*Article*

# Development of Cost and Schedule Data Integration Algorithm based on Big Data Technology

**Daegu Cho [1], Myungdo Lee [2] and Jihye Shin [3,*]**

[1] Department of Construction Big Data Research Center, Ninetynine Inc., South 01905, Korea; bignine99@naver.com
[2] Department of BIM Research Center, Yunwoo Technologies Inc., South 05854, Korea; md.lee@yunwoo.co.kr
[3] The Centre for Spatial Data Infrastructures and Land Administration, Department of Infrastructure Engineering, The University of Melbourne, Victoria 3010, Australia;
[*] Correspondence: jihyes@student.unimelb.edu.au; Tel.: +61 03 8344 0234

**Abstract:** In the information age today, data are getting more and more important. While other industries achieve tangible improvement by applying cutting edge information technology, the construction industry is still far from being enough. Cost, schedule, and performance control are three major functions in the project execution phase. Along with their individual importance, cost-schedule integration has been a significant challenge over the past five decades in the construction industry. Although a lot of efforts have been put into this development, there is no method used in construction practice. The purpose of this study is to propose a new method to integrate cost and schedule data using big data technology. The proposed algorithm is designed to provide data integrity and flexibility in the integration process, considerable time reduction on building and changing database, and practical use in a construction site. It is expected that the proposed method can transform the current way that field engineers regard information management as one of the troublesome tasks in a data-friendly way.

**Keywords:** Big Data, Data Integration, EVMS, Construction Management

## 1. Introduction

### 1.1. Background and Purpose

At present, it is a big data era. With the continuous development and progress of the information age, the demand for information is getting larger and more important. It has been perceived that the significance of data and its latent value in the construction industry will become greater. In this context, there has been a range of efforts to manage information systematically in construction projects by introducing various information management systems, such as continuous acquisition & life-cycle support (CALS), project management information system (PMIS), enterprise resource planning (ERP), building information modeling (BIM). Although these systems have provided a relatively low increase in the efficiency rates over recent decades, data management is still trying to catch up with the exponential data growth in the sector.

The amount and complexity of data being gathered are continuously growing; however, data itself is not valuable. Data in construction management has characteristics with a vast amount, various formats, complicated structure, and frequent changes as projects proceed. It leads to challenges in standardization and generalization of data. Consequently, data management in construction projects is still under a paper document-oriented process, which results in (1) loss of critical construction management data after project completion, (2) lack of available and meaningful data, and (3) absence of database for knowledge share in future projects. To respond to the limitations, it is essential to develop a data management methodology that can facilitate the processing and analysis of construction project data, reflecting its characteristics. This paper aims to propose a data processing algorithm for construction management, which extracts and analyzes the required data by various stakeholders from a wide range of data sources associated with construction projects.

### *1.2. Research Process*

Construction management is a series of systematic procedures to manage the design, schedule, cost, quality, and safety of a project. Cost and schedule management are regarded as project success factors; the data associated with them show direct connectivity and can be used as a basis for managing productivity, labor, periodic payment, and material of construction projects. In this context, the proposed algorithm in this research focuses on managing quantity, cost, schedule, and payment, which are intertwined with each other, from the database perspective.

To achieve the research aim, this research is structured as follows. First, the current status of data control in construction management is investigated; root causes for its unsystematic practice are identified in the context of difficulties of cost-schedule data integration. Second, existing methods for integrating cost and schedule data in construction management are reviewed, together with their limitations in application. Third, characteristics and challenges in cost-schedule data integration are identified, and the feasibility of big data technology is examined as a solution for the challenges. Forth, a multi-dimensional and -level data structure incorporating three requirements for cost-schedule data integration is proposed. Based on it, an algorithm for the integration and its supporting modules for integrating, extracting, analyzing, and visualizing are developed. Lastly, the implemented algorithm is applied to a case study to analyze its validity from three perspectives: data integrity and flexibility, data building and transformation time, and practicability.

## **2. Literature Review**

### *2.1. Current Status of Data Control for Construction Management*

Cost-schedule management is one of the most critical management items in construction projects. It can be used as a quantitative performance indicator that allows evaluating the success of projects [1]. The informatization process for managing construction projects, particularly associated with cost and schedule, is as follows. First of all, the cost report is prepared by applying unit cost for each construction work to the quantity take-off on building elements. Based on the quantity of construction part, schedule management is performed according to materials and work productivity. Measurement of the progress to plan over time allows operating construction projects through progress rate measurement and performance payment. It indicates that quantity, payment, schedule, productivity, progress, and resource management (i.e., labor, material, equipment) have strong correlations and need to be updated as projects proceed.

During the project life-cycle, the vast amount of data associated with construction management are produced by individual systems for different purposes, such as quantity take-off, schedule management, cost management. Despite the correlations among them, the created data from the systems shows a lack of interoperability. As a result, projects are mainly operated based on the paper document with different templates, diverse file formats, and various information levels. It causes difficulties in synchronizing and updating data with the lack of data connectivity, as well as redundant input of data, repeated data management processes, and discrepancy in data. Figure 1 represents the current practice with (1) various documents for construction management, (2) software for independent functions to produce the documents, and (3) relational database management system (RDMS), which is hard to process and synchronize the complicated correlation among the data across the documents. During construction projects, quantity take-off has been conducted at least six times statistically [2]; the same data has been generated seven times on average due to differences in formats [3]. According to the authors' interviews with construction managers, most of them perform information management at the minimum level since they perceived it as troublesome tasks.

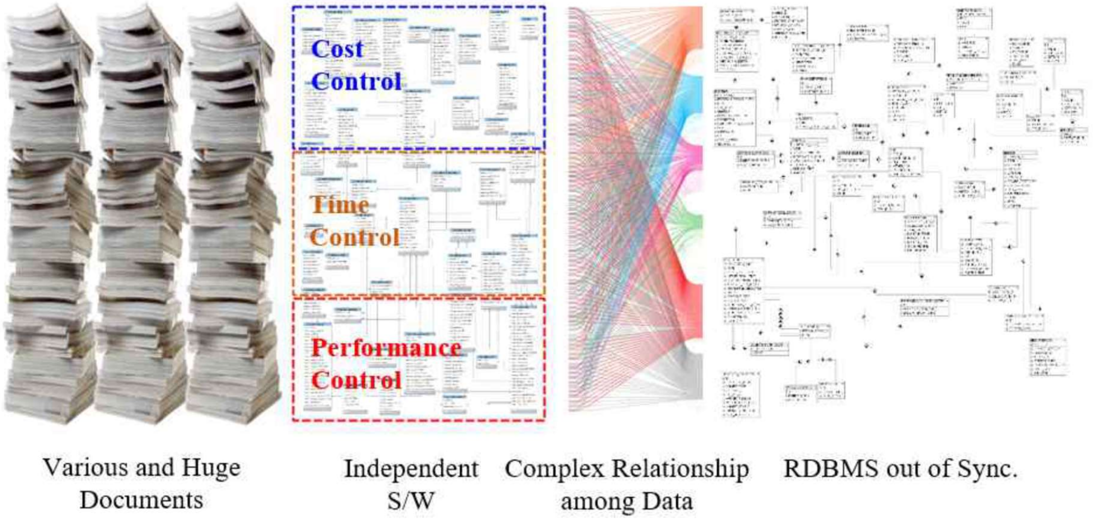


**Figure 1**. Current Data Control for Construction Management

### *2.2. Difficulties in Integrating Cost and Schedule Data*

It may be ideal for both cost and schedule data to be represented by a single hierarchy or be controlled by a single parameter. Unfortunately, lowest-level information units for cost data (consisting of cost items) in the traditional bill of quantities (BOQ) and for project schedule data (comprising schedule items) are on different information hierarchies at different levels. The considerable mismatch among the information units requires multiple hierarchies with complicated parameters for cost-schedule integration [4].

For cost control and reporting in construction projects, cost of account (COA), chart of accounts, or cost breakdown structure (CBS) are used for a logical breakdown of a project into controllable items [5]. Cost items in a detailed cost estimate report are usually represented at an *Operational level*. The operation-oriented cost items are useful to indicate responsible crews and their performance, productivity, periodic payment. Table 1 describes a typical cost item composed of one operation (*HOW*: Formwork) and one element (*WHAT*: Column) information unit at the microscopic aspect of data structure. The cost item focuses on the total cost based on the total quantity of an element, regardless of its locations. Although it seems simple to track an accomplished cost at any point in time, several sets of redundant recalculations are inevitable.

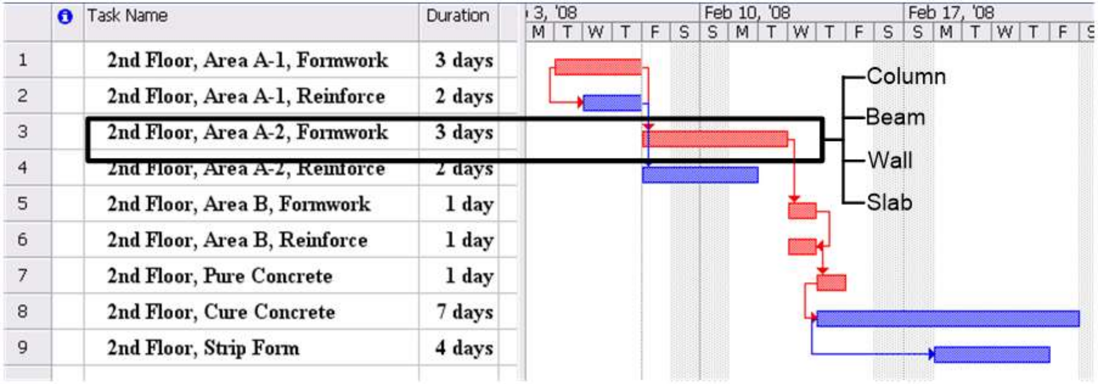
**Table 1.** A Typical Cost Item in the Detailed Cost Estimate [6]

| Cost Code | Description | Unit | Crew | Daily Output | Unit Pricing | | | Total Quantity | Total Cost |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Unit Material | Unit Labor | Unit Equip -ment | | |
| 03110. 410 -6150 | Forms in Place (Job-built plywood, 16" wide columns, 4 use) | SFCA | C-1 | 235 | 0.64 | 4.73 | 0 | 4,263 | 23,302 |

Total Cost = Total Quantity of 16" Column (WHAT) × Sum of unit pricings of Formwork (HOW)
= 4263 SFCA × $ 5.47 = $ 23,302.

In schedule reports, scheduling items are usually represented in a scheduling diagram by focusing on deliverable items (*WHAT*) in a specified location (*WHERE*). Work breakdown structure (WBS) can be leveraged for deriving a scheduling item. As illustrated in Table 2, a typical schedule item is composed of one zone (*WHERE*: 2nd Floor, Area A-2) and one operation (*HOW*: Formwork) information unit. A systematic duration depends on the productivity rate of a schedule item, as shown in the third duration item calculated in Table 2.

**Table 2.** A Typical Schedule Item in a Bar Chart Diagram



| | Task Name | Duration |
|---|---|---|
| 1 | 2nd Floor, Area A-1, Formwork | 3 days |
| 2 | 2nd Floor, Area A-1, Reinforce | 2 days |
| 3 | 2nd Floor, Area A-2, Formwork | 3 days |
| 4 | 2nd Floor, Area A-2, Reinforce | 2 days |
| 5 | 2nd Floor, Area B, Formwork | 1 day |
| 6 | 2nd Floor, Area B, Reinforce | 1 day |
| 7 | 2nd Floor, Pure Concrete | 1 day |
| 8 | 2nd Floor, Cure Concrete | 7 days |
| 9 | 2nd Floor, Strip Form | 4 days |

$Dur3 = Dur_{column} + Dur_{beam} + Dur_{wall} + Dur_{slab}$

Dur column (a duration to complete the column located in the 2nd Fl. Area A-2 (*WHERE*))

= Partitioned Quantity of Element (*WHAT*) / Daily Output of Formwork (*HOW*)

= 214 SFCA / (29.38 SFCA / Crew · Hour)

= 7.3 Work hours · Crew

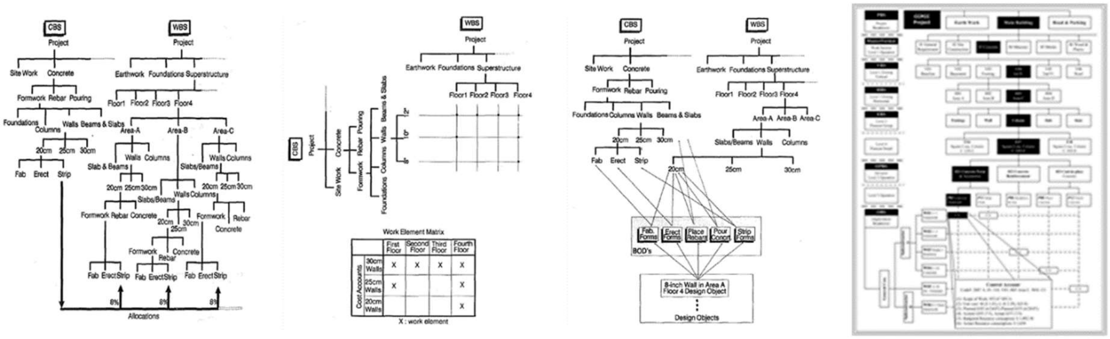
where, $Dur_i$ represents a duration to complete the i item.

The representation of a typical cost item in a detailed cost estimation report typically focuses on the *Operation and Element level* regardless of their location (i.e., Formwork, Column). Then, a bundle of cost items is summarized by the *Operation level* item in a cost estimate report (i.e., Formwork). Finally, the operational cost items are summed into the *Work Section level* in the project budget report (i.e., Concrete) as follows: "Concrete work (Work Section level: 03) > Formwork (Operational level: 031) > Formwork, Column (Operation and Element level: 03111)". On the other hand, the representation of a typical schedule item in a detailed network diagram usually focuses on the *Zone and Operation level* regardless of an element (i.e., 2nd Floor, Framework), and a bundle of schedule items is summarized by the *Location or Operation level* in a master schedule (i.e., 2nd Fl. Construction). Both cost and schedule data may have various levels of detail. Operation data are shared between cost and schedule items, but element or zone data are shared differently. The difference has been regarded as the primary source of integration difficulty

*2.3 Historical Review on Theories of Cost and Schedule Integration*

Cost-schedule integration is required for systematic management of quantity, payment, productivity, material, and labor. The unified management system of cost and schedule data can be used as a basis for (1) securing consistent, accurate, and objective information management, (2) systematic and well-informed construction management, and (3) establishing a database for major projects. Much research has contributed to this area of study and proposed various model for the integration, such as Percent Allocation Model [7], Work Element Model [8], Design Object Model [9], Work-package Model [10], Earned Value Management System [11], Faceted System [12], Flexible WBS System [13].

Although there are differences in detailed procedures and methods, the existing models show common features as follows. They utilize data structure in hierarchical forms and represent data in a top-down approach, ranging from projects to organization (Project – Spaces – Systems/Elements – Tasks – Organization). In addition, they link WBS, CBS, and organization breakdown structure (OBS) by incorporating numbering systems in standard construction information classifications. These connections represent data for construction management in a 2-way matrix with rows and columns. The matrixes establish a complicated relational database, which ties various planned or progress data for quantity, schedule, and cost. However, these models have limitations in achieving data consistency and show insufficient flexibility in the data structure, which produces a vast amount of data whenever low-level information units are added. Yet, there is a dearth of cost-schedule integration theory that is practical and applicable in construction management.



| | First Floor | Second Floor | Third Floor | Fourth Floor |
|---|---|---|---|---|
| 30cm Walls | X | X | X | X |
| 25cm Walls | X | | | X |
| 20cm Walls | | | | X |

(a) Percent Allocation Model (b) Work Element Model (c) Design Object Model (d) Work-package Model

**Figure 2.** Existing Methodologies for Integrating Cost and Schedule Data: (a) Percent Allocation Model [7] by Teicholz; (b) Work Element Model [8] by Hendrickson and Au; (c) Design Object Model [9] by Kim; (d) Work-package Model [10] by Rasdorf and Abudayyeh.

### *2.4 Development of Big Data Technolgy*

The definition of big data in the encyclopedia refers to "a collection of data that cannot be captured, managed, and manipulated with conventional software tools for a period of time, requiring new processing models to have greater decision-making power, insight and discovery and process optimization massive capacity, high growth rates, and diversified information assets" [14]. It is regarded as a promising futuristic technology that generates user-customized and fit-for-purpose information and a more accurate prediction. According to statistics [15], the global big data market is projected to grow to 103 billion dollars by 2027, which indicates a 10.48% annual increase from 2018. It has been expected that this technology brings a new paradigm for competition, productivity growth, innovation, and consumer surplus [16].

In South Korea, the government forecasted the economic ripple effect of big data worth 9.5 billion dollars [17] and selected it as a core technology in the Government 3.0 initiative, a new paradigm of the government workings launched in 2013 [18]. The Korean Ministry of Land, Infrastructure, and Transport established a strategic plan for developing big data application technologies to facilitate innovation in construction project management in 2013 [19]. In line with it, 60 million dollars and 250 million dollars had been invested by 2017 in establishing national spatial data infrastructure and developing infrastructure for big data and spatial data, respectively. Government-led research and development of big data technology for analyzing, using, retrieving construction data continues; a growing number of government projects well explains this phenomenon. The usage of big data in the construction industry mainly focuses on the macroscopic aspect, such as national spatial data, transportation, and communication, while its use at the microscopic level for specific purposes in construction management is limited. Despite ongoing attention to big data, there is a lack of research on its application in construction management. [20]

## 3. Construction Data Analysis Using Big Data Technology

### *3.1. Data Characteristics for Integrating Cost and Schedule Information*

Compared to other industries, the application of integrated cost and schedule management in the construction industry still at a nascent stage. Its core obstacles can be found in characteristics of cost and schedule data in construction projects, as follows.

- **Mismatched level of cost and schedule data**: As addressed in section 2.2, cost data and schedule data in construction projects consist of different information units at different levels, mismatched to each other [4]. In practice, schedules at various levels are created and utilized, such as overall schedules, monthly schedules, weekly schedules, and subcontractor schedules. Each of them requires the cost assignment and aggregation at the appropriate level for corresponding activity levels addressed in the schedule. The database structure flexible to various schedules is critical to achieving data consistency for the integrated management of project cost and schedule.
- **A vast amount of data**: For cost-schedule integrated management, multiple dimensional data for construction projects, such as unit quantities of materials, spaces, building elements, tasks, organization, and time, need to be defined as lowest-level information units. Each dimension contains multiple levels; for instance, the task dimension has more than five levels: construction (level 1) – concrete (level 2) – concrete forming (level 3) – Euro formwork (level 4) – Wall for less than 30-storey (level 5). In addition, task data associated with quantities need to be connected to unit and unit pricing (unit material, unit labor, unit equipment). According to the authors' case studies, around 10 million building elements exist in a 90-million-dollar building project located in Korea, such as structure, envelope, fire safety, mechanical, electrical, and plumbing. Each element need to be decomposed into lowest-level information units and linked to (1) space data (project, section, floor, zone, room); (2) task data (i.e., forming, reinforcing, concrete curing); (3) working organization data (subcontractors, crew); (4) management organization data (contractors, management team, manager, construction manager); (5) time data (year, quarter, month, week, day); and (6) unit pricing data. The level of data decomposition can vary according to the required information level and management level. However, it has been identified that approximately one billion data cells should be prepared for achieving high data consistency for cost-schedule integrated management.
- **Variable data**: The lowest-level information units are generated while operating a construction project. They are updated from time to time as design, construction methods, material, pricing, or contracts change. These updates need to be reflected in the database for cost-schedule integrated management in order to keep data integrity.
- **Complicated data structure**: There have been efforts to establish databases for lowest-level information units using WBS, CBS, and OBS. The hierarchical approaches in introduced methods are not sufficient to incorporate required data at all levels on all dimensions and show limitations in providing enough flexibility and connectivity of data for cost-schedule integration. In addition, the limited flexibility causes an enormous amount of lower data created by data decomposition at a higher level, which results in challenges in information management.
- **Semantic issues**: The existing methods used construction classification standards and their codes for data standardization as well as data mining. Key data represented as a complicated combination of codes hinder intuitive understanding and communication of the data. From the technical perspective, these methods rely on a database to store and manage the codes, whose establishment requires significant time and effort. The specialized software and new procedure of information and administrative management to operate the database cause installation and maintenance costs. These requirements, along with the use of nonintuitive codes, have obstructed the active introduction of cost-schedule integrated management.

### *3.2. Feasibility Analysis of the Use of Big Data Technology for Integrating Cost and Schedule Data*

Compared to traditional data information, big data information has its own outstanding features, which can be summarized as Volume, Variety, Veracity, and Value (as known as 4Vs). The feasibility of the application of big data technology to the cost-schedule data integration has been analyzed.

- As discussed in the above section, the cost-schedule data is vast and created in different forms by different stakeholders in construction projects. Data including quantity, cost, schedule, payment, material, labor is generated by various subcontractors in varied formats (i.e., quantity calculation, BOQ, schedules) for multiple purposes. It changes as the project proceeds according to the changes in design, quantity, material, and construction methods. The 4Vs of big data clearly explain vital attributes of construction data, which is large, heterogeneous, and dynamic with value in decision making [21]. Big data technology can be regarded as a suitable method to integrate and manage the 'big' cost-schedule data in construction management.

- The data accumulated in construction firms are mostly associated with cost and financial management and schedule management. According to the survey with 89 firms in Korea in 2014, 76% and 49% of them have stored cost-related data and schedule-related data, respectively [22]. It infers that the available data for cost-schedule integration is accessible and easy to collect.
- The technical level of handling big data can be categorized according to a data structure (i.e., structured, semi-structured, unstructured), data type (i.e., text, log, sensor, image), data format (i.e., RDB, HTML, XML, Json). Most of the cost-schedule data is structured text data in document format; it facilitates data processing, transmission, store, and evaluation. Technology for collecting, storing, extracting, analyzing, visualizing data has been developed rapidly and published as open-source packages. The high accessibility of technology and the low difficulty in handling cost-schedule data indicate the possibility of the active application of big data technology to the integrated management of cost-schedule.
- Big data technology, which enhances the accuracy of analysis and prediction, is ideal for engineering analysis, including construction engineering [20]. From a statistical perspective, big data infer the patterns within a population using random sampling. The sampling technique is useful for the analysis of limited data size, especially overcoming challenges in collecting and processing data; however, it might have the disadvantage of the overfitted prediction to sample, which leads to the failure to capture real patterns of the population.

### *3.3. Requirements for Data Integration Method*

The improvement of cost-schedule integration leveraging big data technology needs to incorporate the characteristics of cost and schedule data in construction management while resolving the limitations within existing cost-schedule integration models. For developing the applicable method in construction projects, the three requirements have been identified as follows.

- **Data integrity and flexibility**: the integrity of lowest-level information units for cost and schedule data is fundamental for addressing mismatched information levels between them. Data structure needs to be flexible to accommodate various information units at different levels on multiple dimensions to allow data decomposition and aggregation in response to varying levels of project schedules.
- **Efficiency in establishing and transforming data**: The enormous time and effort required to develop the integrated database of cost and schedule data play a role as the root cause of the failure to adopt existing integration models in practice. A novel approach for establishing a database and updating it as projects proceed is essential to achieve its wide application in construction management.
- **Practicability**: the new method needs to support a range of data extraction and analysis associated with cost and schedule management; it can be easily by practitioners. In addition, the method should be fit for the work process and documentation in construction projects. The adoption of new technology in the industry changes not only work activities itself but also work paradigm, including work processes, collaboration methods, information and administrative activities, working knowledge, and organization networks [23]. The model should reflect sufficiently current construction management practices to minimize these changes, leading to the practitioners' hesitation to use existing cost-schedule integration models.

## 4. Big Data Algorithm of Cost and Schedule Data Integration

### *4.1. Proposed Data Structure*

This research proposed a new approach for cost-schedule data integration, based on multiple dimensions comprising the 5W1H data structure, instead of the hierarchical data structure of WBS, CBS, and OBS. Cost and schedule management is a series of operating activities construction projects while asking the fundamental questions continuously – Who did which work in which location with what construction method? How much has it done? When has it done? The 5W1H data structure can be a tool to understand the questions by standardizing them into six dimensions (Where, What, How, When, Who, and Why) [24]. The dimensions here also play a role as metadata, a description of data for supporting efficient data analysis. Figure 3 describes the proposed conceptual data structure for integrated cost-schedule management.

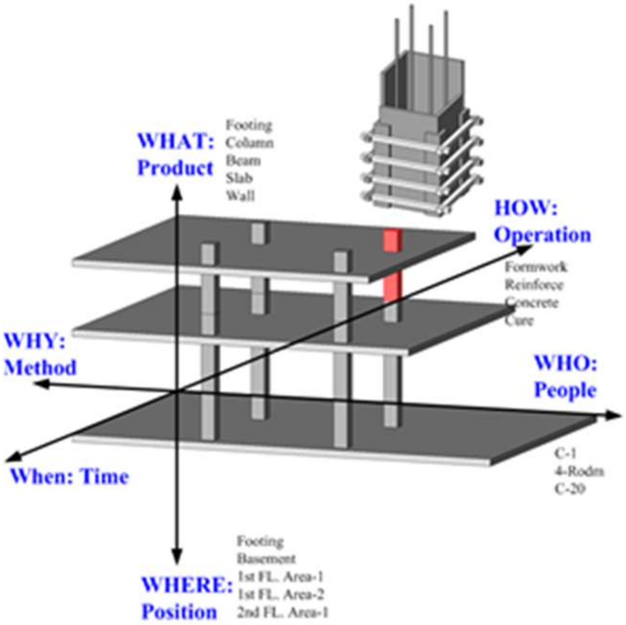


**Figure 3.** Conceptual Data Structure for Integrated Cost-Schedule Management

Two construction classification standards are applied to this research – UniFormat [25] and MasterFormat [6]. UniFormat, a standard for classifying building specifications for cost estimating and analysis purposes, is used as a benchmark for information units in the *WHAT* dimension. In the *HOW* dimension, the level of details for information units is defined based on MasterFormat, which is widely used for specifications for construction contract documents. The interrelationships among the dimensions consisting of stratified information levels are illustrated in Figure 4, except for the *WHY*.

The *WHAT* dimension involves physical or functional elements (i.e., building components and systems). Construction project data comprises the sum of physical elements, together with other information units linked to its information units. In this context, this dimension provides a basic information unit in the proposed structure. Based on the UniFormat, physical elements in this research are decomposed into four levels of details: Major Group > Element Group > Element > Element-detail.

The *HOW* dimension involves construction operations to complete the physical elements. It represents construction methods and contractual relationships. An operation (1) consists of a series of activities, (2) has identical information properties (i.e., material, labor, equipment, indirect cost, and productivity rate), and (3) is performed by the same organization. Based on MasterFormat, information units are be decomposed into four levels of detail: Work Section > Division > Operation > Activity.

The *WHERE* dimension provides the visual perception of project progress and specifies levels of detail of the core units, the physical elements. The zone breakdown structure (ZBS) concept was adopted here. ZBS in building projects can be decomposed into a physical breakdown structure (PBS) and functional breakdown structure (FBS). PBS is composed of the vertical breakdown structure (VBS), indicating vertical positions, and the horizontal breakdown structure (HBS) representing horizontal positions. An individual physical element can be expressed by a horizontal position on the x-axis and a vertical position on the y-axis. The position of every element in the *WHAT* dimension can be represented by using four levels of detail zoning: Facility > VBS only (i.e., 2nd floor) > VBS and HBS (i.e., 2nd floor - Area A) > VBS, HBS, FBS (e.g., 2nd floor - Area A - Elevator Hall).

The *WHO* dimension involves organizations responsible for each operation. Once physical elements and operations have been identified, associated responsibilities and authorities need to be assigned to persons in the project organization. Levels of detail of the organization units can be subdivided into three levels of detail: Contractor > Crew > Individual.

The *WHEN* dimension contains a time factor, a duration to complete a work item. Each work's specific duration is directly related to the exact quantity partitioned by a zone; it becomes a major resource of schedule control. Required levels of detail of *WHEN* unit can be divided into five levels: Year > Quarter > Month > Week > Day.

The *WHY* dimension addresses data required for specifying detailed information of each work item. The dimension is utilized to distinguish items from the database, such as materials supplied by an owner, works completed by a special construction method, and materials carried by a crane. It requires special data management. The dimension is not mandatory but optional.

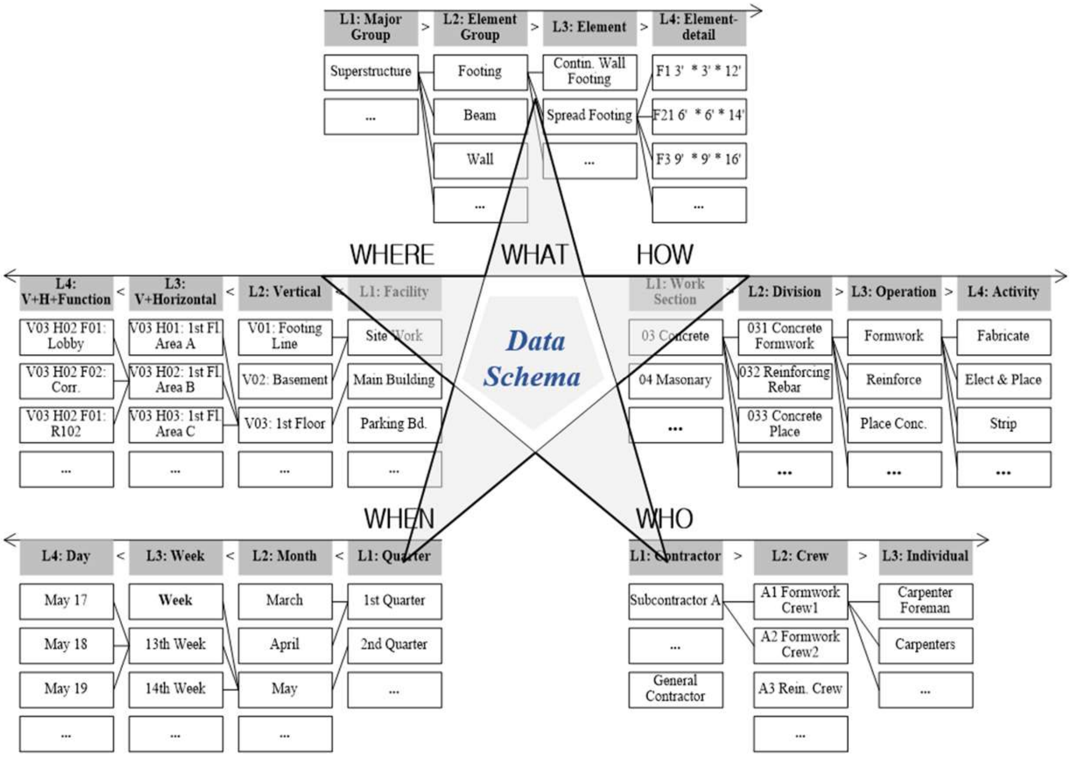


**Figure 4.** Data Star Schema Composed of 5W1H Dimension Focusing on Mandatory Dimensions

### *4.2 Level of Detail of 5W1H Data Structure*

The proposed 5W1H data structure for integrated cost-schedule management can address construction project data at different levels on six dimensions. Table 3 shows how lowest-level information units for cost and schedule data are defined within the suggested structure. The data decomposition in the structure allows answering fundamental questions for performing construction management. For instance, the questions regarding work progress in two different detailed levels can be answered within the data structure – "What is the quantity of concrete that needs to be placed this month?" and "What is the quantity of concrete that needs to be placed for a slab on the third floor of Building A?". For the first question, simple extraction of two information units is necessary: this month (*WHEN* Level 2) and concrete placement (*HOW* Level 3). In the case of the second question, the composition of three extracted data is required: this week (*WHEN* Level 3), Building A (*WHERE* Level 1), third floor (*WHERE* Level 2).

**Table 3.** Levels of Detail of Respective Information Dimension

| Info. Unit | Level of Detail | | | |
|---|---|---|---|---|
| | **Level 1: L1** | **Level 2: L2** | **Level 3: L3** | **Level 4: L4** |
| WHAT | **Major Group**<br>A10: Substructure<br>B10: Superstructure<br>B20: Roof Construction | **Element Group**<br>Footing<br>Wall<br>…<br>Roof | **Element**<br>Continue wall footing<br>Spread Footing<br>…<br>Roof truss | **Element Detail**<br>F0: 24" Conc. Wall Footing<br>F1: 6' ×6'×14" Spread Footing<br>…<br>R3: Metal Fink 11" Span |
| HOW | **Work Section**<br>03: Concrete<br>04: Mansory<br>05: Metals | **Division**<br>031: Conc. Form & Acc.<br>032: Conc. Reinforce<br>…<br>053: Metal Decking | **Operation**<br>0311: Conc. Forming<br>0315: Conc. Accessories<br>…<br>0531: Steel Decking | **Activity**<br>031113: Structural Cast-in-Place Concrete Forming<br>…<br>053123: Steel Roof Decking |
| WHERE | **Facility**<br>Site Work<br>Building<br>Parking Lots | **Vertical**<br>V01: Baseline<br>V02: Basement<br>…<br>V06: Roof | **Vertical + Horizontal**<br>V01 H01: Baseline Area A<br>V01 H01: Baseline Area B<br>…<br>V06 H04: Roof Area D | **Vertical + Horizontal +Functional**<br>V04 H01 Lobby<br>V05 H02 Corridor<br>V05 H03 Room 212 |
| WHO | **Contractor**<br>Subcontractor A<br>Subcontractor B<br>General Contractor | **Crew**<br>A1 Formwork Crew 1<br>A2 Formwork Crew 2<br>A3 Reinforce Crew<br>…<br>G5 E-2: Field Workers | **Individual**<br>A01: 1 Carpenter Foreman, 4 Carpenters, 1 Laborer<br>…<br>A03: 1 Rodman Foreman, 4 Rodman, 1 Equip. Operator | |
| WHEN | **Quarter**<br>1st Quarter<br>2nd Quarter … | **Month**<br>March<br>May… | **Week**<br>12th Week<br>13th Week … | **Day**<br>**17 May**<br>**18 May** |

### *4.3 Three Principles of the Proposed Data Integration Algorithm*

Based on the 5W1H data structure, this research develops a cost and schedule data integration algorithm using big data technology. The three principles have been applied to the development. The first principle is the assumption that the data for construction management exists somewhere in any format. It means that organization charts, contract documents, schedule reports, quantity take-offs, and cost reports are available. Information units for the *WHO* dimension can be retrieved from organization and contract documents, while the units for the *WHEN* can be extracted from schedules. In addition, data for the *WHERE*, *WHAT*, and *HOW* dimensions come from quantity take-offs; one for the *WHY* dimension is captured from cost reports.

The second principle is that a higher level can be decided when a lower-level information unit is defined. For example, F2 6'×6'×14" (Level 4 of *WHAT* information unit: Shortly L4 WHAT, Element Detail) belongs to Spread Footing (L3 *WHAT*, Element), Footing (L2 *WHAT*, Element Group), and finally Superstructure (L1 *WHAT*, Major Group), as represented in the *WHAT* in Table 3. It indicated that the upper levels are automatically defined when lowest-level information is chosen, using standard construction classifications, such as MasterFormat and UniFormat. Slip forming from the cost report is matched to "031113.13 Concrete Slip Forming" in Masterformat; from its code, all four level in *HOW* dimension can be assigned: Structural Cast-in-Place Concrete Forming (L4, MasterFormat code: 031113) < Concrete Forming (L3, MasterFormat code: 0311) < Concrete Forming and Accessories (L2, MasterFormat code: 031) < Concrete (L1, MasterFormat code: 03).

The third principle is as follows: big data technology should be used to collect, structure, extract, and analyze cost and schedule data, which is scattering across different documents. Figure 5 shows the conceptual process of extracting and analyzing cost-schedule data using big data technology. The detailed processes will be discussed in section 4.4.

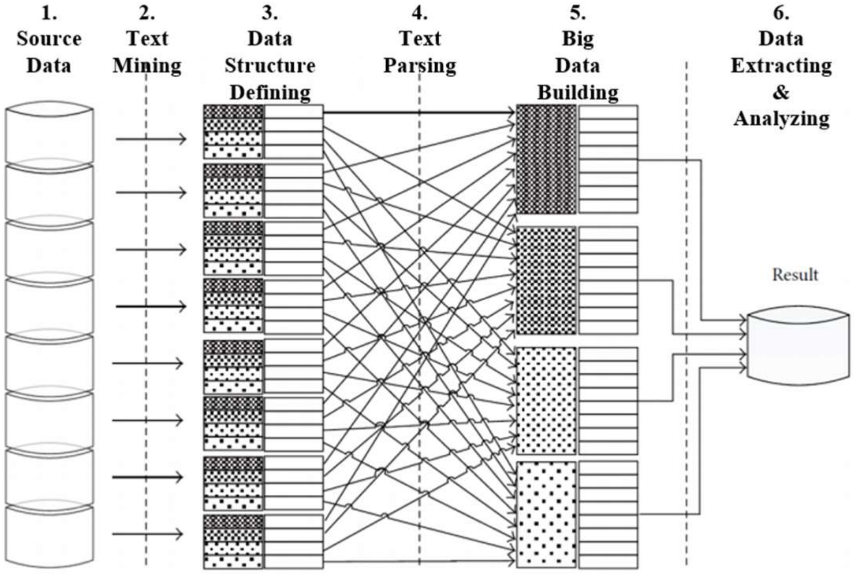


**Figure 5.** Big Data Building Steps for Cost and Schedule Data Integration

### *4.4 Big Data System Architecture for the Integration*

The cost and schedule data integration algorithm driven by big data technology consists of seven modules, as shown in Figure 6. These modules provide functions for (1) building big data from various data sources based on the W5H1 data structure (Module 1 - 4), (2) extracting data for users' requirements (Module 5), (3) analyzing data using extracted data (Module 6), and (4) visualizing analysis results by creating graph, charts, and diagrams (Module 7).

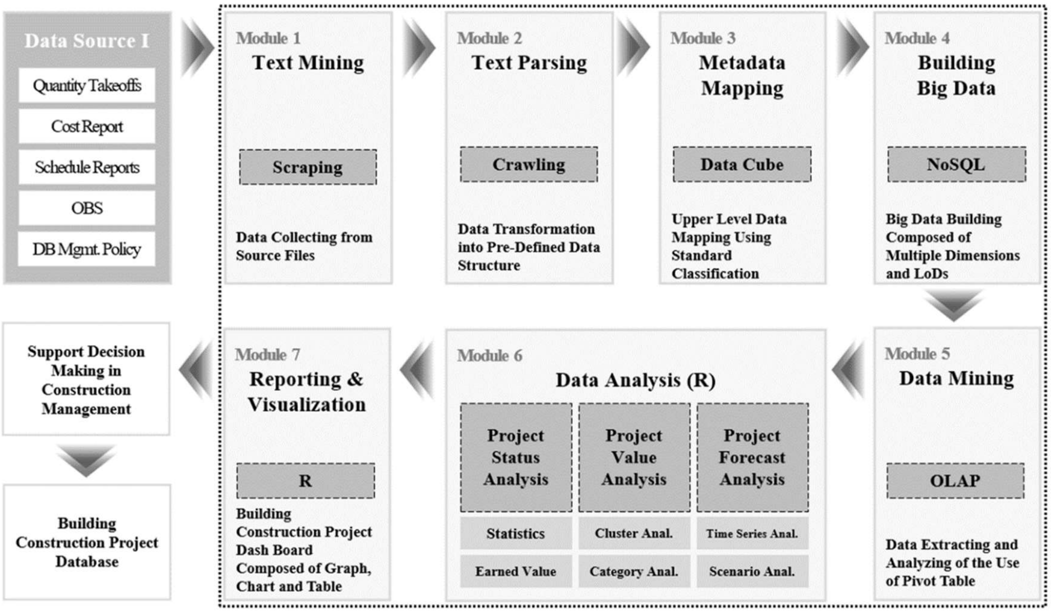

Data Source I
Quantity Takeoffs
Cost Report
Schedule Reports
OBS
DB Mgmt. Policy
Module 1
Text Mining
Scraping
Data Collecting from Source Files
Module 2
Text Parsing
Crawling
Data Transformation into Pre-Defined Data Structure
Module 3
Metadata Mapping
Data Cube
Upper Level Data Mapping Using Standard Classification
Module 4
Building Big Data
NoSQL
Big Data Building Composed of Multiple Dimensions and LoDs
Module 5
Data Mining
OLAP
Data Extracting and Analyzing of the Use of Pivot Table
Module 6
Data Analysis (R)
Project Status Analysis
Project Value Analysis
Project Forecast Analysis
Statistics
Cluster Anal.
Time Series Anal.
Earned Value
Category Anal.
Scenario Anal.
Module 7
Reporting & Visualization
R
Building Construction Project Dash Board Composed of Graph, Chart and Table
Support Decision Making in Construction Management
Building Construction Project Database

**Figure 6.** Big Data System Architecture Composed of Seven Modules for the Data Integration

The detailed definition and function of each module are as follows. Module 1 is *Automatic Text & Semantic Mining* module, which collects information units for cost and schedule data from the data sources (including quantity take-off, cost reports, and schedule reports). The big data scraping technology, which collecting structured or unstructured data in natural language, is applied to this module. Module 2, named *Data Parsing* module, incorporates crawling technology to transform and process the collected data in Module 1 into structured data. Module 3, *Metadata Mapping* module, automates the process of assigning values for higher-levels from a lowest-level information unit based on construction standard classification. This part implements the second requirement in section 4.3 using Python. Module 4 is called *Big Data Building* module and establishes big data of all level information units in six dimensions created in Module 3, according to the 5W1H data structure.

Module 5, *Big Data Mining* module, is designed to extract the required data from the big data. It adopts functions of online analytical processing (OLA) (slice, dice, rotation, rollup, drill-down, query optimizer) to allow data extracting from multi-level of details and multi-perspectives. As Module 6, *Big Data Analysis* module provides various analyses, such as project status analysis, project value analysis, and project forecast analysis, using the extracted data by Module 5. Lastly, Module 7 for *Data Reporting & Visualization* visualizes the analysis results with the help of R or Microsoft Power BI.

The implementation of Module 1 mainly uses Python; the scraped data from sources is translated into Open XML format and checked its validity. Module 2 has been developed by the C# dictionary to define the data structure of the collected data. Module 3 - 5 have been constructed using C# in order to map appropriate metadata to collected text data and to support data extraction with the metadata.

## 5. Case study of the Proposed Algorithm

### *5.1 Summary of Case Study*

To examine the feasibility, the proposed cost-schedule data integration algorithm has been applied to a real construction project of one apartment complex in South Korea. The complex involves nine buildinggs with 774 units. This case study only focuses on the structural part of the complex, which corresponds to 03 Concrete in MasterFormat. All data associated with foundation, walls, slabs, stairs, columns, and beams of nine buildings have been demonstrated.

To enhance understanding of the project data, several email discussions and structural and informal interviews and with project managers were conducted. The full set of project documents (drawings, specifications, quantity-takeoffs, project budget reports, detailed cost estimate report, master schedule, bar chart schedule, and monthly progress curve) was provided, as data sources files for the proposed Module 1. The 20,315 pages of quantity take-offs, 28 pages of cost report, overall and monthly schedule report, and project organization chart are mainly used to develop the integrated big data.

**Table 4.** Levels of Detail of Respective Information Dimension

| **Component** | **Column Items** | | |
|---|---|---|---|
| | **Independent Variables** | **Core Unit Attributes** | **Dependent Variables** |
| Row Items | 4 Level of *WHAT* Unit | Quantity Unit | Quantity |
| | 5 Level of *HOW* Unit | Unit Material Cost | Total Material Cost |
| | 4 Level of *WHERE* Unit | Unit Labor Cost | Total Labor Cost |
| | 8 Level of *WHEN* Unit | Unit Equipment Cost | Total Equipment Cost |
| | 5 Level of *WHO* Unit | Hourly Output | Total Cost |
| | 2 Level of *WHY* Unit | - | Total Required Hours |

The structure of the implemented W5H1 database of collected data from the source files is described in Table 4. The whole database is generated in a big table by aggregating these dependent and independent variables and attributes in the column items. It consists of three major parts; six independent variables, five attributes, and six dependent variables. Therefore, each row item in the database includes independent variables composed of 5W1H dimensions having several levels of detail, attributes, and dependent variables. The whole database comprises 39 columns and 256,540 rows, generating a total of 10,005,060 cells in the spreadsheet, as shown in Figure 7.

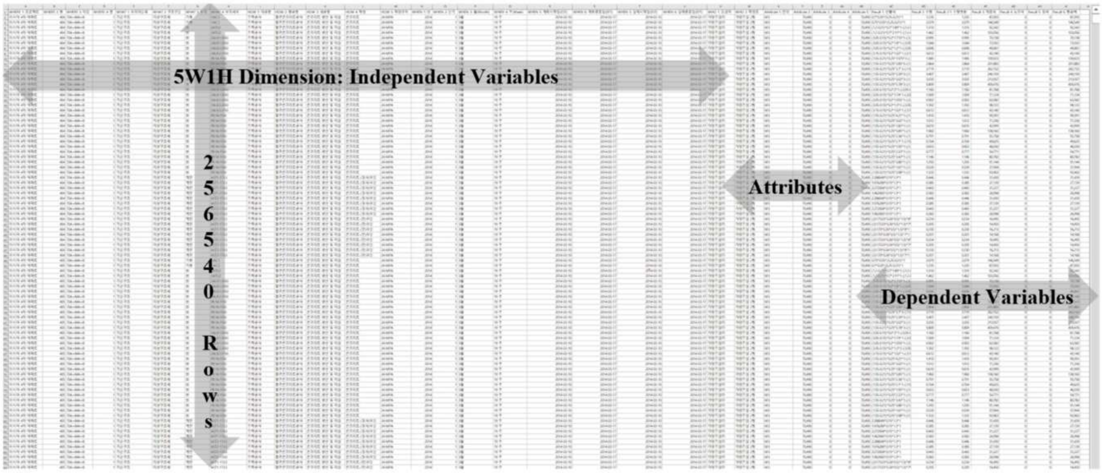


**Figure 7.** Big Data System Architecture Composed of Seven Modules for the Data Integration

*5.2 Verification and Discussion*

The validity of the cost-schedule data integration algorithm based on big data technology is analyzed by evaluating the established big data system. The evaluation focuses on compliance with three requirements for cost-schedule integration methods, discussed in section 3.3.

- **Data integrity and efficiency**: Data in the established system has its basis on the lowest-level of quantity data; it results in a high degree of data integrity in cost-schedule data integration. Furthermore, all lowest-level quantity data in the W5H1 database have 39 types of attributes and relationships with multiple upper-level data via metadata. This data structure created schedule reports at any level of detail by flexibly combining the related information units at different levels on multiple dimensions. For instance, one schedule item in a monthly schedule report, "Structure concrete work on the 13-floor of Building 401 (early start time: 4 March, early finish time: 10 March)", was well subdivided into a weekly schedule by the proposed system. The system automatically generated a corresponding schedule at the week level, "Concrete formwork A Zone on the 13-floor of Building 401 (early start time: 4 March, early finish time: 6 March)". It is an outcome of combining the query results from search conditions, the *WHERE* L3 (Zone A) and the *HOW* L4 (Concrete formwork). Although processing time varies from activity number, the system showed five man·minutes as the average response to time to data synchronization for 100 activities.
- **Efficiency in establishing and transforming data**: In this case study, the demonstration of big data consisting of 10,005,060 cells took less than 20 man·minutes. It is a representative processing time since data building and transformation time differs depending on data engineers' proficiency. However, it can be said that the proposed algorithm and developed modules improve efficiency in integrating cost-schedule data, considering the fact that this task takes typically at least ten man·hours using existing systems. In addition, updating changes in data caused by project changes (i.e., design, quantity, construction methods) requires less than one man·minutes by searching metadata.
- **Practicality**: The developed big data system provides Microsoft Excel as its main interface. Users outside of data engineering can retrieve and extract required data from the big data by using the

Pivot table in Excel; it allows more than 50,000 types of data extraction with simple mouse drag-and-drop. Figure 8 shows the multi-dimension and multi-level data extraction from the demonstrated big data based on OLAP analysis. In the case of big projects, Hadoop or NoSQL can be used instead of Excel to incorporate more amount of data supported by Excel (maximum number of cells: 17,179,870). From the interview with project managers, the high practicality of the proposed system has been identified. After the 10-minute instruction, new users are enabled to use the system and show no-burden in adopting this new system to their projects.

## 6. Conclusions

This research proposed a novel method to integrate cost and schedule data, a long-lasting challenge in the construction industry, by adopting big data technology. The introduced cost-schedule data integration algorithm based on big data technology allows securing consistency and integrity in complicated, various, dynamic, and vast data of construction management. In addition, its flexible data structure provides the ease of data extraction and analysis and high efficiency in building and transforming data that reduce significant time and effort required for information managing in construction projects.

It is expected that the introduced algorithm will contribute to the leap of the construction industry toward a data-driven industry. This research facilitates the accumulation of construction management knowledge, informed decision-making, and the creation of a higher value in the management.

**Author Contributions:** Conceptualization, D.C. and M.L.; algorithm development, D.C. and J.S.; software, M.L.; validation, M.L. and J.S.; formal analysis, D.C.; investigation, M.L.; resources, D.C. and J.S.; data curation, M.L.; writing—original draft preparation, J.S.; writing—review and editing, J.S.; funding acquisition, D.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 20CTAP-C152263-02).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Y. Jung; G. E. Gibson. Planning for computer integrated construction. *J. Comput. Civ. Eng.*, 1999, vol. 13, no. 4, pp. 217–225.
2. M. Oinas. The utilization of product model data in production and procurement planning. In *The life-cycle of construction IT innovations - Technology transfer from research to practice*, 1998.
3. D. Davis. LEAN, Green and Seen (The Issues of Societal Needs, Business Drivers and Converging Technologies Are Making BIM An Inevitable Method of Delivery and Management of the Built Environment). *Journal of Building Information Modeling (JBIM)*, 2007.
4. A. Perera; K. Imriyas. An integrated construction project cost information system using MS AccessTM and MS ProjectTM. *Constr. Manag. Econ.*, 2004, vol. 22, no. 2, pp. 203–211.
5. J. Makepeace. Cost Codes and the Work Breakdown Structure. *US Department of Energy (DOE),* 1997.
6. Construction Specifications Institute (CSI); Construction Specifications Canada (CSC). MasterFormat. 2007.
7. P. M. Teicholz. Current needs for cost control systems. In *Project controls: needs and solutions*, 1987, pp. 47–57.
8. C. Hendrickson; C. T. Hendrickson; T. Au. *Project management for construction: Fundamental concepts for owners, engineers, architects, and builders*. Facsimile edition, Chris Hendrickson, 1989.
9. J. J. Kim, An object-oriented database management system approach to improve construction project planning and control, Doctoral Dissertation. University of Illinois, Urbana, 1989.
10. W. J. Rasdorf; O. Y. Abudayyeh. Cost-and schedule-control integration: Issues and needs. *J. Constr. Eng. Manag.*, 1991, vol. 117, no. 3, pp. 486–502.
11. Q. W. Fleming; J. M. Koppelman. *Earned Value Project Management*. 3. Newt. Square, PA, USA Proj. Manag. Insititute, 2005.

12. L. S. Kang; B. C. Paulson. Information management to integrate cost and schedule for civil engineering projects. *J. Constr. Eng. Manag.*, 1998, vol. 124, no. 5, pp. 381–389.
13. Y. Jung; S. Woo. Flexible work breakdown structure for integrated cost and schedule control. *J. Constr. Eng. Manag.*, 2004, vol. 130, no. 5, pp. 616–625.
14. Y. Ma. Research on Technology Innovation Management in Big Data Environment, in *IOP Conference Series: Earth and Environmental Science*, 2018, vol. 113, no. 1, p. 12141.
15. A. Holst. Big data market size revenue forecast worldwide from 2011 to 2027, Big data market size revenue forecast worldwide from 2011 to 2027, 2018. [Online]. Available: https://www.statista.com/statistics/254266/global-big-data-market-forecast/. [Accessed: 20-Oct-2019].
16. J. Manyika et al. *Big data: The next frontier for innovation, competition, and productivity*, 2011, McKinsey Global Institute.
17. Strategic Committee of Korean Information Technology (CKIT). SMART Government Development Using Bigdata, 2011.
18. Korean Ministry of Finance and Economy. Initiative of IT in the Convergence Smart Era, 2012.
19. Korean Ministry of Land Infrastructure and Transport. 4th Master Plan of Construction Continuous Acquisition & Life-Cycle Support (CALS), 2013.
20. M. Bilal et al. Big Data in the construction industry: A review of present status, opportunities, and future trends. *Adv. Eng. informatics*, 2016, vol. 30, no. 3, pp. 500–521.
21. G. Aouad; M. Kagioglou; R. Cooper; J. Hinks; and M. Sexton. Technology management of IT in construction: a driver or an enabler?. *Logist. Inf. Manag.*, 1999.
22. H. Kim; W. Kim; Y. Yi. Awareness on Big Data of Construction Firms and Future Directions, 2014.
23. Ž. Turk; R. Klinc. A social–product–process framework for construction. *Build. Res. Inf.*, 2020, vol. 48, no. 7, pp. 747–762.
24. D. Cho. Construction Information Database Framework (CIDF) for integrated cost, schedule, and performance control. Doctoral Dissertation, University of Wisconsin, Madison, 2009.
25. Construction Specifications Institute (CSI); Construction Specifications Canada (CSC). UniFormat, 2010.