus on obtaining services from a remote network

connection. This type can be regarded as the successor of the Software as a Service (SaaS) concept. Cloud computing technologies focus on using virtualization and automation technologies to create and popularize a variety of computing resources in a computer. This type can be regarded as an extension of a traditional data center. Compared to the commonly-used personal information systems, cloud computing can be considered to have powerful computing capabilities, which cannot be matched by a personal computer. Therefore, cloud computing can handle a large amount of data and provide various types of software services.

Through the use of cloud computing technology, this study hopes to conduct statistics and analysis on a massive amount of BIM data. In recent years, big data analysis has become one of the hot topics related to cloud computing. Big data means that the file capacity data reaches above the petabyte (PB) level. Regardless of whether this data above the PB level is composed of a file or multiple files, both can be classified as big data. Big data analysis means the use of this data above the PB level generated from long-term accumulation to conduct in-depth analyses. The main aim of these analyses is to use the large amount of experience data accumulated over the past to dig out useful hidden values.

Our team has proposed and developed a cloud-based BIM system called CloudBIM [1] that can perform storage and viewing on the data of massive BIMs. As shown in Figure. 1, it uses cloud computing technology to store massive BIM data and adopts IFC as the BIM file upload format of the CloudBIM system. Based on the IFC format, we developed a commonly-used BIM upload interface and then developed the Web interface for BIM viewing using WebGL, such that the BIM can be reached on any device through a standard web browser online viewing system. Though this system solves the problems caused by the project management mode of existing commercial BIM software based on specific file formats and the compatibility of files between manufacturers, as its main function is only to use cloud computing technology on data storage and the three-dimensional visualization viewing level of massive BIMs, the storage of the data of massive BIMs in the CloudBIM system still has many possibilities for use in analysis. Some such examples are the use of cloud computing technology to conduct statistics and analysis on the property data of massive BIMs, or the addition of the dynamic data of BIM inside the buildings' space for joint operations, and other possibilities that are yet to be realized. If these functions can be added into CloudBIM, its functionality will be more complete.

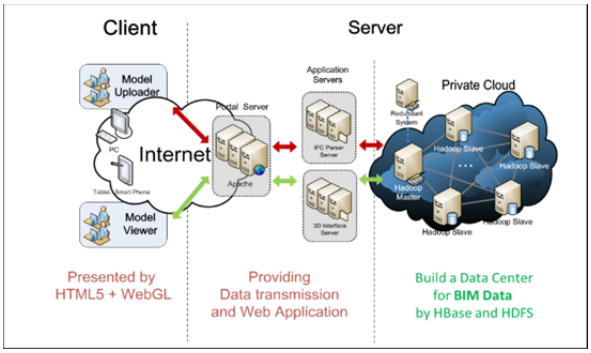


Figure 1. System framework of the CloudBIM.

In this study, we took considerations to improve the value-added applications of BIM. Apart from the three-dimensional visualization model and the design data of each discipline originally covered by BIM, dynamic data that changes over time, including historical records of the environmental state and the usage of facilities inside the building space, can also be added into the BIM to perform further simulation and analysis. So, by integrating a constructed BIM that has geometric physical property parameters and spatial relations for modeling entity space with the historical records of the environmental state and the usage of facilities by monitoring inside the space, the parameter model can be used to simulate the interactions of behaviors within the building space in order to achieve value-added applications of BIM. The dynamic data composing such value-added applications is not only a large number, but will also constantly accumulate and expand with continuous monitoring on the space. These data must not only be captured and applied effectively, but must also be able to be specifically presented for understanding and sharing.

2 Background

Today's computer-aided design trend is moving towards the creation of a design information integration mode of BIM for project design. Furthermore, the construction industry has gradually adopted BIM technology in the design and development process in larger and more complex construction projects. Examples of BIM-related applied research in recent years include Chen and Luo [2], Motamedi et al. [3] and Sanguinetti et al. [4]. These studies integrate BIM with data and method for applications.

The challenge of big data is the problem of properly dealing with the management of three dimensions: Volume, velocity (real time and data batch increase), and variety (structural and non-structural). With greater increase in the data quantity, as well as more variety in data formats, the demand of enterprises being eager to obtain real-time analysis is also clearer. They are

depending on real-time intelligence to obtain advantages in the competition and increase profitability.

Regardless of being enterprises or academics, both can conduct analysis on the big data required in its discipline in order to obtain valuable results to assist in decision-making. For instance, real-time network traffic monitoring and analysis are dynamic and huge. Singh et al. [5] combined Hadoop, Hive, and Mahout, open source big data processing tools, to create a scalable quasi-real-time intrusion detection system. They used Hive to create a distributed framework for dynamic network tracking, as well as made use of the parallel processing capabilities of Mahout to create random forests, using the machine learning approach to detect peer to peer botnet attacks. Cohen et al. [6] stressed the importance of Magnetic, Agile, Deep, (MAD) analysis on big data analysis and proposed their own parallel database systems and parallel algorithms, thereby allowing users to use SQL and MapReduce flexibly over a variety of data storage mechanisms.

By applying cloud computing and WebGL technology, this study is based on the proposed network type BIM system, named CloudBIM, to perform extensions and expansion. Apart from using Apache HBase to integrate the original BIM data and newly-added dynamic data that changes over time, such as historical records obtained by simulating and monitoring the environmental state and facilities' usage in the building space, the cloud computing service of massive BIMs' data analysis is newly-added into its webpage platform which originally provided cloud computing services. This study applied cloud computing and big data analysis technology on the storage and management of massive BIMs and related dynamic data. This is an unprecedented attempt.

3 Data analysis of massive BIMs by integrating basic BIM data and dynamic data

In this study, a big data processing framework based on Storage, MapReduce and Query, (SMAQ) as shown in Figure 2 was adopted. The cloud-based big database of CloudBIM was set as Storage; the big data analysis utilized MapReduce [7] as the main computing framework, whereas Web UI was the platform providing users with a method to conduct analysis and carry out Query on the results of the analysis. Moreover, when performing big data analysis, the process shown in Figure 3 was followed, conducting further extension and application on the massive BIMs stored in the CloudBIM system.

MapReduce is a software framework that allows developers to conduct parallel computing on big structured or unstructured data in a distributed environment. The main characteristics of MapReduce are highly reliable computing, fault tolerant mechanisms, local computation, and automatic load balancing. When local computation allows developers to execute a program, the program will carry out computing above the node storing the data to be processed, and the program code to be executed is transmitted over the network; as for data that is not to be processed, the program code is often much smaller than the file size of the data to be processed. This method can reduce network transmission bandwidth requirements and was also applied equally to the big computation.

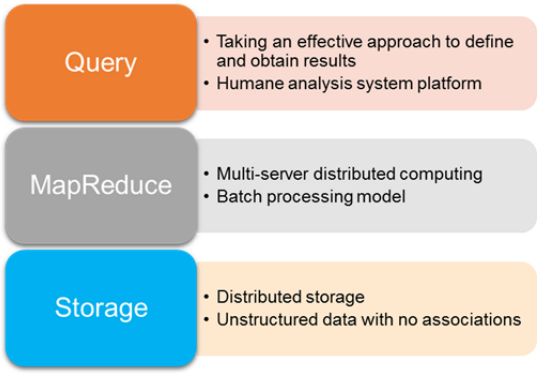


Figure 2. Big data processing framework.

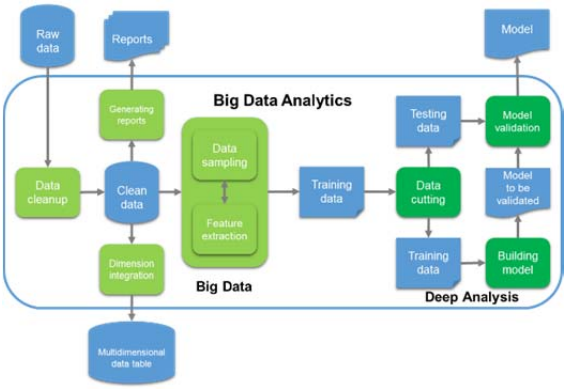


Figure 3. Big data analysis process.

To provide more analysis value for the data in massive BIMs, this study tried to take the simulation and monitoring approach on the BIM stored in the CloudBIM system to simulate dynamic data that changes over time, such as historical records of the environmental state and the use of facilities inside its building space, and then equally stored them in the CloudBIM system. The uploaded dynamic data will be distributed and stored in the cloud-based Big Table. Then, in accordance with MapReduce distributed computing as the main data analysis mode for massive BIMs to expand the CloudBIM system, the BIM data stored in the CloudBIM system will hence including the basic BIM data and the dynamic BIM data. Furthermore, the CloudBIM system will become a cloud system platform that simultaneously possesses the data storage

management of BIM as well as provides data analysis services for massive BIMs. This allows users to select related task support services through mobile phones, tablets, laptops, desktops, or any device that can access the internet, in order to enhance the interaction and integrity of the CloudBIM system.

This study aims to fulfill the purposes mentioned above by using the cloud-based massive BIMs data storage and analysis system framework to develop the system function requirements, and then to propose and develop related plans for the system of this study. We will not only take the simulation and monitoring approach to create dynamic BIM data to expand the CloudBIM system, but will also propose four types of data processing modes based on the MapReduce distributed computing framework for the data of massive BIMs stored in the CloudBIM system, as shown in Figure 4.

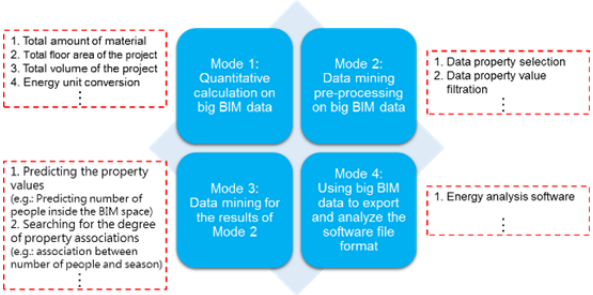


Figure 4. The four parallel data processing modes proposed in this study for massive BIMs.

4 System framework and operation mechanisms

In this study, the overall system framework is as shown in Figure 5. It is based on the big data processing framework of SMAQ. It is divided into three layers. These are the cloud service platform framework on the client side corresponding to the Query layer of SMAQ, the big data analysis framework corresponding to the MapReduce layer, and the Hadoop cluster corresponding to the Storage layer.

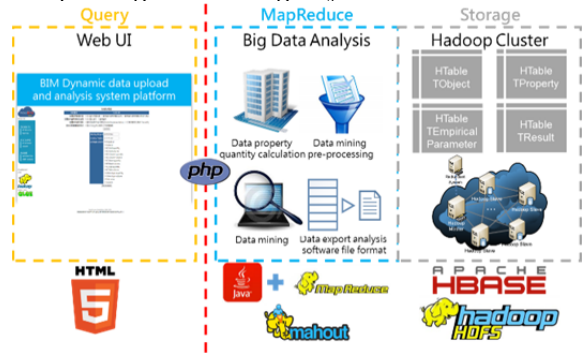


Figure 5. System framework for big data analysis on massive BIMs

4.1 Cloud service platform framework on the client side

The webpage framework of the cloud service platform on the client side for this system is as shown in Figure 6. The pre-existing groups of CloudBIM are the five groups of Menu, Project, Display, Account, and About. The sixth Analysis group is newly added by this study. The Menu group is responsible for providing the webpage index. The Project group allows users to upload BIM and view uploaded BIM lists. This study added the function of dynamic data uploading to this group. The Display group is responsible for presenting three-dimensional visualization models of BIM projects. The Analysis group is a big data analysis platform, allowing users to run four types of distributed computing analysis services on the uploaded BIM object property data and conditional dynamic data. The Account group provides account review and logout capabilities. The About group contains website introductions and tutorials. Through the newly-added Analysis group, the user only needs to set up the parameters required by the back-end computing module in the analysis platform of the group. The user then can perform cross-platform data analysis of massive BIMs and can even view the analysis results on this platform in order to correspond with the aims of simplifying the MapReduce operation of Query layers in SMAQ and view the computing results.

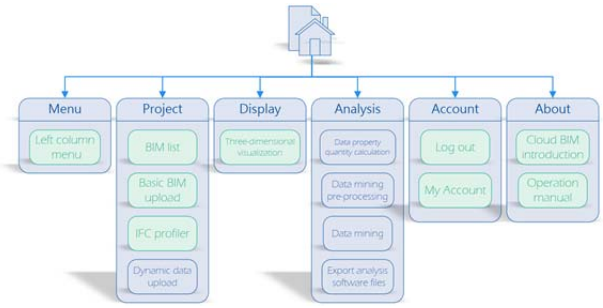


Figure 6. The Web UI framework of CloudBIM.

4.2 Big data analysis framework on the server side

The big data analysis framework of this study consisted of four analysis mode types based on the MapReduce computing frameworks. The basic operation of each mode is described as follows:

1. Quantity calculation on massive BIM data: The user can specify and choose a set of BIM through a webpage analysis platform, selecting the property data of BIM for quantity calculation of geometric properties or the dynamic data of BIM conditions for unit conversion of the energy consumption property.

The former includes computing values such as the total amount of material, the total volume of material, the total BIM floor area, and the total BIM volume. The latter can be divided into the two kinds of energy consumption pre-set units contained in the dynamic data by converting GJ into kWh or million Btu. For the two types of quantity calculations above, the user can specify the main properties of the output in descending order. The system will perform computations based on the user's request in the webpage analysis platform back-end. The final computing results will be in the comma-separated values (CSV) file format for users to download, and users can choose to browse a plain text file or in the form of a report.

2. Data mining pre-processing on massive BIMs: Using the webpage analysis platform, users can choose and specify a set of BIM as well as intentionally select the property of BIM objects or the dynamic data property of the BIM condition. After the selection is completed and sent out, the platform will perform computations in accordance with the parameters selected in the back-end. Finally, the pre-processing results will be stored in the HDFS. The platform not only allows users to browse the pre-processing results, but it will also inform users that data mining pre-processing has been completed and proceed to the mining task in the next step.

3. Data mining on the results of Mode 2: When Mode 2 is successfully completed, this mode will utilize data mining tools on the data of massive BIMs for data mining. In this mode, the desirable method of data mining is first selected by the user. The analysis platform will provide two types of data mining methods: One is making use of many decision trees to create random forests for predicting property values of massive BIMs; Another is calculating the frequent item collection of the property values of massive BIMs in order to allow users to use this frequent item collection to determine the association rules between the properties. The data mining results will be presented in the analysis platform for browsing. Finally, through the data mining result access module, this study will store results into a big table in order to facilitate the subsequent query task, and to enable users to download the data mining results through the same module.

4. Exporting the data of massive BIMs in analysis software file formats: The platform will enable users to specify and select a set of BIM. In accordance with the specifications, it will obtain the property data of BIM components and dynamic data of BIM in order to create the file format required by the analysis software. After the platform computation is completed, the exported analysis software file will be provided in the webpage platform, allowing users to download it for analysis.

As each data analysis module above is based on the operation mechanism of MapReduce, the distributed processing logics are roughly the same. This paper takes the operation mechanism for the property quantity calculation on massive BIMs as an example for description. The operation mechanism is as shown in Figure 7. It takes the Property Table as an input, and uses MapReduce to perform the quantity calculation on the geometric property. The BIM components property covered by the Property Table has length, width, height, positions in three-dimensional space, rotation angle, scaling ratio, etc. The computing sequence is divided into the Map stage and the Reduce stage:

1. Map stage: For analyzing the set of BIM specified by the user. A Mapper only handles one BIM at a time, and there will be multiple Mappers running simultaneously. Each Mapper will first obtain the property values of length, width, height, material, and component types belonging to the local BIM positioned at the same node as the Mapper, and then conduct individual calculations on each type in order to obtain the total amount and the total volume of material of each type in the BIM. For floor types, apart from individual volumes, it still needs to calculate the area and accumulate them to obtain the total floor area of the BIM. Finally, summing up the total volume of each type enables us to obtain the total volume of the BIM. The end results of a BIM will be presented in a line. The presented content includes the total amount of material for walls, pillars, beams, and boards of the BIM, as well as the total volume of the individuals, in addition to the total floor area of the total volume of the BIM.
2. Reduce stage: Reducer integrates all the results of Mapper. It integrates every individual result, generated by analyzing each case of the BIM, into a file containing all the results of the BIM.

When Reducer computation is completed and sent out, it will be handed over to the sorting program in the module for sorting. The sorting program will be based on the result columns specified by the user to perform the sorting process in descending order. For example: A user selects the total volume of the BIM as the column for sorting. The end results will take the total volume of the BIM as a major column to be sorted out against all results in descending order. The output is in the CSV file format, and users can browse it in report form or in the form of plain text.

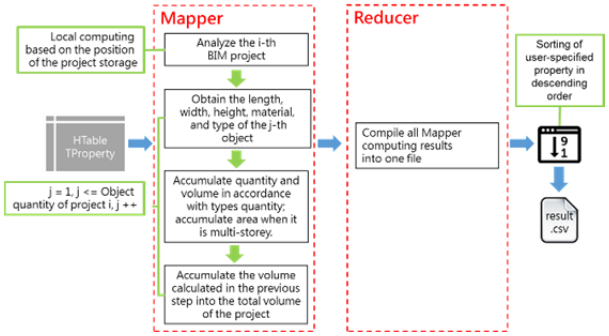


Figure 7. MapReduce-based operation mechanism for property quantity calculation on massive BIMs.

4.3 Big data storage framework on the server side

In order to increase the value-added effectiveness of BIM, this study conducted expansions on the big database HBase of the CloudBIM system. Developed based on HDFS, HBase is a NoSQL database system in BigTable form. Since HBase is a non-relational database, the storage and creation of data do not depend on the primary key, but according to the structured characteristics of the data. Since the form and concept of HBase are very different from a relational database, the important points focused on by outline design of its data are also different. In this study, we added two new data tables: one is a dynamic data table, TempiricalParameter, and the other is a data mining results table, TResult.

The TEmpiricalParameter table is as shown in Figure 8. In accordance with the characteristics of the HBase database, this study wrote the conditions of BIM dynamic data in a continuous manner to improve the efficacy of HBase in reading the data. Since data with higher associations will exhibit closer data distance, when reading one of the conditional data, all the conditional property data will be simultaneously read into memory. For the row key, BIM code is used as the index for data cutting, and so it can separate different BIM under different row keys, prompting all the content of the same BIM to be stored in the same row keys.

The TResult table is as shown in Figure 9. This study takes the code of one set of BIM for data mining as a row key for separating different data mining results of BIM sets. The column name is abbreviated using the data mining method. This storage method mainly uses the column dynamic expanding characteristics. When a new data mining method is incorporated in the future, it can be directly written into the table without separate definitions.

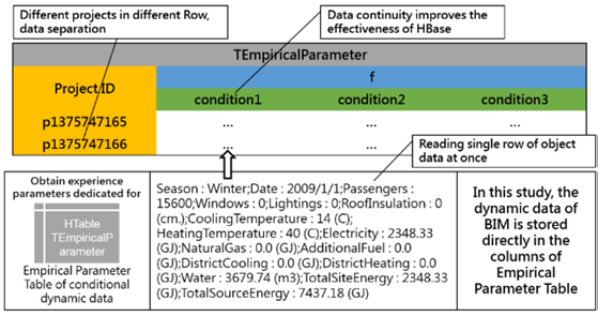


Figure 8. HBase storage framework for dynamic data of BIMs.

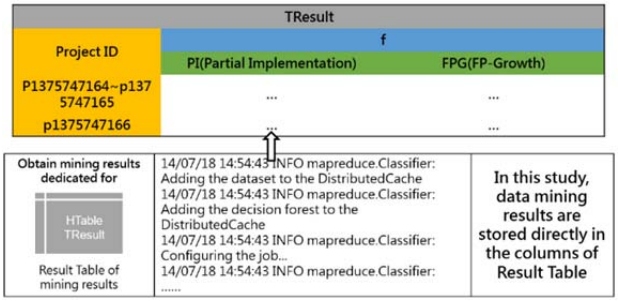


Figure 9. HBase storage framework for the results of data mining.

5 System performance test and validation

This system adopted the MapReduce computing framework on all data analyses of massive BIMs, and the computing works were all carried out on server-side cluster computers. Through this computing framework, not only can the execution time of data analysis on massive BIMs be significantly reduced, the hardware requirements of the user's device can also be reduced. The system pressure of using this system has no difference compared with browsing a general webpage.

In order to present the benefits from MapReduce computing framework on various types of data analyses on the massive BIMs, this test respectively conducted comparisons on the efficacy of a stand-alone version and a MapReduce version for the implemented BIM property quantity calculation, data mining pre-processing and export in an analysis software's file format, and the three functional modules. As the analysis target is big data, to ensure that the amount of data is capable of horizontal expansion, the data source of standalone and distributed computing were all the same as HBase, while the distributed cluster in this study was composed of 15 computing nodes with the same specifications. Table.1 shows hardware specifications of each computing node. The following will compare the difference in efficacy under conditions of each BIM in varying scale. That is, each BIM has different object quantities and number of conditions.

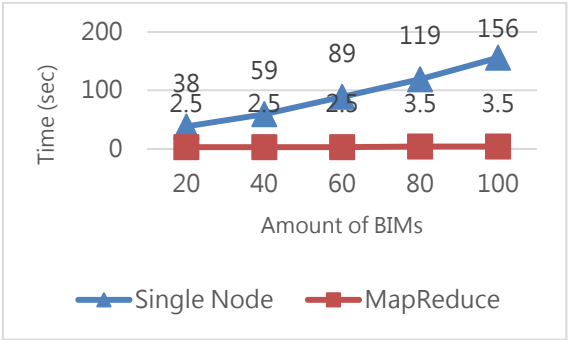
Table.1 Hardware specifications of each node

Hardware item	Specification
CPU	Intel i5-4570 Processor 3.2GHz
RAM	DDR3-1600 16GB
Storage	HDD 1TB 7200rpm
Network	Ethernet 1000 Mbps

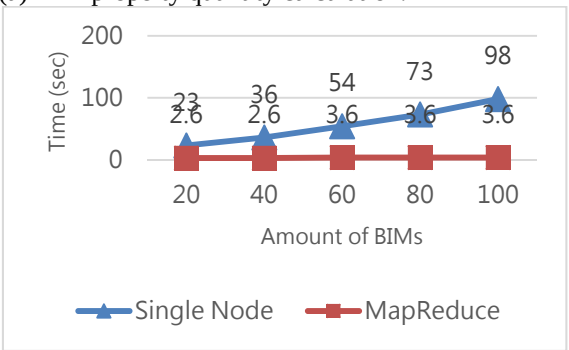
The computing time differences of the module in the MapReduce version and the stand-alone version when this system carried out property quantity calculation on massive BIMs are as shown in Figure 10(a) and Figure 10(b). Both are respectively the analysis time line graphs for the BIM object property data quantity calculation and the energy consumption conversion from dynamic data of BIM conditions on the data of massive BIMs. From the graphs, it can be seen that the analysis time of the stand-alone version is much longer than that of the MapReduce version. Furthermore, when the BIM quantity for analysis is greater, the efficacy of the MapReduce version grows more significantly.

When this system carried out data mining pre-processing tasks on the data of massive BIMs stored in the CloudBIM system, the computing time differences between the standalone version and the MapReduce version were as shown in Figure 12. The data source of Figure 11(a) is BIM object property data, whereas for Figure 11(b), it is the dynamic data of BIM conditions. From the two graphs, it can be seen that when the BIM quantity increased, the analysis time for the stand-alone version continued to increase, while the analysis time for the MapReduce version exhibited almost no difference and is far lower than that of the stand-alone version.

This test also compared the efficacy in exporting analysis software file formats on the stand-alone version and the MapReduce versions. MapReduce was also divided into two versions. Ver. 1 only carried out MapReduce data locality computing for the BIM object property table, whereas Ver. 2 conducted MapReduce data locality computing on both the BIM object property table and the dynamic data of BIM conditions. The analysis time of the three versions was as shown in Figure 12. It can be seen that the performance was significantly better when adopting the MapReduce version. Furthermore, from the graph, it can be seen that the performance of Ver. 2 was superior to Ver. 1. This is because Ver.1 only conducted data locality computing of MapReduce for the object property table. There was no application of MapReduce computations on the conditional dynamic data table. The overall performance of Ver. 2 was not only far better than that of the stand-alone version, it was also far better than that of MapReduce Ver. 1. Therefore, a complete version of MapReduce will obtain the best execution efficiency.

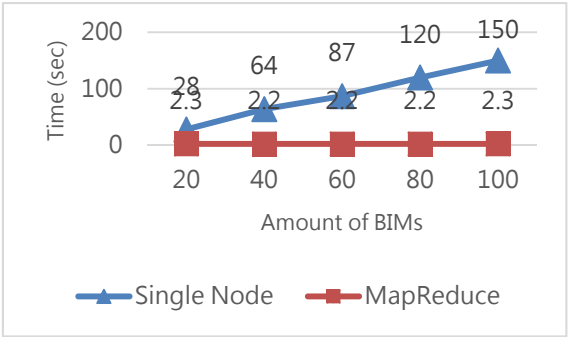


(a) BIM property quantity calculation.

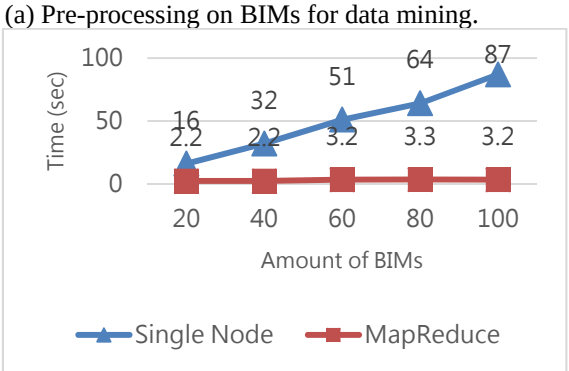


(b) Energy consumption conversion from dynamic data of BIMs.

Figure 10. Time comparison on BIM property quantity calculation.



(a) Pre-processing on BIMs for data mining.



(b) Pre-processing on dynamic data of BIMs for data mining.

Figure 11. Time comparison on pre-processing for data mining

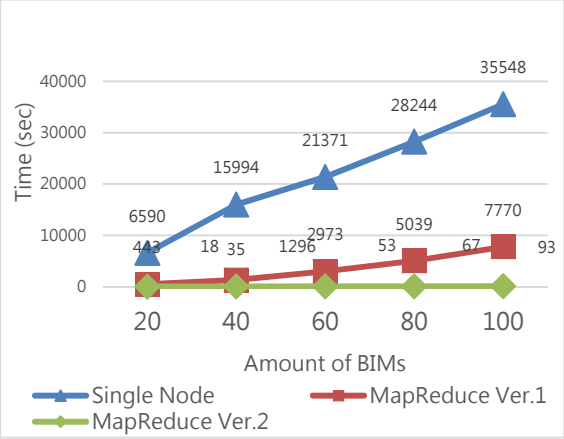


Figure 12. Time comparison on BIM conversion for exporting the input file of analysis software.

6 Conclusions

This study proposed the cloud-based network BIM real-time service framework by combining Apache Hadoop cloud computing technology, WebGL 3D display technology, and HTML5 webpage technology. The system is divided into server-side and client-side. The server-side is composed of HDFS, Hadoop MapReduce, and Hbase. HDFS provides large storage space based on a distributed approach; Hadoop MapReduce can parallel process large data sets; HBase makes use of several big tables to store the information from massive BIMs. For the client-side, it applies WebGL 3D display technology and HTML5 webpage technology. A user can use all types of network-accessible devices to connect to the internet to log into CloudBIM anytime, anywhere to upload or view BIM. Through the support of a HTML5 webpage browser and WebGL 3D display technology, 3D models of BIM data stored on the server-side can be presented.

In this study, BIM was extended to dynamic BIM, which also includes dynamic data such as the historical records from the monitoring of the facility environment and usage, as well as the user experience parameters from continuous observation of the interaction between users and the space. Due to such an extension, dynamic BIM has become a parametric model which can be used to simulate user behaviors. Regarding the applications of dynamic BIM big data, this study proposed four data processing mode types based on the MapReduce distributed computing framework for the data of massive BIMs stored in the CloudBIM system, such that a user can not only view the three-dimensional visualization model and object properties on the platform, but can also easily conduct diversified analyses on the data of massive BIMs stored therein. This study conducted actual validation tests on the prototype system and the results show that the proposed system framework and operation mechanism are able to provide big data storage space and big data analysis capabilities in distributed and parallel processing of the large amount of BIM data. Compared with the stand-alone version, the results prove that the MapReduce-based main computing frameworks adopted in this study can not only be closely integrated with the distributed storage framework of Hadoop's distributed storage architecture, but can also dramatically improve the efficiency of massive BIM data analysis via distributed data locality computing .

References

[1] Chen,H. M. Hou, C. C. and Lin T. H. A Cloud-Based Framework for Online Management of Massive BIMs Using Hadoop and WebGL. In *Proceedings of the 30th International Symposium on Automation and Robotics in Construction*, Montreal, Canada, 2013.

[2] Chen, L. and Luo, H. A BIM-based construction quality management model and its applications. *Automation in Construction*, 46:64-73, 2014.

[3] Motamedi, A., Hammad, A. and Asen, Y. Knowledge-assisted BIM-based visual analytics for failure root cause detection in facilities management. *Automation in Construction*, 43: 73-83, 2014.

[4] Sanguinetti, P., Abdelmohsen, S., Lee, J., Lee, J., Sheward, H. and Eastman, C. General system architecture for BIM: An integrated approach for design and analysis. *Advanced Engineering Informatics*, 26:317-333, 2012.

[5] Singh, K., Guntuku, S. C., Thakur, A. and Hota, C. Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests. *Information Sciences*, 278:488-497, 2014.

[6] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M. and Welton, C. MAD Skills: New Analysis Practices for Big Data. In *Proceedings of the VLDB Endowment*, 2(2):1481-1492, 2009.

[7] Dean, J. and Ghemawat, S. MapReduce : Simplified Data Processing on Large Clusters. In *Proceedings of the 6th Symposium on Operating Systems Design & Implementation*, pages137-150, San Francisco, California, USA, 2004.