Analytics protocol for data-driven decision-making in the construction industry

Ashwini Jain
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_theses

Recommended Citation
https://docs.lib.purdue.edu/open_access_theses/1290

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
ANALYTICS PROTOCOL FOR DATA-DRIVEN DECISION-MAKING IN
THE CONSTRUCTION INDUSTRY

by
Ashwini Jain

A Thesis
Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Master of Science in Civil Engineering

Lyles School of Civil Engineering
West Lafayette, Indiana
December 2017
THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Makarand Hastak, Chair
  Lyles School of Civil Engineering and the Division of Construction Engineering & Management
Dr. David Johnson
  Department of Industrial Engineering, and Joint Appointment with Political Science
Dr. Hubo Cai
  Lyles School of Civil Engineering and the Division of Construction Engineering & Management

Approved by:
  Dr. Dulcy Abraham
  Head of the Graduate Program
To my father who couldn’t achieve his dream to become an engineer but fulfilled my late grandfather’s wish by helping me in achieving my engineering dream...
ACKNOWLEDGMENTS

I would like to thank my advisor Professor Makarand Hastak for his continuous and patient support and guidance throughout my M.S.C.E. program. Whenever I encounter difficulties with my research and even my life in Purdue, he has always given me helpful advice and guided me into the right path. It is definitely one of my best luck to meet him and study as his student.

Also, I would like to express my gratitude to my committee members, Professor David Johnson and Professor Hubo Cai, who gave me insightful comments and help to complete my research. Also, I would like to specially thank Professor Theodore J. Weidner, Mr. David Kish, Mr. Kartik Goyani for providing their inputs.

I would like to extend my heartfelt gratitude to my family. Without their endless supports and love, I would not have finished my study successfully. Whenever I am in trouble, they always help me figure it out by giving me valuable advice.

I am also grateful to CEM department faculty and staffs, particularly Ms. Rebecca Miller, Ms. Bonnie Sondgeroth and Mr. Brandon Fulk. They all helped me to finish my M.S.C.E. program smoothly. Moreover, I would like to thank all the SPARC members. The discussion with them enables me to see my research from various perspectives. I would like to add my deep appreciation to Juyeong Choi. He has willingly shared his intellectual experience with me, which helped me a lot in this thesis work.
# TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................................................... viii
LIST OF FIGURES ................................................................................................................................................ ix
ABSTRACT ........................................................................................................................................................ x

1. INTRODUCTION .............................................................................................................................................. 1
   1.1 Background and Need of Analytics ................................................................................................. 1
   1.2 Thesis Statement ............................................................................................................................. 5
   1.3 Research Objectives ....................................................................................................................... 5
   1.4 Scope of the Research ..................................................................................................................... 6
   1.5 Research Methodology ................................................................................................................... 6
   1.6 Thesis Organization ....................................................................................................................... 9

2. WHAT IS ANALYTICS? ....................................................................................................................................... 11
   2.1 Introduction ...................................................................................................................................... 11
   2.2 Definition ....................................................................................................................................... 14
   2.3 Driving Factors of Analytics .......................................................................................................... 15
   2.4 Types of Analytics .......................................................................................................................... 18
       2.4.1 Descriptive analytics ............................................................................................................. 19
       2.4.2 Diagnostic analytics ............................................................................................................. 19
       2.4.3 Predictive analytics ............................................................................................................. 20
       2.4.4 Prescriptive analytics ........................................................................................................... 20

3. ANALYTICS IN THE CONSTRUCTION INDUSTRY .................................................................................. 21
   3.1 Current Status of Analytics in the Construction Industry .............................................................. 21
   3.2 The State of Analytics Practice in the Construction Industry ....................................................... 23

4. THE STATE OF ANALYTICS PRACTICE IN DIFFERENT INDUSTRIES ........................................ 26
   4.1 Finance and Banking ...................................................................................................................... 27
   4.2 E-commerce and Retail .................................................................................................................. 28
   4.3 Manufacturing and Supply Chain .................................................................................................. 29
   4.4 Health Care and Medical .............................................................................................................. 31
   4.5 Airlines ......................................................................................................................................... 32
   4.6 Sports .......................................................................................................................................... 33
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7</td>
<td>Other Industries</td>
</tr>
<tr>
<td>5.</td>
<td>FUTURE OF ANALYTICS IN THE CONSTRUCTION INDUSTRY</td>
</tr>
<tr>
<td>5.1</td>
<td>Data- and Information-Related Concerns in the Construction Industry</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Challenge 1: Difficulty in data selection</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Challenge 2: Data fragmentation</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Challenge 3: Inconsistent and out-of-context information</td>
</tr>
<tr>
<td>5.2</td>
<td>Challenges in Implementing Analytics</td>
</tr>
<tr>
<td>5.3</td>
<td>The State of Analytics as an Art</td>
</tr>
<tr>
<td>5.4</td>
<td>Need for Analytics in the Construction Industry</td>
</tr>
<tr>
<td>6.</td>
<td>DEVELOPMENT OF A PROTOCOL TO FACILITATE DATA-DRIVEN DECISION-MAKING IN THE CONSTRUCTION INDUSTRY</td>
</tr>
<tr>
<td>6.1</td>
<td>The Protocol</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Conceptualization: Describe the current business situation</td>
</tr>
<tr>
<td>6.1.1.1</td>
<td>Problem Statement</td>
</tr>
<tr>
<td>6.1.1.2</td>
<td>Business Implications</td>
</tr>
<tr>
<td>6.1.1.3</td>
<td>Current Solutions</td>
</tr>
<tr>
<td>6.1.1.4</td>
<td>Entities Involved</td>
</tr>
<tr>
<td>6.1.1.5</td>
<td>Define the Scope of the Data Analytics Project</td>
</tr>
<tr>
<td>6.1.1.6</td>
<td>Data Available</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Design: Understand the problem</td>
</tr>
<tr>
<td>6.1.2.1</td>
<td>Exploratory Analysis</td>
</tr>
<tr>
<td>6.1.2.2</td>
<td>Problem Structure</td>
</tr>
<tr>
<td>6.1.2.3</td>
<td>Analyses Overview</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Development: Understanding the data</td>
</tr>
<tr>
<td>6.1.3.1</td>
<td>Data Quality</td>
</tr>
<tr>
<td>6.1.3.2</td>
<td>Data Structure Needed</td>
</tr>
<tr>
<td>6.1.3.3</td>
<td>Data Sources</td>
</tr>
<tr>
<td>6.1.3.4</td>
<td>Data Collection</td>
</tr>
<tr>
<td>6.1.3.5</td>
<td>Supplementary Information</td>
</tr>
<tr>
<td>6.1.3.6</td>
<td>Data Cleaning</td>
</tr>
<tr>
<td>6.1.4</td>
<td>Refinement: Data refinement</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>6.1.4.1</td>
<td>Slicing and Dicing:</td>
</tr>
<tr>
<td>6.1.4.2</td>
<td>Splicing:</td>
</tr>
<tr>
<td>6.1.4.3</td>
<td>Layering:</td>
</tr>
<tr>
<td>6.1.5</td>
<td>Analyses: Data analyses and modeling</td>
</tr>
<tr>
<td>6.1.5.1</td>
<td>Data Analysis:</td>
</tr>
<tr>
<td>6.1.6</td>
<td>Outcome: Interpretations and outcomes</td>
</tr>
<tr>
<td>6.1.6.1</td>
<td>Deployment:</td>
</tr>
<tr>
<td>7.</td>
<td>IMPLEMENTATION OF THE PROTOCOL IN CASE STUDIES</td>
</tr>
<tr>
<td>7.1</td>
<td>Case Study 1: The Fire Engineering and Maintenance Department at Purdue University</td>
</tr>
<tr>
<td>7.2</td>
<td>Case Study 2: Housing Market Outlook and Business Expansion Locations</td>
</tr>
<tr>
<td>8.</td>
<td>CONCLUSIONS</td>
</tr>
<tr>
<td>8.1</td>
<td>Summary</td>
</tr>
<tr>
<td>8.2</td>
<td>Research Contributions</td>
</tr>
<tr>
<td>8.3</td>
<td>Limitations of the Research</td>
</tr>
<tr>
<td>8.4</td>
<td>Recommendations for Future Research on Analytics</td>
</tr>
<tr>
<td>REFERENCES</td>
<td></td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3-1: The State of Analytics Practice in the Construction Industry ........................................ 23
Table 6-1: Definitions of Slicing and Dicing from Other Researchers ........................................... 71
Table 6-2: Sample Projects of an Organization .................................................................................. 72
Table 6-3: Sample Projects of an Organization After Slicing ........................................................... 73
Table 6-4: Sample Projects of an Organization After Slicing and Dicing ........................................... 74
Table 7-1: Implementation of the Protocol on the Fire-Safety Case Study (Part 1).......................... 84
Table 7-2: Implementation of the Protocol on the Fire-Safety Case Study (Part 2)............................. 97
Table 7-3: Implementation of the Protocol on the Fire-Safety Case Study (Part 3)............................ 100
Table 7-4: Implementation of the Protocol to Determine the Housing Market Outlook and Assess Suitable Regions for Business Expansion (Part 1) ................................................................. 109
Table 7-5: Implementation of the Protocol to Determine the Housing Market Outlook and Assess Suitable Locations for Business Expansion (Part 2) ................................................................. 119
LIST OF FIGURES

Figure 1-1. Flowchart of the research methodology .......................................................................................... 7
Figure 2-1. The four driving forces behind the analytics movement (Liberatore and Luo 2010) ............................. 16
Figure 4-1. How Fortune 1000 executives use data and analytics (Bean 2017) .................................................. 27
Figure 4-2. How some companies are using analytics in different business areas (Davenport 2006). ......................... 34
Figure 5-1. Challenges faced by organizations seeking to become more data-driven (Lavalle et al. 2011) .......... 41
Figure 5-2. Data characteristics sought by managers (Lavalle et al. 2011) ....................................................... 44
Figure 5-3. Four-step process view of analytics (Liberatore and Luo 2010) ......................................................... 45
Figure 5-4. Process of executing analytics (Jugulum 2014) .............................................................................. 46
Figure 5-5. Business processes improvement framework using analytics (Shankararaman and Gottipati 2015) .... 48
Figure 5-6. The human problem-solving process model (Seddon et al. 2016) .................................................... 48
Figure 6-1. Summary of data analytics protocol for the construction industry ...................................................... 53
Figure 6-2. Phase 1: Conceptualization- Describe the current business situation ............................................... 54
Figure 6-3. Phase 2: Design- Understand the problem ......................................................................................... 61
Figure 6-4. Phase 3: Development- Understanding the data ................................................................................. 65
Figure 6-5. Phase 4: Refinement- Data refinement ................................................................................................. 70
Figure 6-6. SQL joins (W3schools.com 1999) ...................................................................................................... 75
Figure 6-7. The concept of layering (Stevens et al. 2012) ................................................................................... 77
Figure 6-8. Map algebra of layering process (Foote and Lynch 2000) ................................................................. 78
Figure 6-9. Map algebra of layering process for the same analysis (Foote and Lynch 2000) ............................... 78
Figure 6-10. Layering example ........................................................................................................................... 79
Figure 6-11. Placement of layers (University of Washington 2005) ................................................................. 79
Figure 6-12. Phase 5: Analyses- Data Analyses and Modeling ........................................................................... 81
Figure 6-13. Phase 6: Outcome- Interpretations and outcomes .............................................................................. 82
Figure 7-1. Hierarchy of analyzing variability .................................................................................................... 93
Figure 7-2. Payroll amount charged vs. WO type ............................................................................................... 94
Figure 7-3. Historical data of payroll amount charged vs. sliced WO type ......................................................... 95
Figure 7-4. Variability in WO cost accounts. ................................................................. 95
Figure 7-5. Historical data of WO amount charged for PM activity type. ......................... 98
Figure 7-6. Variability in PM activity type cost accounts. ............................................... 99
Figure 7-7. Left joining of two databases. ........................................................................ 104
Figure 7-8. Clustering of buildings vs. unit WO amount charged. .................................... 105
Figure 7-9. Clustering of WO description vs. WO amount charged. ............................... 106
Figure 7-10. Exploratory analysis of construction starts, and permits issued. ................... 116
Figure 7-11. Level and trend for housing permits for single-family houses. ...................... 117
Figure 7-12. The Holt-Winters forecasting model for single-family housing permits in the southern region of the United States................................................................. 118
Figure 7-13. Average household growth rates (2010-2017) at the county level. ............... 124
Figure 7-14. Median home values at the county level....................................................... 125
Figure 7-15. Total crime indexes at the county level............................................................ 126
Figure 7-16. Unemployment rates at the county level......................................................... 127
Figure 7-17. Best suited counties for housing development after the layering process. ....... 128
Figure 7-18. Report on the best-suited counties for housing development in the South....... 129
Due to the unique nature of construction projects and the need for leaders in the business to have knowledge from multiple fields, it has become difficult for decision makers to confidently make choices in their complex working environment. Over time, data availability and computational power have both improved. Moreover, technological advancements from the past few years have overwhelmed the industry by generating a diverse amount data from various sources. This large amount of data makes it possible to find nuggets of information. However, in this digital era, collecting data is not the problem, but how to extract useful information is a big challenge. Existing analytics methodologies do not fulfill the needs of the construction industry because of its unique nature. The involvement of multidisciplinary participants is essential for sound decisions to be made in the industry. There are no integrated or unified solutions for the construction industry. In the absence of such solutions, people are making biased decisions based on their personal experiences. Poor information management is a long-standing problem for the construction industry because most data is buried in papers spread across different departments. Therefore, a new paradigm is needed to facilitate decision-making processes in the construction industry and build an analytical culture for organizations.

This thesis presents the State-of-the-Art and State-of-the-Practice of analytics in different industries. However, existing analytical models and frameworks cannot resolve the construction industry’s data- and information-related challenges. Therefore, a new protocol is developed after understanding the specific needs of the construction industry. Later, to understand and implement the protocol, two case studies were solved using real-world data of the construction industry. This work serves as a foundation of analytics in the construction industry by providing a common platform for industry leaders. This platform will also help organizations overcome their individual, managerial, and cultural barriers.
1. INTRODUCTION

1.1 Background and Need of Analytics

In recent years, the construction industry has sought increased globalization, diversification, geographical expansion, and sources of business to fulfill market demand and improve project performance. The industry has seen continual growth and sustained marginally positive results, even during the economic crisis, due to the availability of large infrastructure projects around the world (Roca et al. 2012). However, shortage in finances, intense competition, low margins, and the increased cost of raw materials have created challenges for many organizations.

The construction industry is a highly competitive one, and construction firms go through a selective and aggressive bidding process when looking to secure work. Therefore, companies are constantly facing tremendous pressure to improve their efficiency and effectiveness. Past financial crises have led to a great deal of uncertainty and reduced business margins in the industry (Zhao et al. 2015; Al-Malkawi and Pillai 2013; Koopman and Székely 2009). Due to the high level of competition and rapid developments in technology, a contractor’s strategy must involve differentiating itself (Ho 2015). Therefore, managers have started evaluating performance measures in order to monitor their business processes. This type of performance measurement is also a form of benchmarking. But benchmarking is not only about analyzing numbers; it also involves understanding the root cause of a problem and taking corrective action as needed.

Due to the unique nature of construction projects and the need for leaders in the business to have knowledge from multiple fields such as business, management, accounting, and law, it has become difficult for decision makers to confidently make choices in their complex working environment. Hence, there is a need to bridge multiple disciplines so that decision makers can maximize potential benefits for their organizations while minimizing any risks (Ibrahim et al. 2013; Aouad et al. 2002; Hegazy et al. 2001). Past researchers have focused on critical parameters, the construction processes, and decision support systems. However, even with this newfound knowledge construction projects still experience significant delays and run over budget, which often puts decision makers in difficult positions. Moreover, construction projects typically have location constraints. Even though many researchers have worked on finding optimal solutions for specific problems, it is difficult to generalize construction project models and create accessible
knowledge at the industry level. For example, Ke (2013) studied the project delivery process for projects in Singapore, Beijing, Hong Kong, and Sydney, but this process cannot be reasonably implemented in other countries. Similarly, Love et al. (2005) explored the time-cost relationship for Australian building projects, but this relationship is not applicable to other countries. The unique location of a given construction project makes it difficult to generalize knowledge based only on data. Finally, it must be acknowledged that different countries have their own building codes, tax structures, accounting systems, and labor laws.

The real problem the construction industry must address does not concern how information and data should be gathered but rather how it should be applied (Brandon et al. 2017; Naderpajouh et al. 2015; Azhar 2005). In the construction industry, decision makers continue to rely on qualitative approaches despite the abundance of available data (Oh et al. 2016; Azhar 2005). Analysts in the industry may use questionnaires and other technological mediums to collect data for a given purpose, but the decision makers themselves will often not use this same data and can end up making choices based on their intuitions, thus leading their organizations to undesirable outcomes (Love et al. 2014). For example, risk analysis is often measured qualitatively (Yoon et al. 2014). The issue at hand is the lack of data-driven decision-making in the construction industry.

The construction industry has a history of focusing on qualitative approaches to problem-solving by depending on executive judgments, historical analogy, and panel consensus (He et al. 2015; Luu et al. 2008; Olander 2007; Chan et al. 2001). The dynamic nature of construction projects and a project manager’s inability to derive useful information from data can lead to reactive decisions being made that ultimately result in the failure of a project. Today, many decision makers in the construction industry still rely on qualitative analysis, their intuition, and their personal judgment because they have fully embraced the decision-making practices of (Ahmadu et al. 2015; Esa et al. 2014; Akintoye and MacLeod 1997)

1. conducting surveys,
2. presenting information to a jury of experts, and
3. implementing the Delphi method.

In the past, collecting accurate data and integrating various sources of data has been difficult because of the numerous complex and interlinked systems in use (Akhavian and Behzadan 2012). Over time though, data availability and computational power have both improved, and now
analytics based on quantitative data is readily available. In the construction industry, however, historical data is not used after a project has been finished, so a great deal of knowledge is obtained from experienced peers and through trial and error (Hammad et al. 2014; Wu and Soibelman 2006). But the large amount of historical data that exists makes it possible to find nuggets of information that can help a person understand the underlying nature of a problem. By analyzing the data and measuring the response to a change in an activity, decision makers can capture new knowledge and get ahead of their competitors. Moreover, combining experiential knowledge with quantitative analysis can offer important contextual knowledge (Dainty et al. 2003). Hence, it is worthwhile to find a way to integrate existing knowledge with data.

It is important for decision makers to understand the context of a given problem and be able to extract relevant information from available data in order to come up with a suitable solution. Different users will have various assumptions, views, and understanding of a problem. Contextual knowledge can help in reducing uncertainty by providing the big picture for a problem and enhancing collaboration. Therefore, there is a need for a modular and extensible mechanism that can facilitate decision-making processes. Today’s existing frameworks and models are unable to meet the demands of various industries because they only explain “what happened” and “why it happened” (Azhar 2005). However, it has become essential to know “what is happening now, what is likely to happen next, and what actions should be taken to get the optimal results” (Lavalle et al. 2011).

The underlying knowledge that people can access and its nature varies according to the relevant industry (El-Diraby 2012). The complex operating environment of a construction project makes it unique from other types of endeavors (Hastak and Koo 2016). Multiple people and parties have a part in executing a construction project, so it is essential to have streamlined communication among them. It is unquestionable that different users will have distinct needs, views, and guiding principles based on their expertise and work. While these differences are necessary for the completion of a given project, they can also create a gap in communication between stakeholders and project teams in the construction industry.

A lot of research on construction informatics in the construction industry has focused on data collection and different aspects of data processing. In addition, many analytical tools are already available to the construction industry. However, industry professionals still feel that they need more tools, especially analytical ones (Armstrong and Gilge 2016; Baloi and Price 2003).
Quantitative data gives users greater flexibility in extracting valuable information. The biggest challenge in the construction industry is that numerous project-related factors are subjective and qualitative, which means they are hard to measure. Moreover, technological advancements from the past few years have overwhelmed the industry by generating a diverse amount data from various sources. In reaction to this situation, many researchers have been working on innovative methods to capture data. However, the lack of clarity regarding construction data sets continues to hinder decision-making processes and can keep individuals from making robust decisions. Therefore, having a systematic procedure to implement analytics can help decision makers in the construction industry embrace the practice of using data to make critical choices (Naderpajouh et al. 2015; Tyne et al. 2002).

Over the past few years, most industries have gained greater access to data. The presence of high-quality data has driven decision makers’ interest in analytics. However, Owen et al. (2010) mention that data or information is often not captured properly in the construction industry because of the fragmented work process. Meanwhile, Törmä et al. (2015) states that the complex flow of information in a given construction project creates only partial solutions. Moreover, it is unclear how the growing availability of data can help decision makers in the construction industry. In this digital era, collecting data is not the problem, but how to extract useful information to create value in a construction project and the construction industry as a whole is a big challenge. Decision-making is an intensive human-based activity, and insights come from people’s minds, not computers. Ultimately, construction project managers are the ones who must make complex decisions. Because of the dynamic nature of projects, it is difficult to create mathematical models to aid in decision-making, and it must be acknowledged that problems cannot be solved using technology. Instead, technology is simply a medium that can be used to achieve people’s goals.

We believe that the existing analytics methodologies do not fulfill the needs of the construction industry because of its unique nature. The involvement of multidisciplinary participants is essential for sound decisions to be made in the industry. In construction projects, design information is shared using 4D modeling and building information modeling. However, no integrated or unified solutions for the construction industry currently exist (Owen et al. 2010). In the absence of such solutions, people are making biased decisions based on their personal experiences (Ahiaga-Dagbui and Smith 2014a). For example, a decision maker may focus on solving a problem without having crucial pieces of information while an analyst may focus on
applying advanced analytics without understanding the big picture for a project. Poor information management is a long-standing problem for the construction industry because most data is buried in papers spread across different departments (Westin and Sein 2014; Gyampoh-Vidogah et al. 2003). Therefore, a new paradigm is needed to facilitate decision-making processes in the construction industry and build an analytical culture for organizations.

But as previously mentioned, the problem is not about the availability of analytical tools in the construction industry. Many past researchers have tried to solve specific construction problems using analytical tools. The current approach, which involves human intervention, requires data translation and interpretation at multiple levels and ultimately leads to the loss of information (Karan et al. 2015). Thus, construction organizations that are collecting and analyzing data are still facing challenges in implementing analytics because of the process’s complexity and the lack of relevant talent in the industry (Skibniewski and Golparvar-Fard 2015; Ahiaga-Dagbui and Smith 2014b). Establishing a defined action plan for analytics use in the industry can improve decision-making processes overall. In fact, a systematic approach can set common ground upon which construction organizations can facilitate data-driven decision-making (Oh et al. 2016; Kargul et al. 2015; Skibniewski and Golparvar-Fard 2015; Yang and Shen 2015; Kärnä et al. 2009). Therefore, there is a need to define a systematic process for analytics implementation within the construction industry in order to bring everyone to the same platform and provide a medium to think through analytics with.

1.2 Thesis Statement

Establishing a protocol for the systematic implementation of analytics in the construction industry will facilitate data-driven decision-making.

1.3 Research Objectives

The focus of this study includes the following objectives:

1. to understand the application of analytics in the construction industry,
2. to investigate how analytics is being used in other industries,
3. to understand the challenges in implementing analytics in the construction industry,
4. to identify what kind of protocol is needed to systematically collect and process information for decision-making in construction organizations. Paired with this objective is the goal of developing a systematic process that can facilitate questions-based thinking for a robust decision-making process.

This work serves as a foundation of analytics in the construction industry by providing a common platform for industry leaders. This platform will also help organizations overcome their individual, managerial, and cultural barriers.

1.4 Scope of the Research
The scope of this study includes the development of a protocol to facilitate data-driven decision-making in the construction industry. The developed protocol will be catered to the industry’s specific needs; hence, it may not be applicable to other industries. This protocol can help construction industry organizations regardless of their size. Finally, the protocol will be systematically evaluated through two case studies. These case studies will prove that the proposed protocol can be applied to diverse types of problems within the construction industry. The first case study enhances the decision-making process of the Fire Engineering and Maintenance Department at Purdue University. The second case study provides a data-driven outlook for the housing market in the United States and proposes a strategy for a land development company’s business expansion.

1.5 Research Methodology
Analytics is not regularly used in the construction industry today. Therefore, conducting a survey using a questionnaire will not add value to the goal of implementing analytics. To understand the current gap in technology, an extensive literature review has been done to facilitate the exploratory research work. The flow of the research methodology is laid out in Figure 1-1.
What is analytics?

Understanding analytics in the construction industry

Understanding how other industries are using analytics

Evaluating the need for an analytics protocol in the construction industry

Understanding the state of analytics as an art

Understanding data- and information-related challenges in the construction industry

Developing the protocol

Implementing the protocol in two real case studies

Figure 1-1. Flowchart of the research methodology.

Improvements in technology have rapidly increased in the amount of data available to the construction industry today. This data can be used for many reasons such as project estimating and costing, project planning, project scheduling, project control, materials management, and material procurement. Analytics has become a buzzword, and various industries are using analytics to make data-driven decisions. Thus, there is also an opportunity for the construction industry to improve its decision-making processes using data analytics.

The first part of the literature review, presented in Chapter 2, seeks to understand aspects of analytics such as its driving factors and various types. The second part of the literature review addresses the present status of analytics in the construction industry. There has already been a lot of research done on informatics, especially with a focus on data collection, and data processing, but these topics are presented in Section 3.1. Moreover, Section 3.2 explores the various studies that have been done on mathematical modeling and data analysis in the construction industry. Despite all of this research, however, construction organizations are still lacking data in their decision-making processes. In fact, the construction industry is considered one of the least
digitized industries (Manyika et al. 2011). As a result, decision makers frequently have to rely on qualitative methods to arrive at sound choices.

Before moving further into the details of analytics, it is important to understand how analytics is being used in other industries, what kinds of problems are being solved with analytics, and what analytical methods those other industries are using. Chapter 4 covers different industries such as finance and banking, e-commerce and retail, manufacturing and supply chain, health and medical, airlines, and sports. Using these industries as case studies will help in the identification of their best practices with analytics.

After gaining knowledge on the widespread use of analytics in other industries, it is important to examine the specific problems that the construction industry faces and think about how they can be solved. These problems will likely differ from the challenges other industries have faced because of the construction industry’s unique nature. Therefore, Section 5.1 discusses some of the data- and information-related problems that managers in the construction industry face in their day-to-day job. Using a systematic and collaborative approach, these problems can be solved and ultimately help improve conditions in the construction industry (Ren et al. 2013).

As previously mentioned, understanding the general challenges organizations outside the construction industry have faced with regard to analytics will help shed light on its implementation in construction organizations, so this topic is addressed in Section 5.2. Researchers in other industries have developed various tools to improve the decision-making process. However, some of these tools were altered based on the needs of a specific industry while others remain too generalized to capture the level of detail needed for construction projects. To understand the critical components in the life cycle of the analytics process, a discussion of analytics as an art is presented in Section 5.3. Based on analytics as an art, it is clear that researchers are already working on creating standard models or frameworks for its implementation. Understanding these existing models and frameworks from other industries better informs the development of a standard procedure for analytics implementation in the construction industry. However, it should be noted that these studies fail to mention the type of questions that should be asked when using analytics and how business value can be created. Although, the models and frameworks introduced in Section 5.3 cannot resolve the construction industry’s data- and information-related challenges, they were taken into account for the development of the protocol in Chapter 6. Finally, Section 5.4
addresses the unique nature of the construction industry and differentiates it from other industries. This consideration also helps in the creation of the new protocol for the construction industry.

After understanding the specific problems that the construction industry must deal with and the general challenges that come with implementing analytics, a new protocol is developed in Chapter 6. Aspects of analytics as an art from Section 5.3 are used as best practices that inform and inspire the new protocol. To understand and implement the protocol, two case studies using real-world data are presented in Chapter 7. The first case study, described in Section 7.1, addresses how the Fire Engineering and Maintenance Department at Purdue University can best gain insights from their available data. This case study also demonstrates how analytics can help in preparing a robust budget. In Section 7.2, the second case study presents a data-driven outlook for the housing market in the United States and identifies the best locations for business expansion. The final results from both cases studies were discussed with relevant decision makers, and the decision makers have agreed that the proposed recommendations could potentially solve their business problems.

1.6 Thesis Organization

This thesis is organized into eight chapters. Chapter 1 provides an introduction of the research work. It explains the need for analytics in the construction industry by covering what problems can be solved and how they can be solved. In addition, Chapter 1 covers the research objectives, which define what will be done in this thesis, and the scope of the research, which clearly sets the boundaries of the research work. Finally, Chapter 1 discusses the steps that will be taken during the research process.

Chapter 2 provides a detailed description of analytics, includes a formal definition for it, and discusses the driving factors behind the increasing use of analytics. This chapter also explores the different types of analytics in detail.

Chapter 3 explains the status of analytics in the construction industry. The chapter includes information on the various research work that has been done by researchers focused on the construction industry.

Chapter 4 covers the application of analytics in different industries such as finance, banking, retail, manufacturing, and health care. The unique characteristics of analytics implementation in
each industry are included along with the methodologies and approaches these industries have used to incorporate analytics into their decision-making processes.

Chapter 5 starts by presenting the data- and information-related problems construction organizations face. Then it explores the general challenges that different organizations outside of the industry have had to overcome while implementing analytics. Although this knowledge is useful, current aspects of analytics as an art cannot be used to resolve the construction industry’s unique challenges. Hence, Chapter 5 looks at the construction industry’s specific challenges in detail and identifies the reasons for them. The chapter concludes with a discussion of why there is a need to develop a systematic procedure to implement analytics in the construction industry.

Chapter 6 explains the guidelines of the proposed protocol in detail. The development of the protocol for construction analytics is presented in a thorough manner to show how the protocol can help facilitate data-driven decision-making in the construction industry.

Chapter 7 covers two case studies to show how the systematic implementation of analytics in the construction industry can be effectively carried out.

Chapter 8 discusses potential areas where analytics can be implemented in the construction industry and presents the study’s contribution to the existing body of knowledge. Moreover, the chapter provides examples of how this research work can help construction organizations. Finally, some of the study’s limitations are discussed for further consideration by other researchers and the continuation of this research work.
2. WHAT IS ANALYTICS?

2.1 Introduction

The increasing amount of complexity in business has escalated the level of uncertainty present in the decision-making process. As a result, many organizations have seen reduced profits. It is in the interest of an organization to find quicker and more robust decision-making processes because of the strategic advantages they can offer. Being able to make more sound decisions can increase an organization’s competitiveness in the market as well. However, the challenge lies in how to gain actionable insights from the overwhelming amount of data available.

Over the past few years, significant advancements in computer technology have made it possible to record, store, manage, access, and communicate a wealth of information. Unique data collection opportunities using the Internet, or the Web have shifted people’s focus to big data. Today, organizations store, process, and analyze a large amount of data in the hopes of extracting meaningful information. Social media and Internet companies are collecting a tremendous amount of personal data for more targeted marketing. Various companies are collecting a considerable amount of data on their own products, about their customers, about their competitors, and about their industries. However, many organizations are still struggling with how to best use and decipher their data in a business context (Walker 2016; Hardy 2015; Ferguson 2013a; Agarwal and Weill 2012; Kiron et al. 2011). Therefore, it has become increasingly important to find a way to translate overwhelming amounts of data into useful insights for a given organization. Using today’s enhanced computational power and advanced statistical methods, it is now possible to see data from different angles and identify novel relationships. Such actions create unique opportunities to extract knowledge and derive wisdom from various departments within an organization. Analytics has made it possible for various organizations in different industries to produce meaningful, innovative results.

In today’s business environment, decision-making is progressively becoming more complex. Managers need to make decisions faster, and their dependency on other systems for crucial pieces of information has made a noticeable impact on their decision-making processes. Complex environments, financial uncertainty, and market volatility all further fuel the need for
fast and dynamic decision-making (Kiron et al. 2011). It is undeniable that decreasing profits and reduced customer loyalty has become a big concern for retailers.

The growing availability of large amounts of high-quality data has drawn the attention of business executives to the data-oriented approach in order to make better decisions and find concrete support for their recommendations. Analytics creates an opportunity for organizations to improve their finances, marketing approach, day-to-day operations, supply-chain management, customer relations, and asset management (Evans 2015). Thus, organizations have started collecting financial-, employee-, project-, and operational-level data in hopes of minimizing their costs and optimizing their business processes (Laursen and Thorlund 2010). This in turn has led to a growing interest in analytics.

Companies applying analytics have embraced fact-based decision-making in order to optimize their business processes and improve their market strategies. Using analytics enables them to target specific markets, build the right culture, and encourage data-driven decisions. Companies that are sensitive to analytics will often consider financial and nonfinancial factors to understand their overall performance. They tend to reduce their inventory and likelihood of stock-outs in order to optimize their supply chains (Davenport 2006). Data-driven decision-making helps such companies make informed decisions by minimizing their risks, improving their strategies, and helping them find new insights (Manyika et al. 2011). Organizational uncertainty can be managed through effective information sharing (Cao et al. 2015a). To gain a competitive advantage in a given industry, there is a need for innovation in analytics since it is strongly correlated with successful information management within a company (Coleman et al. 2016). The construction industry is facing numerous challenges today because of its unique nature. Therefore, business processes within various construction organizations should be redesigned to ensure more effective information processing so that decisions can be made based on high-quality data. The growing use of analytics can help organizations make well-informed choices, assess their performance, identify relevant trends, optimize business processes, and better target potential customers. Analytics should be used to improve an organization’s objectives and business metrics.

That said, data and analytics are continuing to transform various industries. They are being used to solve a variety of problems, including challenges with finding new markets and services, increasing efficiency, reducing risk, driving innovation, and identifying opportunities for growth.
Unfortunately, despite having abundant amounts of data available, the construction industry has remained the industry most resistant to implementing analytics (Manyika et al. 2011).

It must be acknowledged that analytics is about a lot more than simply having data and crunching numbers. It goes beyond the data itself by highlighting how people can act intelligently based on core insights obtained from the data. Kiron et al. (2011) defined three key competencies to enable competitive advantages and gain benefits from analytics: information management, analytics skills and tools, and data-oriented culture. Effective analytics implementation also includes

1. modeling tools to define and categorize standardized work,
2. data integrating for better coordination among different entities, and
3. the implementation of activity-monitoring tools to measure performance and inform decision makers in case of bias.

Business Intelligence, a part of analytics, can be defined as a set of tools, systems, applications, and processes that enable the use of analytics through data, technology, and analysis (Davenport and Harris 2007). Meanwhile, informatics is defined as the collection of various data from multiple sources and the processing of data sets (Lim et al. 2015). Turk (2006) defines informatics as the “representation, processing, and communication of information in natural and artificial systems” (p. 188). Building on this definition, Hersh (2009) says informatics is the “discipline focused on the acquisition, storage, and the use of information in a specific setting or domain” (p. 02).

Thus, business intelligence can be described as the summary of historical data into tables and graphs (Maisel and Cokins 2015). On the other hand, analytics is the art of transforming data into useful information and actionable insights. Seddon et al. (2016) define business analytics as the “use of data to make sounder, more evidence-based business decisions” (p. 01). This differs from business intelligence, which uses information technology (IT) based tools such as data warehouses, online analytical processing, statistical tools, quantitative tools, and visualization tools to support business analytics.

Business intelligence is useful because it can answer simple questions using queries (Maisel and Cokins 2013). However, analytics provides the ability to create questions and thus can
answer more complex questions than business intelligence. Therefore, analytics continues to capture the interest of organizations looking to solve their complex decision-making issues.

2.2 Definition

With its growing popularity, it is unsurprising that researchers and businesses have defined analytics in many different ways. Kiron et al. (2011) say analytics is “the use of data and related insights developed through applied analytics disciplines such as statistical, contextual, quantitative, predictive, cognitive, and other modeling to drive fact-based planning, decision, management, execution, measurement, and learning. Also, analytics may be descriptive, predictive, and prescriptive” (p. 03).

Chen et al. (2012) define analytics as the “techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its market and business, in addition to making timely business decisions” (p. 1166). Analytics is the practice of developing actionable insights from data. It is the process of translating real-life problems into statistical questions and making decisions based on the data available. Because there are so many factors involved, analytics can be described as the intersection of processes, people, and technology in order to get insights from data (Wixom et al. 2013).

Over the years, different industries have adopted analytics to solve complex problems with the help of increasing amounts of accessible data, stronger computational power, and advanced algorithms. The tangible benefit of analytics appears when it improves organizational business processes. It should be clarified that analytics is not just the application of mathematical models and techniques but rather a process of transforming data into actionable insights that allow leaders to make context-based decisions. It is a systematic way of thinking that enables decision makers to identify compelling pieces of knowledge and create value for their customers in order to beat their competitors. Value creation can

1. Make a customer or client willing to pay more or
2. transform or standardize an activity, service, or product.

Delving further into the topic, analytics is also the “process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data” (Cooper 2012a, p. 7). Actionable insights are ones that lead
to practical action instead of simply reporting results or providing theoretical explanations. The actionable insights derived from analytics may also result in a distinct set of actions. They are a unique characteristic of analytics unlike decision support systems, reporting tools, or other traditional practices. Many people have the misconception that analytics requires big data, data warehouses, expensive IT structures, and complicated predictive models. However, analytics is actually about “using information as a strategic resource” to solve business problems or a gain competitive advantage in the market (Laursen and Thorlund 2016). It is an iterative, value-creating process that can be applied at any phase of a project’s life cycle.

Yet another definition of analytics says it is “the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to derive decision and actions” (Davenport 2007, p. 07). Analytics is unique because it represents the integration of three different fields of study: information science/business intelligence, statistics, and quantitative methods (Evans 2015). These disciplines, coupled with various computer tools, can not only help with data visualization and comprehension but also help people make intelligent decisions. Various definitions of analytics have been proposed by numerous researchers in the past. However, to maintain consistency, analytics can be comprehensively summarized as the art and science of converting raw data into information using statistical and mathematical models to turn that information into knowledge by asking the right questions in order to gain valuable insights.

2.3 Driving Factors of Analytics

Liberatore and Luo (2010) identified data, processes, people, and software as the four key drivers of analytics as shown in Figure 2-1.
Big data can be characterized by five features: volume, variety, velocity, veracity, and value (Packiam and Prakash 2016; Arora and Malik 2015; Jin et al. 2015; Libes et al. 2015; Sharda et al. 2013). Computer technologies are improving day by day, so smaller and faster devices are constantly appearing in the market. Owning a computer is no longer a privilege, which means that the amount of data available for analysis is growing rapidly because of technological advancements. The ability to monitor a diverse set of activities has been made possible by the developments made in microprocessors and the reduced cost of sensors. Thus, organizations have started gathering an extensive amount of data. Apart from structured operational data, these organizations are also collecting less structured data from various sources such as e-mails, documents, social media, and websites (Watson 2011). Social media and the Internet are generating a massive amount of unstructured data. Data is regularly created in multiple formats including text data, image data, sensor data, and structured data, which all represent the data variety characteristic of the big data. Such data provides the opportunity for timely feedback and customer opinions from a diverse set of populations. Since technology is advancing at a rapid speed, data is also flowing into organizations at a rapid pace and often in real time, which describes the velocity characteristic of big data.
Efficiently storing and accessing any collected data has also become an important consideration. Open-source software frameworks such as Hadoop, which makes it possible to distribute the storage and processing of large data sets across clusters of computers, have made it easier to handle data in ways that traditional databases could not (MacNeill 2012). Data is collected multiple times in a given project and comes in different formats from various stakeholders, therefore creating more uncertainty concerning the data. As the amount of available data increases with time, its quality and accuracy diminish; this describes the veracity feature of big data. Even though the volume of data being collected can overwhelm companies, there is a great deal of useful information hidden in such data that can be used to create opportunities for the extraction of valuable insights. Those companies that embrace big data often use it to gain actionable insights and create value for their customers. This new movement has led to the more recently established scientific research field of business analytics. The real power of data lies in its ability to create actionable insights and optimize business processes. But, apart from data, it is important to understand the other drivers of the analytics movement as they were defined in Figure 2-1.

The flood of new software vendors has redefined traditional analytics. The new software on the market is more sophisticated and interactive than its predecessors, and this innovation is one factor driving decision-makers’ interest. Subscription-based models have made it possible to implement analytics even at on a small scale since they do not need a massive infrastructure investment. User friendliness and the diversity of software types, ranging from reporting analytics to advanced analytics, has made it possible to gain insights from complex data sets. Lavalle et al. (2011) state that data visualization is one of the trending areas in today’s data analytics market because it simplifies the process of drawing complex insights. This visualization generally takes the form of dashboards and scorecards. Indeed, the dashboards can be used in any decision-making process, from providing real-time monitoring of an activity to a snapshot of an entire project. Data visualization is important for the following reasons:

1. It enables humans to think about and absorb information more quickly.
2. It enables users to interact with the tool and ask different business questions.
3. It does not need sophisticated analysis.
4. It enables data analysts to communicate their results to a general audience.

Once decision makers find useful insights, they can use them to optimize business processes, which in turn helps them better serve and create value for their customers. There are
many ways to optimize processes, including utilizing enterprises resource planning (ERP) systems, Six Sigma programs, and balanced scorecards (Liberatore and Luo 2010). Recent technological developments have increased the use of ERP systems because they allow the integration of various departments and result in the improvement of business processes and automation of required tasks within an organization. ERP systems enable organizations to collect and store all their data on a single platform. Because of ERP systems, many companies have started monitoring their key performance indicators (KPIs) in order to assess their day-to-day activities or business operations.

Many insights obtained through analytics are turned into actions implemented by decision makers. Analytics is about making fact-based decisions using data; however, it needs technically literate executives to understand the information. Simply highlighting different insights from the data is not enough. The real power of analytics lies in the creation of actionable insights, emphasis of value to customers, and optimization of business processes. Therefore, it is important for decision makers to ask the questions that will allow them to find the root cause of a problem and any potential solutions. Ultimately, the goal of any analytics project is to create value for the customers. Thus, even though a lot of effort is required to effectively implement analytics in an organization, it is a worthwhile endeavor. In summary, organizations that successfully adopt analytics need the following:

1. top management commitment,
2. continuous focus on improvement,
3. data-driven or fact-based decision-making processes,
4. employee participation or an open organizational culture, and
5. fast response cycles.

### 2.4 Types of Analytics

In today’s digital era, managers are facing many challenges in their decision-making because of the increased amount of data available for analysis. Such leaders are constantly asking questions about what has happened, why something happened, what is happening, what will happen, and what actions should be taken. These questions have resulted in four different types of analytics. The rest of this section explores them in greater detail.
2.4.1 Descriptive analytics

This type of analysis uses historical data to answer questions about what has happened or what is happening. Descriptive analytics is the summary or collation of all the data sets inside an organization in the form of a report, dashboard, alert, or trend. Data visualization is a useful tool for descriptive analytics. For the most part, both descriptive analytics and data visualization are considered parts of business intelligence (Ouahilal et al. 2016). Jugulum (2014) states that basic profiling, descriptive statistics, and data mining tools should be used for descriptive analytics. In addition, descriptive analytics can be likened to a post-mortem on data (Shankararaman and Gottipati 2015). Hayashi (2014) believes that it is the easiest form of analysis, but sometimes it can be difficult depending on the business question being asked. This type of analytics helps in understanding a business process at any given point of time in the past. It is usually presented in the form of reports and facilitate user understanding about the context of a problem (Hodeghatta and Nayak 2017; Liu et al. 2008). Using historical data, descriptive analytics provides information about topics such as what is happening on an organizational level (Ouahilal et al. 2016; Hayashi 2014). Mean, median, mode, range, quantiles, standard deviation, variance, and a summary of data are the general parameters of interest in descriptive analytics (Hodeghatta and Nayak 2017). Factor analysis, cluster analysis, and association analysis are considered advanced descriptive analytics techniques (Leventhal 2010).

2.4.2 Diagnostic analytics

Diagnostic analytics facilitates the reasoning behind the results obtained from descriptive analytics. It answers questions concerning why something has happened or is happening. Diagnostic analytics is a more difficult process than descriptive analytics. It seeks to understand the root cause of a problem or a change in a business process through the use of visualization tools, content analytics, and natural language processing applied to documents, e-mails, the Web, and social media (Ouahilal et al. 2016). Jugulum (2014) states that correlation analysis, hypothesis testing, analysis of variance testing, and control charts are some other possible analytical techniques that can be applied for diagnostic analysis. Laursen and Tholrud (2016) assert the importance of KPIs in diagnostic analytics because they can show a correlation between different activities and help leaders understand the effect of a process on those activities.
2.4.3 Predictive analytics

This is the most popular type of analytics in the business analytics market today. It answers questions about what will happen using mathematical models built on historical data (Shankararaman and Gottipati 2015). Predictive analytics is about moving to the next level and getting ahead of competitors by understanding what will happen or what is likely to happen in the future. It involves statistical techniques and data mining techniques such as regression and classification algorithms (Ouahilal et al. 2016). Predictive analytics involves modeling, regression, and simulation techniques (Jugulum 2014; Leventhal 2010). Simulations are helpful because they allow the development of future scenarios using what-if analysis. As its name suggests, predictive analytics helps leaders make predictions and understand future outcomes.

2.4.4 Prescriptive analytics

This is a complicated type of analysis that focuses on questions about what a business can do to make something happen, what a business should do to make something happen, or what the optimal answer to a given problem is. The final output of prescriptive analytics is a set of recommendations or actions that can be derived from predictive analytics (Shankararaman and Gottipati 2015). Ouahilal et al. (2016) suggest that for prescriptive analytics, we should look at current trends to forecast into the future and then use this information to get the best answer possible. Prescriptive analytics aims to optimize business processes or improve the performance of a particular system. Analytical tools such as scenario planning, designed experiments, and simulation analysis should be used for prescriptive analytics (Jugulum 2014). This is the hardest type of analytics and not only involves other types of analytics but also calls on the knowledge and decision-making power of an analyst to fully understand a given problem.
3. ANALYTICS IN THE CONSTRUCTION INDUSTRY

3.1 Current Status of Analytics in the Construction Industry

Sensor-based mobile devices can feature integrated RFID, bar codes, and radio tags. Hence, they have opened a new stream of Internet of Things for analytics. Internet-enabled mobile devices can collect location-based, person-centered, and context-relevant data. As a result, new opportunities especially in construction and facilities management are arising. For example, construction equipment manufacturers can embed sensors to collect real-time data on items such as usage patterns, product usage information, and the geography of a site. This data can be helpful in predicting future product development and providing a better estimation of productivity. By using this data, we can predict when a piece of equipment will no longer be functional and can prevent productivity loss by replacing parts in a timely manner (Manyika et al. 2011). As a result, Internet-enabled mobile devices and sensor-based technologies have created brand new opportunities for the construction sector to gain insights and make discoveries using highly contextualized and detailed data.

A construction project consists of a series of numerous complex interrelated activities and will be affected by design changes or changes in the overall project’s scope (Yang and Shen 2015; Jarkas 2013). The dynamic nature of construction projects has always been a problem for managers and decision makers because the fast pace at which construction process change often leads to poor decision-making. This in turn can lead to project delays and issues with going over the proposed budget. The risks generated by poor decision-making processes affect the quality, cost, time, budget, and performance of a given project. If these risks are not properly managed, they could eventually cause the failure of a project (Abderisak and Lindahl 2015; Zhao et al. 2015; Zeynalorian et al. 2013). Due to high client expectations, the increasing complexity of projects, the unique nature of projects, and rapid changes in technology, it has become difficult for managers and executives to quickly make rational decisions. Management’s inability to make sound choices is an internal organizational risk. Therefore, a faster decision-making process is necessary for managers in the construction industry if they intend to thrive and see their organizations succeed. The main problem that must be fixed to achieve this outcome is the inefficient retrieval of existing

The slow process of decision-making regularly causes substantial budget overruns and delay in construction projects (Martínez-Rojas et al. 2015; Kulemeka et al. 2015; Westin and Päivärinta 2011; Azhar 2005). At present, many decision-making processes in the industry occur manually, which limits analytical capability to what-if analysis (Behzadan et al. 2015). During the risk-handling process, a decision maker will apply his subjective judgment along with business constraints and market conditions to the analysis (Yoon et al. 2014). But the subjective nature of decision-making means that a systematic bias is sometimes introduced into the process (Ahiaga-Dagbui and Smith 2014a). Still, companies today are dependent on experts and subjective decision-making in various project management tasks instead of data-driven decision-making. Despite having access to various solutions and intelligent system for decision-making, managers are still struggling with how they can best make use of the tools on hand (Irani and Kamal 2014).

Currently, there is a large amount of information available to support the decision-making process; however, the construction industry lacks a systematic means of identifying and putting useful knowledge to work. Most historical data from past projects is not used to inform best practices in future work (Schwindt and Zimmermann 2015; Elazouni and Salem 2011; Wu and Soibelman 2006). Project managers have a large amount of data and information available to them, but construction industry data is often not integrated in a manner that makes it easy to manipulate. Instead, such data is stored in various silos and formats, which creates a problem for decision makers looking to quickly arrive at sound conclusions (Zhao et al. 2015; Azhar et al. 2014; Ghassemi and Becerik-Gerber 2011; Azhar 2005). The construction industry needs the latest advancements in decision-making tools. Without a systematic information structure and visual explanations for completed data analysis, it is difficult to make informed decisions (Songer 2010).

In the past, researchers have focused on finding the optimal solutions for specific problems without establishing any general knowledge. In the interest of making balanced decisions, both analytics and established knowledge are necessary because sound decision-making requires the integration of data and information from various sources (Blanco et al. 2017; Akhavian and Behzadan 2015; Karan et al. 2015; El-Diraby 2012). What would be a sound conclusion in a given situation changes over time because of the inherent uncertainties in construction processes. Competitive pressure and limited resources also represent barriers to making well-informed
decisions. Furthermore, construction projects typically involve multiple stakeholders; hence, any decisions made by one party are likely to affect other parties (Ren et al. 2013). It should be noted that decisions in the construction industry should be robust rather than optimal (Irani and Kamal 2014). Robustness is a key factor because it encourages solutions that are flexible and involve questions-based thinking and input from decision makers. Therefore, there is a need for a standard methodological approach that yields an effective and consistent decision-making process in order to solve complex problems and enable managers to make robust decisions. It is essential to understand exiting challenges in the construction industry before developing this standard approach because these elements are the ones hindering analytics implementation. The rest of this chapter presents key components of analytics and aspects of it as an art. The remainder of Chapter 3 also identifies the key reasons that have kept the construction industry from implementing existing analytics formulations and methodologies.

3.2 The State of Analytics Practice in the Construction Industry

The construction industry has maintained its focus on critical parameters and construction processes over time. However, prevalent problems such as low productivity, budget overrun, and project delay still exist as well. In the past, researchers have worked on finding optimal solutions to solve specific problems as shown in Table 3-1.

Table 3-1: The State of Analytics Practice in the Construction Industry

<table>
<thead>
<tr>
<th>Authors</th>
<th>Focus</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmadu et al. (2015)</td>
<td>Models are dependent on the</td>
<td>Using multivariate models, it is possible to predict construction time</td>
</tr>
<tr>
<td></td>
<td>characteristics of the region,</td>
<td>considering both project scope factors and non-project scope factors.</td>
</tr>
<tr>
<td></td>
<td>country, or construction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>industry</td>
<td></td>
</tr>
<tr>
<td>Patel and Jha (2015)</td>
<td>Prediction of safe work behavior</td>
<td>An artificial neural network approach was used to predict safe work behavior</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and efficient management of employees.</td>
</tr>
<tr>
<td>Author(s) (Year)</td>
<td>Description</td>
<td>Methodology/Findings</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Shehu et al. (2015)</td>
<td>Malaysian construction projects are experiencing schedule overruns</td>
<td>A regression model was used to predict project duration based on contract period.</td>
</tr>
<tr>
<td>Ahiaga-Dagbui and Smith (2014)</td>
<td>Causes of cost overrun and dynamics between cognitive dispositions, learning, and estimation</td>
<td>Using data mining and artificial neural networks, a cost prediction model was built to address cost overruns in the construction industry.</td>
</tr>
<tr>
<td>Tsehayae and Fayek (2014)</td>
<td>Whether there is a gap between experts’ knowledge and data-driven analysis of parameters and labor productivity field data</td>
<td>Using the RELIEF feature selection algorithm, identified a major discrepancy between expert perspectives and data-driven results.</td>
</tr>
<tr>
<td>Chaovalitwongse et al. (2012)</td>
<td>Lowest bidder can cause significant increase in cost because of change orders</td>
<td>Neural network models were used in project owners’ decision-making during the final bid selection.</td>
</tr>
<tr>
<td>Al-Tmeemy et al. (2011)</td>
<td>Inadequate to judge a project’s success based on cost, time, and quality</td>
<td>Project success is a multidimensional concept. Using principal component analysis, it is possible to measure project success by considering short-term goals and the company’s long-term financial objectives.</td>
</tr>
<tr>
<td>Xu et al. (2011)</td>
<td>Accurate measurement of construction cost index (CCI) with other variables in the short and long run</td>
<td>A co-integrated vector autoregression model was used to forecast a more accurate CCI.</td>
</tr>
<tr>
<td>Ashuri et al. (2010)</td>
<td>Short-term variations in the CCI are problematic in the preparation of accurate bids</td>
<td>Several time series can be used to accurately forecast the CCI by incorporating predicted price variations in bids.</td>
</tr>
</tbody>
</table>
There is no generalized approach to solving problems in the construction industry using data analytics. So how can companies maximize their opportunities while minimizing their risks? Organizations should look beyond the market boundaries and industry structure by taking data-driven actions and applying the analytics-related beliefs of key players in the industry. This might also lead to a new and uncontested market space that nobody has previously identified. It is important to understand the basis of analytics and the opportunities created through its application. However, the construction industry is still facing various challenges in finding generalized knowledge about analytics and understanding the value of generated data. The proposed protocol creates this foundational knowledge to facilitate data-driven decision-making in the construction industry. To show the systematic implementation of the protocol, two case studies are presented in Chapter 7.
4. THE STATE OF ANALYTICS PRACTICE IN DIFFERENT INDUSTRIES

Over the years, many organizations have considered their business requirements in the context of data analytics in order to optimize their business processes and create value for their customers. They have used data to understand the root cause of a problem, streamline operations, and solve their business problems. Having an understanding of different data sources, such as external, internal, and new sources, is essential to extract the most value out of analytics (Marr 2016; Franks 2014; Sharda et al. 2013; Barton and Court 2012; Schweitzer 2004). Today, organizations in various industries are continuing to speed up their decision-making processes, manage enterprise risks, and improve their understanding of customers using sophisticated analytics. Meanwhile, the construction industry remains far behind in implementing analytics compared to other industries because construction organizations do not understand how analytics can create value for their clients (Manyika et al. 2011). It is important to think about how information can be utilized to facilitate data-driven decision-making in the construction industry since other industries are already benefitting from this practice. In order to implement analytics in the construction industry, it is necessary to look at other industries and study how they are using data and information in their decision-making processes. Chapter 4 provides a brief review of analytics implementation in other industries. There are multiple data analytics domains, such as marketing analytics, operations analytics, customer analytics, service analytics, human resources analytics, supply chain analytics, risk analytics, and financial analytics (Holsapple et al. 2014). These domains are applicable in different industries, and other domains may exist based on a given industry’s unique characteristics.

Understanding how organizations outside the construction industry are using analytics to create value and achieve measurable results can help shed light on the implementation of analytics in construction companies. Analytics is used in a variety of tasks including financial forecasting, supply chain optimization, process and operations streamlining, location optimization, and annual budget allocation. But analytics affects much more than cost cutting or revenue generation. Indeed, it is also used in customer acquisition, business strategy creation, and human resources (Kiron et al. 2011). More importantly, analytics is used to create value for clients and customers. Figure 4-1 shows how Fortune 1000 companies are adding more value to their services or products using analytics.
We can see from Figure 4-1 that companies use data and analytics not only for cost-cutting purposes but also for innovation, increasing profit, and setting up a data-driven culture. Some companies use analytics to minimize their operational expense while others use analytics to understand customer demand. Various industries may use analytics for different purposes. For example, many industries use analytics to understand product demand in the market. The construction industry is a project-based industry, which means that construction projects are executed based on predetermined designs and are driven by clients. Therefore, using analytics to understand product demand would not create value for a construction organization. Still, analytics has created promising results for various industries, and establishing a data-driven culture has been a major challenge for numerous organizations. Before moving further, it is important to assess how other industries are using analytics and identify the unique problems the construction industry is currently facing.

### 4.1 Finance and Banking

Global financial crises have pushed financial companies to improve risk management and support continual innovation. Prices, returns, and trading volumes are the characteristics of uncertainty in the financial market (Flood 2009; Koopman and Székely 2009). Financial
companies have steadily been shifting their focus to applying analytics and asking fundamental questions about a client’s business requirements, business model changes, and organizational capabilities. Moreover, financial companies are trying to better manage risk and deal with regulatory changes. They are applying analytics to find new sources of returns and investments, identify new asset classes, and expand their markets (Ferguson 2014a). The finance industry is keen to find new insights about investment processes in order to lower their costs. Time-series modeling and other forecasting techniques are currently being used to make stock market predictions (Ouahilal et al. 2016).

In the finance industry, there are generally two types of data: time series, which shows the value of different market variables over time, and cross sectional data, which explains the state of multiple market factors and companies at a specific time (Pachamanova and Fabozzi 2014). Data visualization, forecasting models, and optimization tools are generally used in the industry to understand different markets, allocate assets, and define investment strategies.

Outside of the finance industry but still on the topic of finances, analytics is being used to better manage the finances of a company because actual performance can be measured by financial metrics instead of relying on intuition and market trends. Finance was the first area where other industries adopted analytics. Various companies use it to manage their financial forecasting and annual budget allocations (Kiron et al. 2011). For example, utility companies focus on financial management to increase revenue assurance and minimize avoidable debt write-offs. Meanwhile, in the health care industry, companies can predict the associated financial implications of certain risks by understanding the behavior and unique characteristics of an individual patient. Finally, retailers are increasing their competitive advantage by applying analytics to their financial management and budgeting (Lavalle et al. 2011).

### 4.2 E-commerce and Retail

Analytics is used extensively in the e-commerce and retail industry. Prior to the implementation of analytics, capturing data on customer behavior was very difficult. Then the introduction of the Web, social networking sites, and e-commerce businesses made it possible to gather large amounts of data on customer behavior. People in sales and marketing can collect such data quickly through a digital medium, and analytics has made it possible to gain useful insights about customers. Today, e-commerce and retail companies spend billions of dollars every year to secure repeat orders from
their customers and promote customer loyalty. In addition, customer satisfaction has become a widely used metric in assessing customer loyalty since satisfied customers bring more business to a company (Keiningham et al. 2014). Using cluster analysis, it is possible to find untapped customer segments. This information can then be used to develop new products or adjust a company’s pricing strategy, thus giving the company a competitive advantage over its rivals (Liberatore and Luo 2010). The e-commerce industry employs data visualization, text analytics, and social media to better understand their customers and gain useful insights (Lavalle et al. 2011).

However, it is only possible to collect a huge amount of data on sales and marketing in the e-commerce and retail industry because relevant businesses have large numbers of customers. Unfortunately, because the construction industry has fewer customers and clients, it can be difficult to collect data on them. The construction industry is not a consumer-centric industry but a project-based one. Therefore, best practices in analytics implementation in the sales and marketing departments of e-commerce and retail companies will likely be difficult to apply to the construction industry.

### 4.3 Manufacturing and Supply Chain

The manufacturing industry has already streamlined its operations, but productivity is not the biggest issue in the industry. In the slow-growth era and uncertain environment, many manufacturers are still not putting their data to use. Still, analytics would be especially beneficial in predictive maintenance, increasing yield, and optimizing supply chains (Dilda et al. 2017).

Today, the manufacturing industry has shifted its attention to minimizing inventory. Companies have adopted strategies such as product standardization to minimize uncertainty and cater to the needs of their customers at the local level. This is helping manufacturing companies to streamline their processes while maintaining a competitive edge in the market. It should be noted that there are three main sources of uncertainty in a manufacturing supply chain cycle.

1. Delivery of incoming materials: The known capacity of a plant and the limited number of stakeholders in the manufacturing industry make it possible to estimate and anticipate the delivery of incoming materials. In contrast, construction operations are dynamic in nature, require coordination among multiple parties, and are executed in a limited space. Moreover, construction projects are executed based on the design of a given project. It takes a considerable amount of time to construct or assemble any one
part of the project so that it meets the design requirements, and storing materials in advance at a construction project site is usually difficult. Therefore, this creates uncertainty in material delivery during a construction project and can lead to large amounts of wasted materials and inefficient supply chains (Marks 2017).

2. Internal manufacturing processes: The manufacturing industry has adopted a standardized product policy in order to streamline internal manufacturing processes. Unfortunately, the unique nature of construction projects does not allow for standardized products unless the construction project design is being based on the theory of intelligent planning units recently proposed by Hastak and Koo (2016). Because of the unique features of a given location and client requirements, implementing standard design on buildings remains a trying challenge. Therefore, off-site construction on a mass level is not possible in the construction industry.

3. Demand: Product demand is difficult to estimate in the manufacturing industry because it depends on the outside world and various product characteristics. Due to the unique design of a given construction project, managers spend time estimating the quantity of materials that will be needed for the entire project because a shortage of materials can lead to significant delays in the construction project while having excess materials can lead to reduced profits. However, calculating labor productivity is a challenging task because construction operations often differ for two similar projects using similar craft styles. Unless a task is repetitive, it is difficult to estimate labor productivity for a construction project.

Any variation in the delivery of incoming material and an internal manufacturing process will affect the lead time of the manufacturing process. Meanwhile, a change in demand can lead to inventory buildup and back orders. Thus, the growing availability of data can help decision makers better handle these sources of uncertainty.

At the moment, manufacturing companies use short-term price discounts and trade promotions to manage demand. They focus on capacity, inventory, subcontracting, and backlogs to manage their supply. But construction projects are comparatively small in number, are unique, and are design-specific endeavors, so it is not feasible to implement a pricing and promotions strategy in the construction industry. However, it is possible to control the supply part of a
construction project and minimize the total cost of the project’s supply chain. Because design changes happen frequently in any construction project and at any project stage, it is not advisable to have excess inventory. This is not typically a problem in the manufacturing industry.

Technology, people management, process changes, and optimization are essential in a process-based industry. Therefore, the fusion of data analytics with human expertise is essential to transform the manufacturing industry (Dilda et al. 2017). Even though the construction industry is also considered a process-based industry, we argue that analytics concepts developed in the manufacturing industry cannot be readily implemented in the construction industry. It is important to understand the differences that exist between the two industries in order to better grasp our assertion.

In the manufacturing industry, most processes are machine driven while in the construction industry, most processes are people driven; the unique design of a specific construction project is what causes this difference. Also, in the construction industry, the constructability of a design plays a significant role. On the other hand, in the manufacturing industry, most activities are repetitive and carried out by machines, so it is relatively easier to control cost and obtain a more reliable forecast of project parameters. Because construction projects are dynamic and complex in nature, most construction-related activities are unique in a complex project, and people’s inherent biases can lead them to underestimate the project cost while overestimating its value or benefits (Ahiaga-Dagbui and Smith 2014a; Flyvbjerg et al. 2014). Therefore, it is critical to establish transparent or uniform procedures in the construction industry in order to keep everyone involved on the same page.

4.4 Health Care and Medical

As more electronic systems begin to collect data, the use of analytics will increase in the health care and medical industry. Currently, the health care industry is focusing on statistical methods and summary reports. However, in the future, there are good prospects of analytics being used to predict future trends and outcomes (Abusharekh et al. 2015).

There are various applications of analytics that exist in today’s health care industry. For example, health care insurance companies analyze health insurance claims by using analytics to detect fraud or errors. (Srinivasan and Arunasalam 2013). Pfizer uses analytics to its highest potential in its planning and execution processes. The company tracks how each sales
representative presents material and how such presentations are received by clients. The company also monitors the Web patterns of its physicians when they are writing prescription in order to adapt to a specific customer’s needs and maintain a competitive environment. Moreover, Pfizer uses tablet PCs and clickstreams to track its strategies (Krion and Shockley 2011). Meanwhile, hospitals use analytics to translate written notes to patient dashboards so that they can improve patient care and results (Harris et al. 2014).

Some of the Analytical techniques used in the health care industry are biostatistics and epidemiologic analysis, Monte Carlo and discrete-event simulations, casual modeling, Bayesian statistics, optimization modeling, social network analysis, and agent-based simulations (Ward et al. 2014; Lee et al. 2013; Apte et al. 2012). Because of the health care industry must deal with complex operations and multiple stakeholders, analytics development in the sector still needs to be systematically studied (Khanal et al. 2016). Health-care companies collect data from multiple departments, such as clinical, operations, and finance, which has led to a great variety in their data. Using data from numerous sources can lead to better medical treatment and new discoveries (Hiller 2016). The health care industry is different from other industries because its data can be generated from multiple entities, including hospitals, laboratories, and insurance companies. However, there remains a lack of actionable insights because researchers have mostly focused on generic big data analytics frameworks in the health care industry.

4.5 Airlines

The airline industry is already benefiting from using analytics in multiple fields. For example, American Airlines has increased its revenue and strengthened its reputation tremendously in the last few years through the use of analytics (Davenport 2006). Other airline companies are also taking advantage of analytics to optimize pricing and allocate resources using sophisticated mathematical models that help them efficiently make revenue management decisions (Liberatore and Luo 2010). Various airline companies are also using analytics in their operations to reroute customers in an optimal manner when unexpected disruptions occur (Franks 2014). By employing operational analytics, airline companies can help their customers save time and money when problems arise.

Kiron and Ferguson (2012) explored how Cathay Pacific Airways has been using data to guide its daily operations. Analytics is employed for tasks such as determining the fuel efficiency
of planes, managing flight crew schedules, improving customer experience, and luggage handling. The company has implemented analytics in three main areas.

1. Operational efficiency or performance improvement: Cathay Pacific uses operational analytics to determine the engine performance of planes in its fleet and to optimize shifts for its crew members.

2. Customer intimacy: The company uses analytics to provide high-quality services to its frequent fliers. In addition, clickstream data allows the airline to gain a better understanding of customers who are looking for flight information on the Web. Finally, the company also uses data to look for trends in complaints about topics such as meals, seats, and various services.

3. Innovation: Design thinking is being employed to encourage innovation. For example, the company is developing new bag tags that will be able to track individual bags in order to address the ongoing problem of lost and misplaced baggage.

Lastly, it should be noted that Cathay Pacific has also been applying advanced analytical methods by using what-if scenarios and blending data collected from different sources.

4.6 Sports

In recent years, sports analytics has gained a lot of popularity. The initial use of analytics was highlighted with the baseball movie *Moneyball: The Art of Winning an Unfair Game*. Analytics is mainly used in sports to monitor player performance and determine the overall performance of a team (Ferguson 2013b). It is also possible to use analytics to understand strategies for athlete injury prevention, team composition, and player profiles. It is even possible to capture a player’s every movement during a game using GPS and tracking applications. Light (2013) discusses how sequence analysis is a useful tool for understanding the pattern of injuries an athlete has sustained in different body parts. Davenport (2006) explored how Tony La Russa, a baseball analyst, used analytics in combination with his intuition to substitute high-energy players into a batting lineup. The US women’s Olympic cycling team members have improved their health and performance by relying on analytics, and this change arguably led to their winning the silver medal in the London 2012 Olympics (Marr 2016). In addition to improving player performance and better aligning players to the needs of a game, analytics also shows up in sports during talent evaluation, especially
when it comes to predicting the performance of an individual player (Brady et al. 2017, Davenport 2014).

4.7 Other Industries

Apart from the previously explained applications of analytics, Davenport (2006) has also looked at how various companies are taking advantage of analytics by using statistics and mathematical modeling in different business areas as shown in Figure 4-2. Moreover, analytics is being used in other fields for product pricing, crime detection, spam filters, increased organizational efficiency, econometrics, marketing, and fraud detection (Cooper 2012b).

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>DESCRIPTION</th>
<th>EXEMPLARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply chain</td>
<td>Simulate and optimize supply chain flows; reduce inventory and stock-outs.</td>
<td>Dell, Wal-Mart, Amazon</td>
</tr>
<tr>
<td>Customer selection, loyalty, and service</td>
<td>Identify customers with the greatest profit potential; increase likelihood that they will want the product or service offering; retain their loyalty.</td>
<td>Harrah’s, Capital One, Barclays</td>
</tr>
<tr>
<td>Pricing</td>
<td>Identify the price that will maximize yield, or profit.</td>
<td>Progressive, Marriott</td>
</tr>
<tr>
<td>Human capital</td>
<td>Select the best employees for particular tasks or jobs, at particular compensation levels.</td>
<td>New England Patriots, Oakland A’s, Boston Red Sox</td>
</tr>
<tr>
<td>Product and service quality</td>
<td>Detect quality problems early and minimize them.</td>
<td>Honda, Intel</td>
</tr>
<tr>
<td>Financial performance</td>
<td>Better understand the drivers of financial performance and the effects of nonfinancial factors.</td>
<td>MCI, Verizon</td>
</tr>
<tr>
<td>Research and development</td>
<td>Improve quality, efficacy, and, where applicable, safety of products and services.</td>
<td>Novartis, Amazon, Yahoo</td>
</tr>
</tbody>
</table>

*Figure 4-2. How some companies are using analytics in different business areas (Davenport 2006).*

Over the last few years, different industries have seen significant productivity growth, but the construction industry is still lagging behind in terms of productivity because of industry fragmentation. Creating a unified approach by working across management systems and gaining a better understanding of technical systems can change the mind-sets of people in the industry and reinvent construction processes (Barbosa et al. 2017).
5. FUTURE OF ANALYTICS IN THE CONSTRUCTION INDUSTRY

5.1 Data- and Information-Related Concerns in the Construction Industry

The construction industry remains focused on electronic data transfers and information management systems to facilitate communications (Ng et al. 2001). However, any analytics project or application requires the identification of useful data and combinations of data sets to achieve the desired results. Moreover, a proper data management plan and strategic clarity are needed before the initiation of any analytics project (Coleman et al. 2016). Both technical skills as well as business knowledge are required to effectively manage data sets.

Construction organizations rely on multiple sources of data, which causes the fragmentation of information. In addition, since the construction industry is a project-based one, its organizations need information at a fast pace in order to properly manage their projects. Typically, a construction project will include a significant amount of engineering work, thus creating an abundance of unique engineering data. Lin et al. (2007) assert that engineering data comes in different formats and has unique characteristics while business data is usually fixed in format. In the construction industry, there is a substantial risk that engineering data will be of poor quality because of low-quality engineering drawings and errors in design documentation (Westin and Sein 2014). Due to the presence of multiple stakeholders in a construction project, a good amount of data is collected from numerous sources in a variety of formats, which further increases the complexities associated with project management. Martínez-Rojas et al. (2015) point out that on-site production, the project durations, and differences in objectives are all causes of data fragmentation, lack of integration, and complex processes in the construction industry. This means that relevant organizations are continually adopting new information and communication technologies. However, they still lack robust data for their decision-making processes, and these data-related concerns are ultimately restricting the growth of analytics. For construction organizations, obtaining a clear view of the life cycle of data and relevant information remains a cumbersome task (Skibniewski and Ghosh 2009).

5.1.1 Challenge 1: Difficulty in data selection

In this digital era, the amount of data available is increasing rapidly because of technological advancements. Therefore, it has become more important than ever for construction organizations
to select the right pieces of data and put them to use in order to identify actionable insights (Zeid 2014). Moreover, Laursen and Thorlund (2016) argue that automatic reports are interactive and have the ability to drill down into more meaningful details for a company. However, in the construction industry, many reports are too complex in nature to have their creation fully automated.

Still, organizations can increase their competencies by creating value from existing data. Prioritizing value potential by selecting only useful data for analysis is necessary because including low-quality data can skew results. It must be remembered that managers are the individuals who define a strategy along with measurable outcomes to build a competitive advantage in the market, but different departments within a given organization are the ones who accomplish these goals. Thus, individual departments will define their needs based on their targets. Any measurable outcomes or benchmarking processes created by these departments will have to be constrained by time. So, it is fair to say that an entire data set represents the historic view of an organization, which is why it is not an efficient way of reporting. By selecting only the required data for monitoring purposes and market competency evaluation, an organization can conduct its business more efficiently. The multidimensional perspective of a given business helps to provide optimal insights and enables individuals to slice and dice data as needed.

By nature, a construction project involves multiple processes during its execution phase, and monitoring individual operations can improve the health of the project. Selecting the right data to analyze helps in monitoring any essential processes. Laursen and Thorlund (2016) state that a high-level idea concerning a project helps people better understand the project’s stakeholders and any required competencies. However, Stubbs (2013) points out that it is not always possible to sample data on a regular basis because any complete set of data will increase processing time and computational cost. Therefore, data selection is an important factor for large and complex projects in the construction industry.

It must be acknowledged that big data is not always good data since it does not reveal any existing bias in the data collection process and does not explain the data’s context (Boyd and Crawford 2011). Kaisler et al. (2013) say that finding the right type of data for statistical analysis is often difficult and that this problem can be solved by either reducing the size of the data or improving the search or decision-making process. It is important to ask robust questions instead
of making assumptions during the data collection or analysis process. Asking intelligent business questions can help in acquiring the right set of data.

5.1.2 Challenge 2: Data fragmentation

Most construction projects will have multiple stakeholders with different objectives. Data can be collected by various entities throughout the various phases of a project’s life cycle. In the construction industry, data is regularly collected from sources such as sensors and geographical information systems. However, integrating various data sources, analyzing the data, and extracting knowledge from it remains a big challenge (Becerik-Gerber and Siddiqui 2014). In today’s world, organizations are surrounded by information and data silos that cause a gap in information sharing between departments. Data collection and ultimate control of project information will often depend on which stakeholders in a given project are collecting the data (Craig and Sommerville 2006).

Due to poor coordination and the absence of a coherent objective, data collected by different entities can cause unnecessary processing problems (Ruddock 2002). Many construction organizations currently have multiple analytical efforts in place, which creates many redundant steps in the implementation of an analytical project instead of generating actionable insights. Dainty et al. (2010) discuss how the construction industry is a highly fragmented industry in particular because

1. seasonality creates fluctuating demand cycles,
2. product demand is highly dependent on projects,
3. on-site production causes uncertainties in the production itself,
4. construction projects involve a diverse range of skills within a small geographical area, and
5. project stakeholders take any knowledge learned away with them after a project’s completion.

However, Dainty et al. (2010) also mention that use of an integrated supply chain can minimize the existing fragmentation problem in the construction industry. At the moment, the majority of the work in the construction industry is done by small and medium-sized enterprises (SMEs), and each party collects its own data based on its particular interests, which means that it is challenging to integrate information at a project or organization level. In the past, the automobile
and electronics industries also had to tackle the issue of data fragmentation, and they were able to solve the problem by leading projects with a single team (Marks 2017). In the entertainment industry, Caesars Entertainment addressed the issue by standardizing its practices (Ferguson 2013c). Thus, in the construction industry, supply chain professionals should work to exchange materials, services, and information in a more open manner if they wish to solve the problem of data fragmentation. Due to the lack of integration in supply chains, construction projects today can create a lot of disparate pieces of information that make projects difficult to manage.

Skibniewski and Ghosh (2009) assert that the construction industry’s performance measurement issues can be rectified through the development of KPIs. Existing KPIs often require data generated from multiple sources since there is a lack of data integration and information accessibility. This is an undesirable situation because it can lead to reduced profitability. For example, product lifecycle management will generate heterogeneous data from the conceptualizing, design, building, service, and demolition stages. To apply business analytics in product lifecycle management, it is important to integrate data, processes, people, and systems (Rohleder et al. 2013). Laursen and Thorlund (2016) suggest that most complex construction projects fail because of an existing weak link or incompetent department inside a given organization. Therefore, the construction industry should reconsider the existing approach to data and link different departments together for the successful execution of a project. Although information in the construction industry is often collected from multiple sources such as external partners, questionnaires, and websites and then stored in multiple locations, researchers have already shown that it is possible to integrate the data silos of different departments by applying data-fusion techniques (Vasenev et al. 2014; Akhavian et al. 2013). Various data-fusion techniques and their potential applications have been previously described in the construction engineering field (Shahandashti et al. 2010). Therefore, data-fusion techniques can probably be used to solve the issue of fragmented information in the construction industry.

5.1.3 Challenge 3: Inconsistent and out-of-context information

Inconsistent information is a particularly challenging issue for the construction industry because different project stakeholders can have distinct perspectives and associated problems. It is often difficult to pin down a particular stakeholder’s interest, and without this information, decision makers are unlikely to come up with a fitting solution. In addition, a construction project generates
a lot of data that gets taken away by individual parties once the project is completed due to various contractual requirements (Ahiaga-Dagbui and Smith 2014b). Many decision makers are still trying to figure out how data from one project can be applied in future projects without increasing the lack of context for such information. As has been previously mentioned, because construction projects are unique in their designs, specifications, methods, and standards, finding a way to use the data from one project on another is quite a challenging task. The large amount of project information will likely involve several layers of uncertainties ranging from internal factors (e.g., project time, cost variation, unforeseen site conditions, and contractors’ claims) to external factors (e.g., weather and financial market stability) (Akhavian and Behzadan 2012).

The construction industry must still deal with the fact that output data is not consistent in form and there are no standard international definitions for data across the industry. No mechanism exists to facilitate the comparison of construction activity data across different nations, which means it is difficult to measure how economic development of one country against another. Ruddock (2002) says that the construction industry is not able to make use of data at the international level because of the following problems:

1. different regions use their own accounting procedures and statistical definitions, and
2. there are problems with coverage.

In the case of shortages in quality information, people will use their instincts and experiences to make a decision. However, arriving at conclusions based on imperfect information will often lead to bigger problems in the future. Therefore, it is important to come up with a standard reporting system not only for project monitoring but also for day-to-day decisions (Changali et al. 2015).

Data management itself is not a big problem in construction research (Turk 2007). But knowledge development and transfer is a major issue in the industry because of implementation and deployment problems. As it turns out, the lack of sound information integration and management is not a technical issue but an organizational one (Ahmad et al. 2010). Therefore, it is essential to combine analytics with organizational decision-making processes. It is not possible to achieve robust results without such an integration (Cao et al. 2015). Embedded analytics can help focus attention on right pieces of data, make it possible to achieve goals through analytics, and ease the process of making data-driven decisions. While most industries are already using
analytics, they are not obtaining the most useful insights from their data because of the existing lack of understanding concerning the value created by analytics (Henke et al. 2016). Jachimowicz (2017) proposes five steps to maximize the value of data analytics. Moreover, it is also possible to save money by implementing a structured approach to the process, and many organizations in different industries are currently taking this route (Repenning et al. 2017).

The idea of using the increasing amount of data available within organizations to better focus decision-making and action-oriented planning is still a new concept in various industries (MacNeill 2012). Decision makers who wish to effectively implement analytics should focus on building an analytical culture in their organizations. This type of culture can be developed by promoting a systematic approach to reasoning and bringing both decision makers and analysts to the same communication platform.

5.2 Challenges in Implementing Analytics

Despite the increasing popularity of analytics, managers are still facing difficulties in implementing it within their organizations. Collecting data or getting high-quality data is not the problem. Instead, it has been observed that within an organization, analytics barriers are cultural and managerial in form, rather than technological. Lavalle et al. (2011) state that organizations face various challenges when they focus on implementing data-driven decisions, and these challenges are shown in Figure 5-1. Although there are multiple hardships that can occur, Lavalle et al. (2011) emphasize that the managerial and cultural barriers are more serious than the data and technology barriers.
THE IMPEDIMENTS TO BECOMING MORE DATA DRIVEN

The adoption barriers organizations face most are managerial and cultural rather than related to data and technology.

- Lack of understanding of how to use analytics to improve the business
- Lack of management bandwidth due to competing priorities
- Lack of skills internally in the line of business
- Ability to get the data
- Existing culture does not encourage sharing information
- Ownership of data is unclear or governance is ineffective
- Lack of executive sponsorship
- Concerns with the data
- Perceived costs outweigh projected benefits
- No case for change
- Don’t know where to start

Figure 5-1. Challenges faced by organizations seeking to become more data-driven (Lavalle et al. 2011).

In today’s digital era, organizations are collecting large amounts of data, but the true value of data comes when it is used to gain insights. As mentioned earlier, many organizations use ERP software to improve their operations. However, these same organizations may still struggle with making decisions faster. Jin et al. (2015) argue that computational complexity, data complexity, and system complexity are major big data challenges. Meanwhile, Jamiy et al. (2014) cite data security, data integration, and the effective representation of data as some of the challenges in big data analytics.

The collection, efficient storage, and retrieval of data are critical facets of data-driven decision-making. This kind of decision-making is also dependent on the effective and efficient use
of stored data to obtain actionable insights. Some existing computational and visualization tools make it easier for us to extract information from data. The advancements that have been made with such type technological tools is one of the reasons why people have become more interested in analytics (MacNeill 2012). Kiron et al. (2011) define the analytical competency of an organization by its reliance on fact-based decision-making instead of judgment- or intuition-based decision-making. Part of what makes analytics compelling is the fact that it is not a static process. It is the art and science of converting raw data into actionable insights. Since technology is continuing to change rapidly and business organizations are becoming more interdependent, analytics can highlight the places where existing business processes can be improved.

Building a data-driven culture in an organization and breaking down the walls between departments is essential to the process of creating value through analytics (Nicolaus Henke et al. 2016). Organizations need employees who understand both technical and business language in order to bridge such gaps between departments. They should seek decision makers who are willing to ask robust questions and others who are willing to give key recommendations in order to resolve business problems.

In an ideal situation, analysts would work with decision makers or business leaders during the initial phase of the analytics project. This collaboration will help the analyst develop business acumen so that later on business problems can be solved and lessons learned can be turned into proprietary knowledge. However, it is often the case that an analyst will not be interested in solving business problems and making decisions, which means decisions makers may find it difficult to communicate with analysts. There exists a big interpretation gap between analysts and decision makers (Brady et al. 2017). Many organizations are trying to address this data interpretation and communication gap by teaching quantitative methods to domain experts. Such organizations seek to hire people who understand both the business and data worlds, and they may spend a significant amount of money to solve the problem. Therefore, there is a need for a common language to systematically bridge this gap. Embracing a systematic approach will make it possible for everyone in a given organization to understand the changes being made. But it must be noted that technology is not the answer to everything: Humans are an essential part of analytics too. Analytics can achieve its full potential when all employees in an organization are taking part in the process.
5.3 The State of Analytics as an Art

To reiterate, Lavalle et al. (2011) argue that for any organization, managerial and cultural barriers are more difficult to overcome than technological barriers. They have also recommended some strategies to overcome these challenges, and they are presented below (Lavalle et al. 2011, p. 25, 29).

1. “Focus on the biggest and highest value opportunities”: It is important to implement any analytics project from a strategic perspective in order to maximize business value creation.

2. “Within each opportunity, start with questions, not data”: Focusing only on data collection from the start of an analytics journey can divert attention away from data management issues in a harmful manner. Asking questions is a very important part of maximizing the impact of analytics. Moreover, the ability to understand the big picture of business objectives and select the most appropriate data set for the analysis is facilitated by a systematic procedure of asking questions.

3. “Build the parts, plan the whole”: In this digital world, the amount of data being created is rapidly increasing. Thus, it is important to get the right pieces of information at the right time. Organizations are facing various challenges when it comes to managing their data and using it effectively. Having integrated and consistent data is essential to build a solid foundation for information management. Based on a survey done by Lavalle et al. (2011) and shown in Figure 5-2, managers consistently look for certain data characteristics. Because managers deal with multiple stakeholders and entities, they know that each one has its own data and that it is important to understand these different data sets. Getting consistent data from multiple entities and then integrating it into one set is an important step to making robust and insightful decisions.
The application of analytics depends on the underlying nature of the problem being solved. There is no standard tool or analytical method that can resolve all of the problems an organization may have. Instead, before applying analytics, decision makers should clearly understand what type of information is required to yield the post implementation findings that are desired. Thinking about a problem from a multidimensional perspective is what will create value and result in meaningful information. Therefore, decision makers should be focusing on the art of asking questions. Because construction projects involve multiple disciplines, a variety of problems that are different nature can exist at any time in the construction process. Only by having a systematic procedure can these problems be solved. Technology alone cannot make sense of why a project is seeing poor productivity. Instead, a cultural change within the organization must also occur, and it should embrace standard procedures and acceptable systems. Hence, a protocol is needed to
drive construction business and facilitate a decision maker’s ability to generate high-quality questions.

Liberatore and Luo (2010) defined a four-step process view of analytics that is shown below in Figure 5-3. However, this visualization does not cover the importance of asking questions in an analytics project. Still, Liberatore and Luo (2010) did mention that analytics projects start with the collection, extraction, and manipulation of data. In contrast, we argue that analytics projects start from “understanding the problem” with regard to the needs of the business end users or decision makers.

![Figure 5-3. Four-step process view of analytics (Liberatore and Luo 2010).](image)

Asking questions, gaining industry insights, and understanding operational challenges are important considerations in any data analytics project (Kiron et al. 2013). Any business process inside a business value chain can generate analytical insights. For example, when a superintendent submits project data at the end of a day, analytical insights can be defined based on who is supposed to look at the data, what should be measured, and who should take what actions.

Jugulum (2014) describes a process for executing analytics and focused on collecting relevant data and ensuring high quality data, as shown in Figure 5-4. However, organizations in business today have already been collecting large amounts of data on their own. As such, there is a lack of understanding about how to use existing data and formulate useful problems for analysis.
At the moment, construction companies are gathering large amounts of data without asking questions. This has led to cumbersome data management processes in terms of data collecting, management, and cleaning. Data management by itself is a highly time-consuming process and forces managers to make dynamic decisions while they are waiting to gain fresh insights from analytics. In order to meet their business requirements, organizations should focus on defining their business goals and asking relevant questions. Focusing on asking a few key questions related to specific subject areas can create value quickly (Kiron et al. 2013). The amount of value derived from an analytics project depends on how a decision maker defines the problem, whether an analyst asks robust questions, how an analyst structures the analytical solution, and how end users implement the solution (Franks 2014). In the construction industry, project data is often stored in different databases, which creates non-validated and non-integrated data and complicates the decision-making process. Therefore, it is important to ask different questions about who, what, why, what if, and what next when seeking to solve a problem (Azhar 2005).

Information sharing is not being done optimally in the construction industry (Vidal and Möller 2007). Moreover, many construction organizations are missing an analytical culture for information sharing and data-driven decision-making. To overcome these challenges, it is
necessary to increase collaboration and improve communication between project stakeholders (Ren et al. 2013). A unified approach is required to facilitate the smooth flow of organizational data. At the moment, knowledge and information barriers such as a reliance on tacit knowledge and the lack of standardized approach are major barriers to improving any tasks in the construction industry (Martínez-Rojas et al. 2015). The good news is that having a standardized data-driven approach can facilitate data-driven decision-making in the industry. Wu and Soibelman (2006) stress that the industry’s analysis tools still lack general or explicit knowledge. However, many construction problems are already being resolved through systematic approaches that involve asking different questions. These examples can lead to the creation of standardized systems that will help project teams effectively makes decisions and easily identify any issues (Yang and Shen 2015; Jalaei and Jrade 2014). Hence, it is necessary to create processes that can extract existing contextual knowledge to support critical decisions.

Shankararaman and Gottipati (2015) developed the business process improvement framework that is shown in Figure 5-5. In addition, Seddon et al. (2016) argue that decision-making is a complex process, and Figure 5-6 presents their interpretation of the problem-solving process. But neither of the proposed approaches captures the decision-making process in detail nor do they describe how analysts ask questions. Therefore, there remains a need to create a new protocol that will help both decision makers and analysts implement analytics in construction organizations.
Figure 5-5. Business processes improvement framework using analytics (Shankararaman and Gottipati 2015).

Figure 5-6. The human problem-solving process model (Seddon et al. 2016).
Analytics facilitates an organization’s ability to gain actionable insights from data. Even though the term *actionable insights* is frequently used in analytics, many researchers do not address how they can be created (Cao et al. 2015; Barton and Court 2012; Kiron and Ferguson 2012). While organizations want to use analytics to obtain information based on data, they do not have a comprehensive approach to generating insights. Lim et al. (2015) state that the manufacturing industry is a product-based industry and that recently such industries have started adopting the competitive strategies used by service-based industries. But there exists a third type of industry, the project-based industry, and the construction industry is the best example of it. In the construction industry, organizations are already collecting large amounts of data using innovative technologies such as drones, sensors, and laser scanning. However, because construction organizations lack an analytical culture, they spend a significant amount of money on technology and digital transformation without being able to effectively implement analytics. Therefore, it is important to understand the challenges the industry is facing in order to take analytics implementation to the next step. Furthermore, building a systematic approach to enable data-driven decision-making through analytics is an ideal solution since the construction projects are unique by nature.

### 5.4 Need for Analytics in the Construction Industry

Analytics can reduce cost and create value for an organization (Feldmann et al. 2017). However, gaining the ability to deploy it multiple times across various departments is a big challenge. Most organizations have the required domain expertise and IT capabilities, but their organizational cultures and capabilities are often lacking. Therefore, there is a need for a standard approach to facilitate decision-making and analytical thinking throughout a given organization. Anyone in an organization can create value using a systematic and structured approach, thus generating an organizational culture that embraces data analytics. Analytics can create value when a problem has been defined and robust questions are asked (Franks 2014). Having a standard process for analytics implementation in a process-based industry will make it easier to change the mind-sets of employees and make them realize the importance of analytics (Dilda et al. 2017). However, cultural changes will also need to be made to improve productivity and remove silos because technology alone cannot solve the industry’s existing problems. Over the years, there has been a demographic shift in the industry that has resulted in larger populations of migrant laborers and
older managers (Barbosa et al. 2017). While implementing technological solutions may not be a problem for construction organizations, the scalability of any solution is a notable issue since the project location and involvement of multiple stakeholders will put constraints on a solution (Agarwal et al. 2016). Developing a uniform and standard information process to capture more insights in an organization is critical for long-term success (Zeid 2014).

Due to multiple internal data silos that vary by geography, asset class, and business units, construction organizations have information stored in multiple locations and have failed to properly maintain their resources (Blanco et al. 2016). To make matters worse, the head office and project office often use any existing systems in a different manner, which further exacerbates the problem. Even though other industries have found their own solutions to this issue, these solutions do not meet the construction industry’s requirements because no standard construction operations currently exist (Blanco et al. 2016; Skibniewski and Ghosh 2009). The construction industry differs from other industries because construction projects have multiple players who take part with different priorities in mind, and each aspect of construction work is tracked on the task level.

In today’s digital era, the construction industry is considered a traditional industry because many organizations still rely on pen and paper. The lack of focus on analytics and low investment in IT are two factors hindering the industry’s development (Agarwal et al. 2016). Moreover, the lack of coordination between stakeholders who have different objectives means that there is often no consistent view for a construction project.

In addition to operating in a more traditional manner, the construction industry is a risk-averse industry, and construction projects are often executed based on lump-sum contracts with very low profit margins. As a result, contractors tend to apply only tried-and-true techniques, and executives hesitate to implement modern technologies. Construction companies have not yet been able to take full advantage of today’s technologies because of the high initial investment cost and lack of clarity concerning their returns on investment. Therefore, the industry is not innovating as quickly as many other industries.

Although the era of analytics seems certain, the plausibility of analytics implementation in construction is still uncertain due to a lack of understanding about its potential applications and the scarcity of young talent in the industry. However, many researchers in the past have already shown the importance of analytics in the construction industry (Hafiz et al. 2015; Ko et al. 2015; Nik Bakht and El-diraby 2015; Shrestha 2013). But no defined path exists for decision makers in
the industry who wish to make well-informed choices based on data. Hence, there is a need for an analytical movement that can integrate the use of data, processes, and systems while offering a clear pathway to obtain consistent results.

Today, the construction industry is facing many challenges with implementing analytics, and these problems can be caused by complex information, fragmented data, or inconsistent information (Blanco et al. 2017). Decision makers struggle to use analytics in their organizations because they lack a well-defined and repeatable process that can facilitate data-driven decision making (Zeid 2014). A systematic approach with a unique shared mission can reduce the information silos inside a given organization. This structured approach can not only address information silos but also attack departmental and organizational silos (Swann 2017). Turk (2006) argues that construction informatics can transfer information technology to the construction sector. However, there has been no systematic study that enables decision makers to derive value from data using analytics. Hence, a protocol should be defined to facilitate the development of actionable insights and ease an organization’s ability to make decisions based on sound data.

It is important to understand that while other industries are customer-centric, the construction industry is project-centric in focus. Even though other industries can understand customer behavior by analyzing a large amount of personalized and customer-specific data, the construction industry cannot follow the same approach because of its unique nature. The number of customers or construction project owners in the construction industry is smaller than the number in other industries, so it is hard to collect data on people’s behavior. Therefore, the analytical frameworks or decision-making models that have been developed in other industries will not be overwhelmingly helpful to the construction industry. Instead, a new protocol is required to facilitate structured decision-making in construction processes.
6. DEVELOPMENT OF A PROTOCOL TO FACILITATE DATA-DRIVEN DECISION-MAKING IN THE CONSTRUCTION INDUSTRY

6.1 The Protocol

To implement analytics in an organization, decision makers should focus on achieving the greatest competitive advantage and growing the financial success of the organization (Davenport 2006). The primary focus of an analyst who is implementing this protocol should be on solving a problem, not on applying advanced analytics methods. Simple problems should be solved using simple analysis rather than advanced analytics.

To design the road map of a successful analytics project, it is essential to have a fixed structure on hand along with information and a business strategy (Laursen and Thorlund 2016). Therefore, we have taken steps to describe how analytics can be implemented and used in the construction industry. First, it should be noted that this protocol consists of the six parts shown in Figure 6-1 to facilitate data-driven decision-making:

1. Conceptualization,
2. Design,
3. Development,
4. Refinement,
5. Analyses, and
6. Outcome.
Figure 6-1. Summary of data analytics protocol for the construction industry.
6.1.1 Conceptualization: Describe the current business situation

Analysts should work closely with decision makers or business subject matter experts during the conceptualization stage of the analytics project because these people are the individuals who are aware of what is happening inside and outside of the business (White 2013). Later, analysts can work closely with data engineers in the development and modeling stage to ensure the use good-quality data and that the final data for the modeling or analysis solves the business problem. The conceptualization stage is important because it helps in understanding the problem itself. This state also allows analysts to acquire insights from decision makers. It is important for analysts to communicate well with decision makers as they collect information and gain knowledge from the decision makers. Sometimes, analysts are more excited about applying advanced analytics, so they will forget that the goal is to solve a business problem. They should understand that the key to connecting with decision makers is to ask questions. First and foremost, analysts should focus on getting in-depth background knowledge on the problems an organization is facing. The various tasks for this stage are summarized in Figure 6-2.

**Figure 6-2. Phase 1: Conceptualization- Describe the current business situation.**

### 6.1.1.1 Problem Statement:

Analysts should understand that decision makers might not have the time to learn about analytics techniques. These individuals are more interested in getting business problems solved. Thus, it is easier to monitor and measure the various activities that are occurring with the increased amount of data and technological improvements available. But simply using computers and technology will not result in new insights from data. It is an analyst who must understand the problem at hand, find a solution, and improve business processes. To transform an organization, analysts should start by addressing the problems that can be solved and take advantage of the opportunities that can be created using analytics (Dilda et al. 2017). However, if a problem is poorly framed, then it
will be difficult to obtain robust solutions and rectify any mistakes made (Franks 2014). Identifying an organization’s main problem is an important task. A decision maker should evaluate current organizational capabilities and identify the organization’s critical business areas before defining the problem and working to have it solved using analytics. For example, managers can look at a company’s KPIs while defining the problem statement. The exercise presented in Figure 6-2 can help decision makers identify the most challenging business problems.

If a user focuses on a single problem during the conceptualization phase, the cognitive and perceptual load will be reduced (Green et al. 2008). Having a narrow focus will help in getting tangible results through an analytics project (Zeid 2014). Unsurprisingly, it is easy to define the problem statement when the business problem is a simple one. Once the business problem becomes complex, defining and understanding the problem statement becomes more important because it will help determine the reasoning about and analysis of the problem. The problem statement plays the key role of identifying the business problem and understanding its causes. Moreover, the problem statement helps employees grasp the impact the problem is having on the business.

The application of analytics depends on the underlying nature of the problem statement (Arora and Malik 2015). There is no one-size-fits-all solution for implementing analytics. Understanding the business problem should be done before jumping into any advanced analysis so that an analyst can apply knowledge in the correct context. Business analytics is useful when it solves a business problem and creates value for customers or clients.

Managers should take the time to understand the strategy fit and underlying nature of the problem at hand in order to define the problem statement for an analytics project. However, it must be noted that not every solution can create value. Therefore, problems with the highest potential to create value should be given priority. Here are some questions that decision makers can ask when assessing the urgency of a problem:

1. How do we stand with respect to our competitors?
2. Where is the market going?
3. What kind of competencies do we need to have for a certain market condition?

6.1.1.2 Business Implications:

In order to determine business implications, analysts should understand the organization’s primary objective, its defined business criteria, business goals, and how the business operates as a whole.
Analytics can dramatically transform an organization when it is integrated into business processes (Dilda et al. 2017). The business implications of selecting a new project in the construction industry can be identified by evaluating how well the new project aligns with the existing business model, assessing the required competency for the project delivery, determining whether the project will be successful, and understanding the utilization of existing know-how in the company (Pekuri et al. 2015). Business implications seek to understand a business before and after analytics implementation. Once obtained, business implications provide a bigger picture of the company and scope of the analytics project. Ben et al. (2016) gave an example using energy and resources efficiency to show that improvement in this area can help in reducing cost, in enhancing the company image, and in strategic business development.

The business implications of reducing the amount of time spent on a project should address the opportunity cost and potential gains from time savings (Butcher and Sheehan 2010). Because the construction industry is a low-margin industry, it is important to follow a low-cost (competitive) differentiation strategy. To achieve a competitive advantage in the construction industry, organizations should focus on acquiring powerful clients while maintaining intense rivalries in order to create a strong barrier for new entrants in specific markets or locations (Ho 2016). Understanding the capital structure (comprised of debt and equity) of an organization in the construction industry facilitates a better understanding of a company’s business opportunities and limitations (Yee et al. 2006). Meanwhile, business implications considered by the executives tend to be more focused on long-term planning and strategy (Ferguson 2014a). Barbosa et al. (2017) believe that issues with productivity and on-site execution in the construction industry can be solved by using a common set of KPIs. However, in today’s environment, different organizations will likely have their own areas of emphasis in their metrics, so understanding the business implications of a given problem is essential before starting any analytics project.

6.1.1.3 Current Solutions:

Davenport (2006) says that in order to gain a competitive advantage, organizations should spend their resources on maximizing the development and benefits of existing systems. Learning about existing solutions and gathering fact-based information helps in the creation of systems thinking and application of analytics. Understanding existing analytical solutions can also help foster innovative solutions in addition to improving old ones. This approach can even help with
evaluating whether the problem is due to technology, data, process, or company culture. Beyond evaluation, taking the time to comprehend current solutions can aid in the recognition of useful content and characteristics that already exist inside the organization. This is the path that will lead to applicable knowledge about the construction industry even if it still lacks integrated platforms (Agarwal et al. 2016). In fact, the industry has no standardized electronic document management system; instead, companies have their own specific systems (Kahkonen and Rannisto 2015). Understanding existing solutions will shed light on how best they can be integrated with information and communications technology systems and tools. Furthermore, these solution should be simplified to be intuitive in nature so that they do not require significant investment on existing platforms. Franks (2014) says that there are three ways to create value through analytics:

1. by solving an entirely different problem,
2. by solving existing problems using innovative methods, or
3. by adding more value to existing processes.

Thus, taking time to analyze business processes helps to solve existing problems in an innovative way, and knowledge of existing analytical solutions can be used to develop an understanding of the limitations and infeasibilities of alternative solutions.

6.1.1.4 Entities Involved:
The business stakeholders and end users of the analytics project comprise the entities involved. Defining business stakeholders and end users helps identify the analytics project requirements (Stubbs 2013; Hughes et al. 2012). In the construction industry, the design of any work process depends on the quality and amount of information required (Robert et al. 2006; Azhar 2005). Organizations have a continuous flow of information coming from their interdependent subunits. This flow creates complexities and uncertainties in the decision-making process because information must be integrated from diverse sources. And since this information is not consistent and does not consist of high-quality data, it is essential to find the root causes of a problem and ask questions about who owns the data, its sources, and who can affect data quality inside the organization. This process sheds light on the needs of various entities and their work practices. For example, because the equipment department for a construction company will share information on equipment usage cost with project managers, getting the information at the right time is a critical
factor if the project manager is to accurately monitor the progress of the project. Concurrently, this information is important for the equipment department itself when employees are deciding whether to buy a new piece of equipment, rent it from within the company, or rent it from another organization. By selecting the cheapest option, the equipment department helps the project manager keep the project on track. On the other hand, information from the project manager about usage hours, weather conditions, ground conditions, and work conditions can aid the equipment department in financing the cheapest option.

Laursen and Thorlund (2016) assert that good data quality is important in any analytics project. It is essential to start with the data source for validation and understand how it has been used by various entities. Defining entities in any analytics project can shed light on data quality problems. Using this newfound knowledge, an analyst can give data management recommendations. For example, the multiple vendors on a given construction project will deliver their specific goods and materials to the project site. In this scenario, it is the field engineer or superintendent who receives the delivery receipts and then passes them on to the accounts payable department. However, the accounts payable department will also receive invoices from many other vendors. One task of the accounts payable department is to create a routing slip for delivery receipts obtained from the project manager so that the proper amount can be issued to the vendor. During this process, multiple e-mails and calls are exchanged among different people, and there is a high possibility that the company may not be capturing all available data concerning the transactions or that the collected data will not be in the proper format for analysis. To solve this type of communication problem at the organization level, Caesars Entertainment identified its stakeholders and provided them with a transparent platform where they could easily see how their operations were running and how the current situation might affect their future operations (Ferguson 2013c). The construction industry has similar situations where multiple stakeholders or entities are involved in a single project. To get high-quality data, it is critical to streamline processes by understanding different entities.

6.1.1.5 Define the Scope of the Data Analytics Project:

A construction project involves several operations during its life cycle. Therefore, applying performance management principles is a useful way to guide project operations. The goal or scope of the analytics project in any organization should be decided based on the overall business strategy
(i.e., vision, mission, and objective) because the business strategy will be tied to some KPIs that can help with business evaluation (Laursen and Thorlund 2016). Defining end user requirements is also important during an analytics project because they can help in developing a reliable analytical solution (Kay 2013). A single problem will often be related to other problems and can affect multiple managers since the world is more connected today than it was in the past. Therefore, analyzing the business process chain and defining the scope of the analytics project will keep the task manageable (Liberatore and Luo 2010). The scope of an analytics project can be defined based on following points:

1. What problem should be solved?
2. How will the analytics project solve your problem, and how will it help you in decision-making?
3. Once the project is over, what actions will be taken by the end users?

Having an “analytics vision” is essential to define the scope of an analytics project (Jugulum 2014). The aim of this step is to determine the information needs of the end users (e.g., a report, dashboard, predictive model, or recommendations for optimization). The scope of the analytics project should be decided at the project level as well as the company level, thus resulting in KPIs that can be used to measure the project’s progress and success. Laursen and Thorlund (2016) say that the scope of any analytics project should meet all of the following conditions:

1. Specific: Decision makers should define specific targets to achieve their strategy and gain a competitive advantage in the market. For example, revenue in the next year should be increased by a certain percentage, a company should win a specific number of new projects, or the accident rate should be decreased by a certain percentage.
2. Measurable: If it is not possible to measure or monitor the target, then either the outcome required from the analytics project or the scope of the project need to be changed.
3. Agreed: To maintain ownership of the analytics project, it is important that decision makers and executives agree on its scope and any required changes to business processes.
4. Realistic: Often, decision makers do not think about whether the target is achievable. They should strive to have a realistic strategy and a scope for the project in place; otherwise, there will be no output from analytics.

5. Time-bound: Finally, the amount of time required to implement the strategy should be defined along with the tentative deadline for the completion of the analytics project. These considerations will help the analyst fully grasp the nature of the project and suggest any changes or make corrections as per the timeline.

6.1.1.6 Data Available:
There are many analytics tools on the market. Having a grasp on data availability and business requirements aids in the selection of the right analytical tool. Today, utilizing information resources remains a difficult task in the construction industry (Ma et al. 2011). Proper storage, identification, review, and representation of information is essential for the efficient analysis of an organization’s information systems. This includes the process of listing the data available from enterprise systems that might be providing information during the analytics project to help create value for the organization. Laursen and Thorlund (2016) state that managers or decision makers should evaluate each data source by asking two questions:

1. How useful might the data be?
2. How accessible is the data?

6.1.2 Design: Understand the problem
The design stage should be closely linked to the conceptualization stage and development stage because it helps ensure that people working on the analytics project have a common goal and can facilitate a mutual understanding of high-quality data. This is the stage that requires the technical knowledge, domain knowledge, and business acumen of an analyst. Armed with domain experience and technical skills, the analyst must be ready to think critically. Interdisciplinary knowledge is also important because it further boosts critical thinking and learning from other industries. A key aspect of the design stage is breaking down the problem and putting together an analytical solution for it. The analyst should strive to remain open-minded and pay attention to details in the design stage in order to solve problems intelligently. This stage requires the analyst to apply experiential knowledge to design open-ended questions and evaluate alternative solutions.
The analyst should be cautious about making assumptions and should verify that the designed problem solves the business problem and helps in the decision-making process.

It is possible for an analyst to focus on complex analysis or optimizing models without understanding the bigger picture. However, any insights generated from these analyses will not be useful if the analyst is unable to communicate them properly. Complex analyses and models are useless without the application and value addition parts of analytics. In the design phase of analytics, an analyst should focus on gaining an in-depth understanding of the analytics project and should define the flow of the project as laid out in Figure 6-3.

Figure 6-3. Phase 2: Design- Understand the problem.

6.1.2.1 Exploratory Analysis:

Exploratory analysis is a branch of analytics looks at data to find value. In this digital era, companies are collecting data to monitor and measure their business processes. This data provides a tremendous opportunity to gain insights. During the exploratory analysis phase, an analyst plays with the data while keeping in mind a broad scope of the analytics project. Thus, the analyst can better understand and design the structure of the problem in a concrete manner. In the exploratory analysis phase, an analyst can look for interesting patterns, evaluate correlations, and formulate the hypothesis of the problem. During this stage, the analyst should take some time to explore existing data sources and analytical methodologies in order to formulate a solid picture of the solution (Franks 2014). The purpose of data visualization is to find interesting nuggets of information and then communicate them to other people. Data visualization is the art of connecting novel information in a new way in order to examine data from multiple perspectives. Although exploratory analysis is a time-consuming process, an analyst can use it to identify a continuous flow of information (Zahalka and Worrying 2014). Dainty et al. (2003) say that exploratory analysis is important in the construction industry because construction projects are unique and do not come with much prior knowledge. Exploratory data analysis can help reveal the characteristics of data
and uncover hidden pieces of information along with a potential hypothesis (Jackson 2002). It can even help an analyst find inconsistencies and errors in the data.

Exploratory analysis allows an analyst to evaluate any requirements for information extraction and any data needs for the later analysis. It is the stage of any analytics project during which the analyst makes himself aware of the problem’s context in order to better understand the problem in detail. At this point, an analyst will work with data using visuals or summaries to identify patterns and understand the data’s characteristics. This stage is valuable for the analyst because it acts as a link between the business context and analytics. Moreover, the exploratory analysis helps an analyst become a more open-minded person by encouraging the analyst to ask robust questions from the data and understand the value of analytics in the decision-making process. Using this type of analysis, the analyst can identify patterns within the data and validate assumptions that may be useful when structuring the problem and working on model selection.

Data visualization is a main component of exploratory analytics and involves the visualization of data, model results, and analyses, which can yield astonishing relationships among different variables and may even provide an easy medium through which results can be shared with a general audience in the organization (Evans 2015). This is a key step for people who are interested in exploring the data on hand to see what it may contain. Data visualization is important for any data analytics project because it provides a framework for solving problems that considers topics such as what data should be used for the analysis and what features appear in the study. Apart from these points, data visualization is also important for the reasons described below:

1. It helps analysts understand what is going on during exploratory analysis, can explain many things, and can answer numerous questions in a period.

2. It introduces new questions during the analysis, which helps analysts better understand the problem at hand.

3. Analysts can visualize data in multiple ways and find hidden patterns in the data using data visualizations tools. This makes it possible to pair the most appropriate analysis with the data set.

4. Data visualization is the best way to communicate results to an audience or project stakeholders (Evans 2015).

5. Visualization techniques also improve the validation of system quality (Solomon 1995).
Data visualization is not only important while exploring data but also when sharing results with the public. It is typically not an easy task to address questions through hypothesis testing, and data visualization has become an essential component of any analytics project to demonstrate how certain questions could be answered. Before applying analytics, it is important to have a robust understanding of the problem’s context, which includes considerations such as who the decision maker is, which stakeholders are involved, and what decision makers need to know. Various data visualization tools such as a dashboard or scorecards are used to monitor key business processes. In addition, business leaders use these tools to measure and monitor industry trends, measure performance at various levels of an organization, and understand competitor behavior (Evans 2015).

6.1.2.2 Problem Structure:

In this stage, an analyst defines what information is necessary to understand the behavior of a problem as explained by business managers. A lot of attention has been paid to data collection and fact-based decision-making in recent years. Researchers in the construction industry have developed many theories to build general knowledge and understand the relationships of dynamic systems. We believe these theories may play a vital role in implementing analytics, structuring the problem strategically, and understanding the root cause of the problem. While data analysis can easily answer the question of what is happening, business experience or a theoretical knowledge can address why the problem is happening and how it can be readily resolved. Having strong theoretical knowledge can help an analyst ask questions from multiple perspectives, which is good since numbers alone cannot provide the entire picture to make a robust strategy. It is important to understand the underlying context and meanings behind the data (Kirkland and Wagner 2017). An analyst can look for multiple viewpoints and widen the picture of the analytics project by correctly understanding and framing the problem (Maisel and Cokins 2015). Therefore, the analyst should intertwine data with learned knowledge and theoretical concepts. To move ahead in this stage, an analyst should look for relevant research and see how other researchers have explained a given relationship. The analyst should be able to translate the business problem into an analytical problem and must be able to specify the type of information required to solve the business problem. If an analyst is not able to ask the appropriate questions, then all other efforts will be of no use (Laursen and Thorlund 2016). This is the information developing strategy step and is a crucial
stage of any analytics project. Here, an analyst specifies the information requirement and gains an understanding of what information is required to solve the business problem based on the business strategy.

6.1.2.3 Analyses Overview:

The idea behind this stage is to evaluate what type of analysis is necessary for the end user and whether it will answer the questions defined earlier in the conceptualization stage. To leverage data and be competent during the final analysis of the analytics project, it is essential to think through what kind of analysis would be most suitable per the available data. This step is required to analyze the data needs. Through this step, the analyst can review the available data within an organization and reduce redundant data collections. The analyst can also define the information requirement for the last stage. Thus, an analyst should focus on how this information will be gathered or how to obtain some actionable insights. The analyst should think about possible analytical methods, the objective of analysis, the results to be achieved, and possible outcomes while designing an analytical solution.

6.1.3 Development: Understanding the data

The fundamental key to understanding data is to recognize the quality of data (Stubbs 2013). One attribute of good analysts is that they never assume data is clean or that they have complete data. To keep consistency, an analyst can follow the proposed approach in this stage of the protocol, which also assists in comprehending, exploring, and collecting data. Many organizations are not following the right approach to data collection, and as a result, they are collecting all sorts of data, which leads to collection of unnecessary data. Meanwhile, a good understanding of the data collection process can yield a robust strategy (Kirkland and Wagner 2017). Organizations should not collect all types of data; rather, they should identify what data makes the biggest difference to them and what data is useful for actionable insights (Marr 2016). After Phase 3 of the protocol, an analyst should have a good understanding of any organizational data silos. The majority of the time spent on any analytics project will go into data preparation and cleaning. At the end of Phase 3, an analyst should have an outline of the analytics project similar to what is shown in Figure 6-4.
6.1.3.1 Data Quality:

In today’s digital world, organizations are collecting large amounts of data. However, companies do not know whether their data collection methods are reliable and valid (Jachimowicz 2017). Thus, before implementing analytics, it is essential to check the quality of the collected data; however, in many organizations, no one has been assigned to handle data quality (Abai et al. 2015). But it is unquestionable that maintaining good data quality is important for the successful implantation of an analytics project because the outcome will affect the organization as a whole. Jin et al. (2015) assert that in this digital era, big data is providing us with a wide variety of data, which has caused the emergence of data quality problems such as null values and duplicate entries. Ransbotham (2015) says that predictive and prescriptive analytics requires good data quality and goes on to suggest that creating shared platforms can solve the problem of data duplication. Quality and security during the data collection process should be given utmost importance in for the interest of effective decision-making (Caya and Bourdon 2016). Data quality is a dominant issue in the health care industry, and it has been suggested that good data processing is required to solve the problem (Cho et al. 2015).

The construction industry is also facing many challenges when it comes to maintaining good data quality. There is a large gap between the information needed and information available because of the industry’s traditional data handling methods (Mervi and Veli 2002). Agarwal and Weill (2012) say that standardizing and digitizing processes are the best way to obtain high-quality data. However, the construction industry must still address its information and data quality problems, particularly with regard to missing or inaccurate data (Dainty et al. 2001). Therefore, it is important to check data quality before starting any analytics project. Accuracy, completeness, and consistency are important parameters that can be used to measure and assess data quality.
(Cooper 2012; Westin and Holen 2012; Batini et al. 2009; Falorsi et al. 2003), and these parameters can be measured according to the following considerations.

1. Completeness: For data to be deemed complete, the number of null values should be minimized.
2. Consistency: The number of consistent values should be high, and values violating any constraints should be minimized for data to be considered consistent.
3. Data accuracy: The number of correct values should be measured.

### 6.1.3.2 Data Structure Needed:

In many organizations, the collected data is not reliable and may not be in the desired format to solve business questions. It is quintessential to identify the required data structures needed to solve the business problem. This stage calls for an analyst’s critical thinking about topics such as what data is useful and what the structure of data should be to solve the business problem. An analyst should have a quality conversation with decision makers, end users, and other stakeholders to ensure that the collected information is sufficient to solve the business problem. This stage helps in the creation of an efficient data structure, and the analyst should follow the systematic approach proposed below:

1. **Data variables needed:** In this stage, an analyst should ask the question, what variables and attributes are necessary in the analysis? For example, an analyst could come up with the following data structure:
   - Variable1 = \[a1\]
   - Variable2 = \[b1, b2, b3\]
   - Variable3 = \[c1, c2, c3\]
   - Variable4 = \[d1, d2\]

   Once an analyst precisely answers this question, a more efficient implementation of analytics is possible. This arrangement of data is called a data structure.

2. **Data variables available:** In this stage, an analyst should write down all the available data sources that were suggested by decision makers in the conceptualization stage. For example, three different variables with their available data attributes are shown below:
   - Variable1 = \[a1, a2, a3\]
Variable2 = [b1, b2, b3]
Variable3 = [c1, c2, c3]

3. Data strings needed: Finally, this is the process of identification where an analyst finds what data strings are not present in the available data structure and still need to be collected. Based on the needs of the additional data variables and available data variables, an analyst can figure out what kind of data is missing. For example, in this case, there is a need to collect data for Variable4.

6.1.3.3 Data Sources:

From the data structure stage, it is clear that Variable4 is not present in the available data sets. Therefore, an analyst should look for different options to collect Variable4 data and may consider some of the choices below:

1. Internal data sources: Decision makers
2. External data sources: Public sources, other departments inside the organization, and other companies
3. Data that needs to be collected

A key point to note is that the time required to receive data from internal sources and public sources tends to be minimal. However, if data is not available internally or externally, then its collection may take a significant amount of time. Therefore, it is important to avoid any additional data collection and use proxies or proxy variables in the absence of a required variable whenever possible. Proxy variables are not directly relevant but will have a good positive or negative correlation with the required variable.

6.1.3.4 Data Collection:

This is a critical step of the overall analytics project where the focus should be on identifying the limitations of the data after the sources of the required data have been identified. In this step, an analyst should be able to communicate data collection constraints and data limitations to decision makers before moving on to the analysis stage because imperfect information can lead to biased decision-making. Data communication is also important because it allows decision makers to
integrate the assumptions made by an analyst into their decision-making processes (Ferguson 2014b).

6.1.3.5 Supplementary Information: This step is about communicating concerns about the data analytics project to decision makers. In this step, an analyst can gather additional sources of information from decision makers in order to better address the business question. The analyst should try to collect additional information that was not available earlier during the conceptualization phase. The special effort made to get more information is also important because it will affect the integration of technological solutions inside the organization. Once the problem has been designed and the data has been understood, some additional information may be required to provide more data sources and strengthen understanding system integration. For example, KPIs represent the specific business processes of a company and also enable the identification of previously unconsidered applications (Skibniewski and Ghosh 2009). Before moving to the data cleaning step, it is wise to get more information because data cleaning is typically the most time-consuming task of an analytics project (de Jonge and van der Loo 2013; Kalashnikov and Mehrotra 2006; Raman and Hellerstein 2001). Finally, putting in the extra effort to collect more information at this stage can save time later in the process.

Project information flows are in complex nature because of the considerable number of events that occur. Therefore, it is necessary to make a few assumptions about the exact address and order of interaction in order to enable a resilient information flow (Törmä et al. 2015). But sometimes, an analyst will make assumptions that could be harmful to the project. Because organizations have started collecting huge amounts of data, which helps in extracting insights about a problem, it is essential to reconsider any assumptions made during an early phase of analytics such as data collection (Marr 2016). Moreover, it should be noted that once an analyst understands the problem, no further assumptions should be made about the issue. Instead, if any doubts arise, the analyst should seek more information from decision makers.

6.1.3.6 Data Cleaning: Before doing any analysis, it is important to perform data cleaning and data integration so that the data will be in the proper format (Raman and Hellerstein 2001; Spaccapietra and Maryanski 1998). Removing data discrepancies and normalizing the data are important tasks in data cleaning, which
can be described as the “process of detecting, diagnosing, and editing faulty data” (Van Den Broeck et al. 2005). During data cleaning, the following tasks defined by de Jonge and van der Loo (2013) should be completed:

1. putting headers for each column
2. checking data types (e.g., numbers should not be stored as strings), and
3. ensuring categories have been properly labelled.

In a given database, California can be written as California or Cal or CA. This is called data discrepancy, and such discrepancies often occur because data coding is being done by different people with no standardized procedure to follow. Data stewardship is necessary to resolve this problem (MIT Sloan Management Review 2016). Meanwhile, data transformation can be understood as the act of converting data from one scale to another, and there are many techniques for data transformation. However, converting any data set into a relative scale with respect to the mean is the most general approach (Andrienko and Andrienko 2013).

Any data analysis or modeling requires data cleaning because of the presence of inconsistent data such as missing values and duplicate entries. It must be remembered that data is created at various times, by different people, and using different conventions within an organization. This often leads to errors, missing values, and inconsistent values in the data. Although there have been many efforts made to adopt standard formats of data exchange in order to resolve this issue, merging heterogeneous sources of data to create a single database sometimes also creates data cleaning problems (Kalashnikov et al. 2005). To reiterate, this is typically the most time-consuming task of any data analytics project.

### 6.1.4 Refinement: Data refinement

In today’s business world, storing data has become much easier and cheaper. However, processing that data is still quite a cumbersome task. Ransbotham (2016) argues that data processing is an expensive and time-consuming process that takes about 80% of the total analytics project time. The characterization of challenging issues such as this one and the use of appropriate tools is required for effective project management in the construction industry. Big data in the industry is creating information overload for project managers because of the lack of information structures and visual explanations. Existing visual approaches lack information density, and lot of research
has focused on improving the information required for effective decision-making (Zahalka and Worring 2014). Therefore, there is a need for a systematic approach that can improve the presentation of information in order to increase organizational awareness about the business problem and make it easy for people to grasp it. Songer (2010) states that although the construction industry has diversified data, it also has a shortage of visual tools to communicate information.

The low speed at which information is processed is the main reason why large construction projects experience delays (Westin and Päivärinta 2011; Gyampoh-Vidogah et al. 2003). Reducing this time by taking a structured approach based on the business question being posed is a priority. Good data management practices and timely access to information can facilitate operations and decision-making processes. It is notable that construction and project managers are still facing difficulties in their decision-making processes despite having various intelligent systems (Irani and Kamal 2014). More recently, the introduction of data warehouses and online analytic processing tools in the industry has made it possible to find useful information (Martínez-Rojas et al. 2015). Slicing, dicing, splicing, and layering are not new tasks, and many researchers have worked on the topic in the past. Indeed, various definitions of slicing and dicing have been proposed by different researchers and are presented in Table 6-1. Slicing and dicing is a philosophy for analytics that allows an analyst to take only the data needed for analysis, as shown in Figure 6-5. This is an intelligent way to obtain relevant data while minimizing the loss of information. At the end of Phase 4, an analyst should get rid of any unnecessary data so that the organization can save time and money.

![Figure 6-5. Phase 4: Refinement- Data refinement.](image)

**6.1.4.1 Slicing and Dicing:**

Slicing and dicing is not a new philosophy or technique since it is used in online analytical processing. But slicing and dicing depends on the business questions asked by the decision makers. Researchers in the past have defined slicing and dicing, and this information is shown in Table 6-1.
<table>
<thead>
<tr>
<th>Authors</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aalst (2013)</td>
<td>Slicing: Pick a specific value for one of the dimensions. One dimension is then being removed. Dicing: Pick specific values for multiple dimensions. No dimensions are removed.</td>
</tr>
<tr>
<td>Kasinadh and Krishna (2008)</td>
<td>Slicing: Operations reduce the number of dimensions by taking a projection of data in the cube on a subset of dimensions. Dicing: Operation amounts to a range select condition on one dimension or to a selected condition on more than one dimension, where logical operators like “and”, “or”, and “not” connect conditions on different dimensions.</td>
</tr>
<tr>
<td>Hutchison and Mitchell (2001)</td>
<td>Slicing: Reducing the data cube by one or more dimensions. Dicing: Subselecting a smaller data cube and analyzing it from a distinct perspective.</td>
</tr>
<tr>
<td>Han (1997)</td>
<td>Slicing: Extraction from a data cube of summarized data for a given dimension value. Dicing: Extraction of a subcube or intersection of several slices of the data cube.</td>
</tr>
</tbody>
</table>

It is essential to acknowledge that slicing and dicing is not a sampling technique but a data refinement technique to answer business questions. On the other hand, sampling is used when there is a class imbalance problem in the data (i.e., one class dominates the other in a data set). In the era of the Internet of Things, organizations are collecting large amounts of data and are
subsequently finding it difficult to analyze this data in real time (Blanco et al. 2017; Vasenev et al. 2014). Therefore, selecting high-quality data is important for analytical purposes. The goal of slicing and dicing is to select the most useful features of data in order to increase analytics speed and power while simultaneously minimizing the loss of relevant information associated with the problem. It is the process of extracting useful data by understanding the context of the business problem. Without knowing this context, it is difficult to understand the outcomes of the analytics project, and any resulting decisions made based off of the findings may be biased (Ahiaga-Dagbui and Smith 2014b). If the required information is taking too much time to process, it may be classified as inaccessible and an analyst could be wasting his time. Saklani (2017) proposed some strategies to find actionable data and explained the importance of small data. Finally, slicing and dicing enables an analyst to prepare interactive reports and ask questions such as what, when, where, by whom, what project, what location, and for whom (Laursen and Thorlund 2016). Summarized definitions of slicing and dicing are presented below.

- **Slicing:** The selection of data within a meaningful range of one dimension or variable from single multidimensional data.
- **Dicing:** The selection of data within a meaningful range of more than one dimension or variable from single multidimensional data. In the case of dicing, there is no reduction in variables or dimensions.

<table>
<thead>
<tr>
<th>Project ID</th>
<th>Year</th>
<th>Project Delivery Type</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2005</td>
<td>Design Build</td>
<td>Midwest</td>
</tr>
<tr>
<td>B</td>
<td>2006</td>
<td>Design Bid Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>C</td>
<td>2006</td>
<td>Design Build Operate</td>
<td>Southeast</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maintain</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>2007</td>
<td>Design Bid Build</td>
<td>Midwest</td>
</tr>
<tr>
<td>E</td>
<td>2008</td>
<td>Design Build</td>
<td>Midwest</td>
</tr>
<tr>
<td>F</td>
<td>2009</td>
<td>Design Build</td>
<td>Midwest</td>
</tr>
<tr>
<td>G</td>
<td>2009</td>
<td>Design Bid Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>H</td>
<td>2009</td>
<td>Integrated Project Delivery</td>
<td>Southeast</td>
</tr>
<tr>
<td>I</td>
<td>2010</td>
<td>Design Build</td>
<td>Midwest</td>
</tr>
</tbody>
</table>

Table 6-2: Sample Projects of an Organization
For example, as shown in Table 6-2, we have 11 years’ worth of historical data spanning from 2005 to 2015 for all of the projects, and we know the project delivery types and regions for a fictional organization. The company is operating in 2 locations, Midwest and Southeast. If we are interested in only Southeast projects, then we slice Southeast location projects from Table 6-2. Finally, we have only Southeast projects as shown in Table 6-3 as shown below:

Table 6-3: Sample Projects of an Organization After Slicing

<table>
<thead>
<tr>
<th>Project ID</th>
<th>Year</th>
<th>Project Delivery Type</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>2006</td>
<td>Design Bid Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>C</td>
<td>2006</td>
<td>Design Build Operate Maintain</td>
<td>Southeast</td>
</tr>
<tr>
<td>G</td>
<td>2009</td>
<td>Design Bid Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>H</td>
<td>2009</td>
<td>Integrated Project Delivery</td>
<td>Southeast</td>
</tr>
<tr>
<td>J</td>
<td>2011</td>
<td>Design Bid Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>M</td>
<td>2012</td>
<td>Design Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>O</td>
<td>2013</td>
<td>Design Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>P</td>
<td>2014</td>
<td>Design Bid Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>Q</td>
<td>2015</td>
<td>Design Bid Build</td>
<td>Southeast</td>
</tr>
</tbody>
</table>

Using slicing, we select the data from one variable. Now, Table 6-3 shows data that is not dependent on location because all the projects noted are in in the Southeast. The final data set in Table 6-3 has the year and project delivery type as variables. For dicing, we are only interested in
projects that occurred after the year 2009 and have an integrated project delivery or design build delivery method. After slicing and dicing, our final data set is shown below in Table 6-4.

Table 6-4: Sample Projects of an Organization After Slicing and Dicing

<table>
<thead>
<tr>
<th>Project ID</th>
<th>Year</th>
<th>Project Delivery Type</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>2009</td>
<td>Integrated Project Delivery</td>
<td>Southeast</td>
</tr>
<tr>
<td>M</td>
<td>2012</td>
<td>Design Build</td>
<td>Southeast</td>
</tr>
<tr>
<td>O</td>
<td>2013</td>
<td>Design Build</td>
<td>Southeast</td>
</tr>
</tbody>
</table>

It is important to note that all the projects are located in the Southeast region because we already sliced the location component from our data set when we sliced it. Thus, both slicing and dicing are dependent on business questions, and when used together, they solve a unique data analytics challenge in the construction industry (Azhar 2005). Slicing and dicing addresses the issue of data diversity by systematically reducing diverse data to be more specific and ultimately producing more concrete information. Some other examples of slicing and dicing the data include considering revenue by region, injuries by project type, injuries by location, and injuries by floor height. To implement slicing and dicing during an analytics project, a three-step approach should be followed:

1. assess a list of all data features,
2. select only those features that are relevant to the business question, and
3. work with the extracted data.

Slicing and dicing should be done based on the analyst’s understanding of organizational business practices. Important business problems should be given high priority because that in turn will aid in gathering high-quality data. The slicing and dicing of data should be based on the scope of the problem being consider because the information and analysis required depend on the objective of the analytics project. Data managers or report developers can use slicing and dicing to select and present the most meaningful data.

6.1.4.2 Splicing:
The construction industry is often unable to convert data from multiple sources into relevant information and integrate it into decision-making processes, which causes a progressive loss of the
generated information (Turk 2007). Jachimowicz (2017) states that different departments of an organization have their own data sets and it is important to link them. Meanwhile, El-Diraby (2012) argues that effective decision-making and reasoning requires the integration of data and information from various domains. Designing processes to connect fragmented departments and linking various data sets can lead to useful insights. Researchers in the construction industry are already using data fusion techniques to integrate multiple data sources. Industry professionals in the civil and construction industries are use the splicing technique at numerous places to connect two structural elements. Splicing is used to link different pieces or blocks in order to create a new item where emphasis is given to the key consideration. However, the lack of talent in the construction industry capable of implementing data fusion techniques has led to scalability issues (Blanco et al. 2016; Changali et al. 2015). It is necessary to bridge this gap by introducing a more user-friendly approach to the technique. Splicing integrates multiple data sets from various sources and creates a new coherent data set from them. Moreover, splicing is the process of connecting different data sets by using a common variable, and the splicing approach can be taken from relational databases. A relational database is a collection of data in many tables that can be stored, maintained, and retrieved using a relational database management system. To join any two tables or databases, SQL joining commands such as INNER JOIN, LEFT JOIN, RIGHT JOIN, and FULL OUTER JOIN should be used as shown in Figure 6-6.

![Figure 6-6. SQL joins (W3schools.com 1999).](image)

Combining data from many different sources provides an opportunity to gain new perspectives and understand different contexts; such outcomes are not possible using sliced data (Zeid 2014). Splicing helps facilitate understanding of the interaction among business units. It also aids in the integration of organizational strategy with business analytics, which is dependent on

1. internal competencies,
2. technological options, and
3. competitive situations in the market.
6.1.4.3 Layering:

Layering makes it possible to compare different data sources and provides an aggregated view of the problem. A layering-based example can improve the functionality of a single layer. Thus, layering is an important and unique concept that integrates a number of discrete data sets or information and leads to create additional information that otherwise would not be found through traditional analytical methods. This technique also facilitates the visualization of different data sets into a cake shape or the kind of layers that comprise soil. It is possible to extract more information using a layering approach because of the amount of information present in other layers of data. This can also reduce data collection efforts, final analysis time, and computational cost. Layering is a powerful concept that allows information from different layers to be compared as well as the analysis of information in combination with various layers; looking at information this way can uncover the optimal result. Because construction projects are unique in nature, historical data might not have much importance for construction organizations. However, it is difficult to estimate the production rate of a given process because of the absence of the historical data. Therefore, projects are typically planned based mostly on personal experience. Layering is important because it not only helps reveal additional information from existing knowledge but also helps compare things in a way not possible using traditional methods.

The layering process can put an information layer on top of the strategy layer to help an organization gain a competitive advantage. Used regularly, layering will reduce the amount of data required to find additional information. Because each layer is specific to itself and contains information that is different from other layers, when we are applying layering, we want information from each layer to identify an additional layer of knowledge. Layering is the process of converting large data into small pieces of information and then combining that information to get even more information. Even though all layers have different pieces of information, they all have some related features as well. When these layers are combined, and the related features aligned, additional information will again be revealed. Therefore, it is necessary to introduce these layers together in order to capture any inherent relatedness. In the layering process, the relevant information is a part of all the layers present, and focus in the layering process should be on the final information created or final goal.

To reiterate, layering is the process of extracting information from one or more input layers and using that information to create additional information. When layering, multiple data layers
that have independent information are organized to provide new information, which helps in gaining knowledge and achieving business objectives. A generic application of layering features a map where a data set is tied to a location and different layers of data are layered over one another. Tables or graphs can also be used to represent information. In Figure 6-7, layer 1 combines with layer 2 to yield layer 3.

![Diagram of layering](image)

*Figure 6-7. The concept of layering (Stevens et al. 2012).*

To facilitate layering in decision-making, an individual should answer the following questions before moving forward with the analysis:

1. What is the additional information required?
2. What layers should be overlaid?
3. How should different layers of information be overlaid? (That is, the sequence of information.)

It is important to note that not all the information or data available is required for analysis. It is possible to select three different data sets for further study as shown in Figure 6-8.
It is also possible to select two different data sets and uncover additional information that can be used with other data sets in the same analysis, as shown in Figure 6-9.

However, it is still a challenge to identify data sets that overlay. Therefore, we are proposing that only the most useful data should be manipulated in a layering process. Thus, it is
essential to identify all the types of information that can be obtained from each data set. The overall layering process can be understood through the example shown in Figure 6-10.

<table>
<thead>
<tr>
<th></th>
<th>Data 1</th>
<th>Data 2</th>
<th>Data 3</th>
<th>Data 4</th>
<th>Data 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective 1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Objective 2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Objective 3</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Objective 4</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6-10. Layering example.*

Suppose there are five different data sets and each data set provides different information. Therefore, layering those data sets will help us discover additional information. For example, as shown in Figure 6-10, suppose we would like to achieve objective 1. Then we should select data 1, data 2, and data 4 since each data set provides different information and can help us fulfill objective 1.

Let us consider a second example where we want to find a house in a city, and the relevant information is being presented in different layers as shown in Figure 6-11.

*Figure 6-11. Placement of layers (University of Washington 2005).*

Here, a layer represents a geographical plane or space, and the individual layers are further described below.

1. Layer 1: Infrastructure or road network
2. Layer 2: # of schools within 2 miles
3. Layer 3: # of hospitals within 2 miles
4. Layer 4: Weather data
Thus, the layering can also be written as:

Selection of the best house: \( \alpha \) (Roads) + \( \beta \) (# of schools) + \( \gamma \) (# of hospitals) + \( \delta \) (weather)

It is important to note that all of the layers are providing different information. However, they also share a feature of information that connects them to each other because each layer has the same objective of providing the maximum level of comfort possible to a person. In this example, the selection of the best house is a linear combination of roads, schools, hospitals, and weather. Moreover, \( \alpha, \beta, \gamma, \) and \( \delta \) are constant. For a person looking for a new house, \( \alpha, \beta, \gamma, \) and \( \delta \) are fixed and can be selected according to individual priorities.

The construction industry has many examples that would benefit from layering. Risk management is a subjective approach in the industry that involves development risk, construction processes risk, operational risk, and financial risk, making it an ideal candidate for layering. Using a multilayered concept in this case makes sense because it can show the likelihood of an event occurring during the life span of a project (Abderisak and Lindahl 2015). Location-based examples can also be visualized and manipulated using layering. To increase data-driven decision-making in an organization, a leader should accept that there is no shortcut for success and that it is essential to integrate data across the organization. The data refinement process is the foundation of an analytics project and helps to produce more actionable insights. Splicing and layering is essential to gain knowledge and execute actions within the broader organizational context. Laursen and Thorlund (2016) argue that the integration of different products helps in the process of future product development. We believe layering is a part of the data refinement process that facilitates the extraction of new robust information to solve a business problem.

6.1.5 Analyses: Data analyses and modeling

6.1.5.1 Data Analysis:

Once the data is in the required form, the analyses and modeling step focuses on extracting information or knowledge from the current data set. Analytics has various applications in multiple disciplines and different industries. There are various analytical techniques that fit various situations. It is essential to have some idea about the business problem and an understanding of the type of analysis required for the problem. Analytics is not only about crunching numbers; it also involves other disciplines such as psychology and behavioral science. For example, Ransbotham (2015b) states that for any analysis in the marketing area, segmentation analysis is
the best approach because it divides customers into different groups, which helps with comprehending the characteristics or behaviors of the groups. Segmentation analysis is a predictive approach that deals with finding new customer categories to inform prices and promotions. Many organizations are frustrated because they are using personal judgments or management experience over data-driven analysis. However, it is data analysis that provides more benefits than management experience (Kiron et al. 2014).

The analyses and modeling step includes a wide variety of analytical techniques. Because there is no one-size-fits-all solution for different problems, analysts should use their own knowledge to come up with suitable analytical techniques. The goal of Phase 5 is presented in Figure 6-12 and ultimately centers on implementing analytical models after spending time understanding the problem as defined in Phase 2.

---

**Figure 6-12.** Phase 5: Analyses- Data Analyses and Modeling.

### 6.1.6 Outcome: Interpretations and outcomes

#### 6.1.6.1 Deployment:

The aim of this stage is to embed analytics in the decision-making process. A good analyst with strong critical thinking will always be curious to understand the root cause behind the problem and can tell how the outcome can help decision makers arrive at sound choices. The analyst can focus on implementing advanced analytics; however, he should understand that decision makers may not be aware of advanced data models. At the end of an analytics project, an analyst should ensure that the final implementation or deployment is providing actionable insights and serving decision makers in the intended way.

One of the biggest mistakes that an analyst can make is applying complicated mathematical models when they are not required to uncover actionable insights. Most of the time, a business analyst will go into data exploration to provide useful insights and recommendations to decision makers so that the company can gain a competitive advantage in the market. Thus, data visualization has become a crucial tool in analytics because it provides fast results with visual
graphs and charts that are easy to share. The analyst should focus on solving the business problem using basic tools and then communicate any recommendations to decision makers. This analytical solution should be implemented at the organizational level with the intention of improving decision-making processes by reporting a solution, dashboard, or predictive model, or by providing recommendations as shown in Figure 6-13.

<table>
<thead>
<tr>
<th>Phase 6: Outcome-Interpretations and outcomes</th>
<th>Task 6a: Deployment - Reports, dashbaords, predictive models, or recommendations.</th>
<th>Key Players: The analyst, decision maker, and end users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output: Facilitate data-driven decision-making using reports, dashboards, predictive models, or recommendations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6-13. Phase 6: Outcome- Interpretations and outcomes.*
It should be clarified that analytics is not a onetime thing, but a repetitive process used to better understand problems. While the world is becoming more complex, technology continues to quickly change, and financial rules are becoming more complicated. It is the analyst’s responsibility to see all this change as an ongoing process. In the end, the goal of any project is to create value and facilitate the decision-making process. If an analyst or a decision maker feels that the analytics solution has not created any value, then the protocol should be revisited, and the analyst should iterate Phases 2, 3, 4, or 5 as needed. Having an established metric or criterion will make it easier to go back to a previous phase as shown in Figure 6-1. For an analyst, it is important to understand the output of each phase, so they have been defined below.

1. Phase 1 outcome: A document detailing top business priorities and analytical goals
2. Phase 2 outcome: Identification of information to be extracted and selection of methods for the analysis
3. Phase 3 outcome: Processed data
4. Phase 4 outcome: Organization of the data into a structured format
5. Phase 5 outcome: Extraction of information or formulation of knowledge

To successfully deploy the analytics project, time should be taken to consider whether the project’s outcome, in the form of extracted information or derived knowledge, is answering the business question and fulfilling the scope of the project as defined in task 1e of Figure 6-1. Since executives generally support analytics projects, the analyst should engage them in evaluating its effectiveness. If the analytics project is not able to answer the business question, then its outcome should be reevaluated. During this process, an analyst should compare the expected outcome with the achieved outcome for each phase as summarized above. If the outcomes differ, then the analyst should go back to the relevant phase and implement the protocol again from that point onward.
7. IMPLEMENTATION OF THE PROTOCOL IN CASE STUDIES

7.1 Case Study 1: The Fire Engineering and Maintenance Department at Purdue University

Purdue University’s Fire Engineering and Maintenance Department is a part of the Physical Facilities Department. The chief engineer in the Fire Engineering and Maintenance Department is the decision maker and decides the budget every year for his department before sending it to Physical Facilities Department. There are total of 15 departments in Physical Facilities, and Physical Facilities decides the final amount that each department will receive based on the requested budget and total available funds. Physical Facilities has to make sure enough money is allocated to all the departments to meet their needs. For the past few years, the chief engineer of the Fire Engineering and Maintenance Department has struggled with putting together an accurate budget. As a result, the department will only sometimes receive full funding. Being short on funding means that the department’s work activities throughout the year may be affected. Therefore, the chief engineer of the Fire Engineering and Maintenance Department is curious to see whether extracting information from historical data can help him compile a more accurate budget.

For analysis and data visualization, a data visualization software called Tableau has been used in this case study. The protocol described in Chapter 6 was implemented in the manner explained below to help the chief engineer in his decision-making process.

Table 7-1: Implementation of the Protocol on the Fire-Safety Case Study (Part 1)

<table>
<thead>
<tr>
<th>Task #</th>
<th>Task Name</th>
<th>Description of the Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>Conceptualization: Describe the current business situation</td>
<td>Key Players: The decision maker and analyst</td>
</tr>
<tr>
<td>1a</td>
<td>Problem Statement</td>
<td>Describe the problem in your own words.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Each year, a certain budget amount is allocated to the Purdue Fire Engineering and Maintenance Department from the Physical Facilities Department. Some years, the</td>
</tr>
</tbody>
</table>
expenses exceed the allocated budget, and other years, they do not. This affects the department’s planning process for the future budget and any possibilities of getting more money in upcoming years because the budget is decided based on the previous year’s expenses.

<table>
<thead>
<tr>
<th>1b</th>
<th>Business Implications</th>
<th>How is it affecting your business?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• The Fire Engineering and Maintenance Department is not able to accurately plan its budget for maintenance work.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The Physical Facilities Department has a fixed amount of money that it distributes to 15 departments also in total. If the Fire Engineering and Maintenance Department requests a large amount of money, then other departments might not get enough for their needs. Therefore, there is an opportunity cost involved for the other departments.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1c</th>
<th>Current Solutions</th>
<th>What existing practices do you follow to resolve this problem?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• Base plus budget: Dollars (this year) + inflation + salary increase (If the expenses and the budget allocated are comparable, then the Fire Engineering and Maintenance Department adds a value for inflation and then submits the new budget.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Change the budget based on the previous year’s expenses. (If there is a large difference between the department’s budget and expenses, then the Fire Engineering and</td>
</tr>
</tbody>
</table>
Maintenance Department will submit a budget based on the previous year’s expenses.)

Write advantages & disadvantages of these practices.

- Both approaches are good but do not give an idea of how much money will be required for the upcoming year. Therefore, a promising idea at the task level or work-order level is required for better estimation of the budget.

<table>
<thead>
<tr>
<th>Id</th>
<th>Entities Involved</th>
<th>Define the project stakeholders of the analytics project.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1d</td>
<td>The Physical Facilities business office, which approves the department’s budget.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Fire Engineering and Maintenance Department, which is receiving the money.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The 14 other Physical Facilities departments because they must deal with individual opportunity costs. (If one department gets more money, then the other departments might get less money than requested because Physical Facilities has a limited amount of money to distribute.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Define the end users of the analytics project.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Physical Facilities business office, which can intelligently allocate money to the Fire Engineering and Maintenance Department. (This will also help Physical Facilities allocate the rest of the money to the other department.)</td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>Scope of the Data Analytics Project</td>
<td>Definition of the analytics project.</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>1e</td>
<td>Define the goal of the analytics project.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Understanding the mix of expenses from different cost accounts.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Identify work orders and any activities that are causing major expenses.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Data Available</th>
<th>What data sets are available for analysis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1f</td>
<td>What data sets are available for analysis?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- In-house: The “Maintenance activities” data file.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Describe the important variables for analysis.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Work-order type = [Null, CONS, RFS] (Out of several work-order types, only three are important as explained at the end of Table 7-1.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- PM activity type, work-order description, payroll amount charged, work-order amount charged, and buildings are important variables from the “Maintenance activities” data file.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 2</th>
<th>Design: Understanding the problem</th>
<th>Key Players: The analyst</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Task 2</th>
<th>Exploratory Analysis</th>
<th>Explore the available data and understand what information can be extracted.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a</td>
<td>Explore the available data and understand what information can be extracted.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Explore the payroll and work orders in the “Maintenance activities” data file (It was found that there is a hierarchy in the data as shown in Figure 7-1. The information in</td>
<td></td>
</tr>
</tbody>
</table>
| Task 3 | Data Quality | Identify the quality of the existing data.  
|---|---|---|
| 3a | Data Quality | • Accurate? – Yes.  
| | | (All the values in the data are correct values and are accurately described based according to Section 4.4.3.)  
| | | • Completeness? - Yes.  
| 2b | Problem Structure | Translate the business questions into analytical questions.  
| | | • Which major cost accounts have high variability?  
| | | • Divide the cost accounts with high variability into a system of smaller units until they reach the operational level (i.e., work-order description or activity level; refer to Figure 7-1.)  
| | | • Identify critical buildings and maintenance activities to take corrective actions.  
| 2c | Analyses Overview | Determine what types of analyses are required.  
| | | • Calculate the coefficient of variation (CV) to find variability in the cost accounts.  
| | | • Use cluster analysis to find critical buildings with high unit work-order amounts charged.  
| | | • Use cluster analysis to find critical maintenance activities with high work-order amounts charged.  

Figure 7-2 and 7-3 was also obtained from exploratory analysis.)
There are some null values in the WO [work order] description column but they reflect general department operating budget activities. These activities do not have work orders assigned to them."

- Consistent? – Yes.
  (All the values are consistent and do not violate any constraint.)

<table>
<thead>
<tr>
<th>3b</th>
<th>Data Structure Needed</th>
<th>Decide on the arrangement of your data.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Data variables needed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Work-order type = [Null, CONS, RFS]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Here only Null, CONS, and RFS work orders are needed as per decision maker requirements because most of the budget has been spent on only these three types of work-orders based on the knowledge of the decision maker.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Payroll amount charged and year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data variables available</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Work-order type = [Null, 18, AFF, CONS, IHD, RFS]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(These work-order types are defined at the end of Table 7-1.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Payroll amount charged and year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additional data strings needed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- None.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(No extra data is required because we are only interested in three types of work orders, and the data for those work orders is already available.)</td>
</tr>
<tr>
<td>Column</td>
<td>Section</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 3c     | Data Sources | Identify the sources of additional data to be collected.  
• None.  
(There is no need for additional data collection, or no data strings are needed as defined in Section 3b of this table. Therefore, there is no need to look for other sources of data.) |
| 3d     | Data Collection | Identify the constraints to capture the data or final data limitations.  
• None.  
(There is no need to collect additional data because there are no constraints or limitations in the data collection.) |
| 3e     | Supplementary Information | Decide on additional information required to address queries/questions.  
• None.  
(Required information is already available based on the exploratory analysis done in Section 2a of this table.) |
| 3f     | Data Cleaning |  
• Check header.  
(Column headers are perfectly labelled, there are no problems with the data type, and categories are properly labelled.)  
• Check data types.  
(All the data types are perfectly defined.)  
• Check category labelling.  
(All column categories are properly labelled.) |

| Task 4 | Refinement- Data refinement | Key Players: The analyst and data engineers |
| 4a  | Slicing and Dicing (Data selection) | Ask business questions to identify important data attributes.  
• Work-order type: Select [CONS, Null, RFS]  
  (Here only CONS, Null, and RFS work-order type variables are needed as per task 3b.) |
| 4b  | Splicing (Connecting different data sets) | What are the different data sources required to address the business problem?  
• No additional data is required because we are interested in only the selected cost accounts with high variability as shown in Figures 7-2 and 7-3.  
How do we merge or combine those data sources?  
• No additional data is required. Therefore, there is no need to merge or combine data sources. |
| 4c  | Layering (Map data) | Identify different layers of information to find additional information.  
• No additional information is required because we are interested in only the selected cost accounts with higher variability shown in Figures 7-2 and 7-3. |
| Task 5 | Analyses - Data analyses and modeling | Key Players: The analyst |
| 5a  | Data Analysis | What analysis and methods are suitable to address the business problem?  
• Calculate the CV to identify which work-order types are causing maximum variability and drill down into that work-order type as shown in Figure 7-4. |
Based on task 2b defined in Table 7-1, once high-variability work-orders have been identified, the analyst should drill down to project maintenance (PM) activity types. After that, the PM activities with high variability should be drilled down to the buildings and work-order description level as shown in Figure 7-1. Based on Figure 7-1, the Fire Engineering and Maintenance Department has divided all of its activities into certain work-order types. These work orders and corresponding descriptions for their codes are detailed below.

1. AFF: Work orders affiliated with Purdue Research Foundation, student organizations, or others
2. IHD: Work orders corresponding to in-house design
3. CONS: Work orders based on construction
4. RFS: Work orders related to request for services
5. Null: All the general department operating budget activities have been reclassified as the Null work-order type
6. 18: Work orders related to building repairs

Based on the hierarchy defined in Figure 7-1, all the work-order types were divided into PM activities. It should be noted that this hierarchy was found during the exploratory analysis done in task 2a of Table 7-1. The selected PM activities shown in Figure 7-1 are explained below.

1. 009: Scheduled or planned work done by the department during new construction projects on campus
2. 018: Schedule or planned work done by the department during major repairs of campus buildings
3. 019: Unplanned work done by the department based on urgent campus building needs (time and materials [T&M] construction)
During the exploratory analysis, it was observed that data was available for 6 years from July 2009 to June 2015. As a result, the date column in the maintenance activities data was changed to the fiscal years shown below.

1. July 2009 to June 2010: Year1
2. July 2010 to June 2011: Year2
3. July 2011 to June 2012: Year3
4. July 2012 to June 2013: Year4
5. July 2013 to June 2014: Year5

Also during the exploratory analysis, as discussed in Section 2a of Table 7-1, the payroll amount charged was drawn for all the WO (work order) types shown in Figure 7-2.
Based on the business requirement, the decision maker is interested in only Null, CONS, and RFS work-order types because they make major contributions to the payroll amount charged as shown in Figure 7-2. The slicing part of the analysis is explained below.

1. Available work-order types = [Null, 18, AFF, CONS, IHD, RFS]
2. Work-order types required = [Null, CONS, RFS]

Therefore, it is necessary to slice the data as per task 4a in Table 7-1. After slicing the data, the sliced data for the payroll amount charged was drawn as shown in Figure 7-3.
Finally, to identify variability in the cost accounts as defined in task 5a of the protocol in Table 7-1, the CV was calculated for each cost account as shown in Figure 7-4. Based on the CV, the CONS work-order type has the maximum variability. Moreover, based on the CV, the Null work-order type has high variability. It should be noted that Null work-order type activities correspond to the general department operating budget. These activities do not have any work-
order descriptions. However, for inclusion in the analysis, these general department operating budget activities have been labelled as “Null” work-order activities.

**Intermediate Conclusion 1:**
As seen in Figure 7-4, the Null and CONS work-order types have higher variabilities than RFS. Therefore, an analyst should drill down into the Null and CONS work orders. The Null work-order type is just a general department operating budget. Null work orders do not have any work-order description or PM activities type; therefore, drilling down into Null work order types is not useful. Descriptions of general department operating activities are missing from the data, which impacts the Null work-order descriptions. Therefore, it is necessary to capture any descriptions corresponding to general department operating activities so that the Fire Engineering and Maintenance Department can track its expenses in the future and minimize them.

The CONS work-order type was also selected for drilling down, and these work orders are related to campus construction projects. The higher variability in the campus construction projects affects the department’s budget needs. The analysis done based on task 5a in Table 7-1 does not fulfill the scope of the analytics project as defined in task 1e of Table 7-1. Therefore, it is necessary to go back to the appropriate place in the protocol. There is no need to apply the protocol again from Phase 1, and the output of Phase 2 is correct because we are aware of the required information and analytical models. Since we have also converted the raw data into meaningful data, the goal of Phase 3 has been achieved. However, better selection of data could be done for drilling down into CONS work orders so that it will be easier to identify the variability for PM activity types within the CONS work-order type. Based on the output of the different phases defined in the protocol, as presented in Figure 6-1, we should return to Phase 4 because we are interested in the better selection of data. First, the CONS work-order data should be sliced and then the variability for the different PM activities should be calculated. Therefore, the protocol is again applied starting from task 4, and the procedure is detailed in Table 7-2.
Table 7-2: Implementation of the Protocol on the Fire-Safety Case Study (Part 2)

<table>
<thead>
<tr>
<th>Task #</th>
<th>Task Name</th>
<th>Description of the Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 4</td>
<td>Refinement - Data refinement</td>
<td>Key Players: The analyst and data engineers</td>
</tr>
</tbody>
</table>
| 4a     | Slicing and Dicing (Data selection) | Ask business questions to identify important data attributes.  
  - WO type: Select [CONS]  
    (We are interested in only drilling down into the CONS work-order type.) |
| 4b     | Splicing (Connecting different data sets) | What are the different data sources required to address the business problem?  
  - Not required. (We are only drilling down. Therefore, we do not need any additional data for the analysis.)  
  How should we merge or combine these data sources?  
  - Not required. |
| 4c     | Layering (Map data) | Identify different layers of information to find additional information.  
  - Not required. (Layering is not required because we do not need any additional information from the existing data.) |
| Task 5 | Analyses- Data analyses and modeling | Key Players: The analyst |
| 5a     | Data Analyses | What analysis (descriptive, diagnostic, predictive, or prescriptive) or methods are suitable to address the business problem?  
  - Calculate the CV to identify which PM activity type is causing maximum variability and drill down into that PM activity (see Figures 7-5 and 7-6). |
After slicing the data as described in task 4a of Table 7-2, the work-order amount charged was drawn for the PM activity type as shown in Figure 7-5. The description of these PM activities is provided below.

1. **009**: Scheduled or planned work done by the department during new construction projects on the campus
2. **018**: Scheduled or planned work done by the department during major repairs of campus buildings
3. **019**: Unplanned work done by the department based on the urgent campus building needs (T&M construction)

*Figure 7-5. Historical data of WO amount charged for PM activity type.*
The 009, 018, and 019 PM activity types from Figure 7-5 correspond to the CONS work-order type because we are interested in drilling down into the CONS work-order type. Based on Figure 7-6, the 018 PM activity type has the maximum variability because its CV is 81%. The CVs for 009 and 019 are also significantly high. Moreover, the 009 and 018 PM activity types are planned or scheduled activities while the 019 PM activity is an unplanned activity. There is a higher possibility of reducing variations in the expenses of work-order amounts charged for the 019 PM activity type, as shown in Figure 7-6, because 019 is not a planned activity. Thus, there is a need to drill down into 019 in an effort to save more money since reducing the variability in the cases of 009 and 018 would not save as much money in comparison to 019.

**Intermediate Conclusion 2:**

To find the variability in the different cost accounts of the PM activity types, the CV is calculated as shown in Figure 7-6. Reducing variability for the 019 PM activity type (T&M construction) is more useful for the Fire Engineering and Maintenance Department than doing so for the 009 and 018 PM activity types because 019 has more money-saving potential. Therefore, to the analyst should drill down into the 019 PM activity type. The analysis done in task 5a of Table 7-2 does not fulfill the scope of the analytics project as defined in task 1e of Table 7-1. As a result, it is necessary to drill down into the 019 PM activity type and apply the protocol from task 3 on as shown in Table 7-3. As per Figure 7-1, variability at the building level should be analyzed after assessing the variability in PM activities. However, we do not have the areas of the buildings to...
analyze variability at building level with. Since additional data is required, the protocol is applied again beginning with Phase 3, as shown in Table 7-3, for an in-depth understanding of the data.

Table 7-3: Implementation of the Protocol on the Fire-Safety Case Study (Part 3)

<table>
<thead>
<tr>
<th>Task #</th>
<th>Task Name</th>
<th>Description of the Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 3</td>
<td>Development: Understanding the data</td>
<td>Key Players: The analyst and data engineers</td>
</tr>
<tr>
<td>3a</td>
<td>Data Quality</td>
<td>Identify the quality of the existing data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Completeness? - Yes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(After slicing, we have only the CONS work-order type. Therefore, there are no null values in the data.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Consistent? – Yes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(All the values are consistent and do not violate any constraints.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Accurate? – Yes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(All the values in the data are accurately described.)</td>
</tr>
<tr>
<td>3b</td>
<td>Data Structure Needed</td>
<td>Decide on the arrangement of your data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data variables needed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• PM activity type = [19]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• WO description, WO amount charged, building code, and area of buildings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data variables available</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• PM activity type = [009, 018, 019]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• WO description, WO amount charged, building code</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additional data strings needed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Areas of buildings data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(We need to calculate the unit maintenance cost for each building.)</td>
</tr>
<tr>
<td>3c</td>
<td>Data Sources</td>
<td>Identify the sources of additional data to be collected.</td>
</tr>
</tbody>
</table>
| 3d | Data Collection | Identify the constraints to capture the data or final data limitations.  
|    |                | • If we start by collecting the areas of buildings data on our own, then it will be a difficult and time-consuming process. However, there is a possibility that Physical Facilities will already have the areas of buildings, in which case there are no constraints. |
| 3e | Supplementary Information | Decide on additional information required to address queries/questions.  
|    |                | • Sources to find areas of buildings.  
|    |                | (Since we are unable to find this data, it is important to ask for the information from the decision makers. In this case, a decision maker provided the data.) |
| 3f | Data Cleaning | • Check header.  
|    |                | (The column headers are perfectly labelled, there are no problems with the data type, and all categories are properly labelled.)  
|    |                | • Check data types.  
|    |                | (All the data types are perfectly defined.)  
|    |                | • Check category labelling.  
<p>|    |                | (Some of the building abbreviations are different in the “Area of buildings” file when compared to the “Maintenance activities” file. Therefore, abbreviations in the “Maintenance activities” file have been changed according to the “Area of buildings” file.) |</p>
<table>
<thead>
<tr>
<th>Task 4</th>
<th>Refinement- Data refinement</th>
<th>Key Players: The analyst and data engineers</th>
</tr>
</thead>
</table>
| 4a    | Slicing and Dicing (Data selection) | Ask business questions to identify important data attributes.  
• PM activity type: Select [19] |
| 4b    | Splicing (Connecting different data sets) | What are the different data sources required to address the business problem?  
• Data present: “Maintenance activities” data and “Area of buildings” data  
How should we merge or combine these data sources?  
• Identify the building codes for each data set and then take left join of “Maintenance activities” data with the “Area of buildings” data as shown in Figure 7-7. |
| 4c    | Layering (Map data) | Identify different layers of information to find additional information.  
• Not required.  
(Layering is not required because we do not need any additional information from the existing data sets.) |

<table>
<thead>
<tr>
<th>Task 5</th>
<th>Analyses- Data analyses and modeling</th>
<th>Key Players: The analyst</th>
</tr>
</thead>
</table>
| 5a    | Data Analyses and Modeling | What analysis and methods are suitable to address the business problem?  
• Use $k$-means clustering analysis to find buildings that have the highest unit work-order amounts charged.  
• Use $k$-means clustering analysis to find work-order descriptions or activities that have the maximum work-order amounts charged. |
<p>| Task 6 | Outcome – Interpretations and outcomes | Key Players: The analyst, decision maker, and end users |</p>
<table>
<thead>
<tr>
<th>6a</th>
<th>Deployment Report, dashboard, predictive model, or recommendations.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Most of the buildings in cluster 1 and cluster 2 (high-cost clusters), as shown in Figure 7-8, are laboratory buildings. It might be expected that laboratory buildings have higher fire maintenance costs, but with proper ventilation design, exhaust systems installation, and careful handling of hazardous elements, these costs can be reduced.</td>
</tr>
<tr>
<td></td>
<td>• The WO description shown in Figure 7-9 does not have a consistent task description. Therefore, it is not useful to analyze the data at the activity level. A data management recommendation for items such as the proper labeling of WO descriptions is required at the organizational level so that everyone will have the same understanding of tasks or WO descriptions. Standardizing task descriptions using a fixed list will not help the department immediately. But in the future, it will provide a transparent platform to all employees inside the organization and will facilitate a data-driven culture.</td>
</tr>
</tbody>
</table>

Based on task 3b as defined in Table 7-3, the areas of buildings are missing in the “Maintenance activities” data. Therefore, to calculate the unit work-order cost for each building, it is necessary to collect their areas. After getting the areas of buildings data from the decision makers, a left join technique was applied as described in task 4b, and this technique is frequently used in a relational database. The left join technique, with results shown in Figure 7-7, was used because not all buildings were present in the “Area of buildings” file. Moreover, left join was used because then all the records in the left table are preserved and only matched with records returned from the right table. When joining two data sets, it is important not to lose any data. Therefore, with left join, all the entries in the “Maintenance activities” file will remain intact.
After slicing and splicing, as described in Table 7-3, the $k$-means clustering technique was applied on:

1. the unit WO amount charged for each building and
2. the WO amount charged for each WO description.

After applying the $k$-means clustering technique, clusters with higher values of unit WO amount charged for each building and WO amount charged for each WO description are selected as shown in Figures 7-8 and 7-9 respectively. We are interested in finding the different cost accounts that are causing high variability. The $k$-means clustering approach was chosen in task 5a of Table 7-3 because it divides all the observations into numbers of clusters. When using the $k$-means clustering approach, it is important to maintain the quality of each cluster (Hung and Kang 2014). There is no fixed approach to define the number of clusters. However, Tableau selects the optimal number of clusters based on Calinski-Harabasz criteria. In this case, a total of three clusters were identified to prevent overfitting. Due to space constraints, only clusters 1 and 2 are presented in Figure 7-8 because the buildings corresponding to clusters 1 and 2 had the maximum unit work-order costs. Similarly, a total of five clusters were identified for WO amount charged. In this case, only clusters 2 and 3 were selected for Figure 7-9 because they represent activities that resulted in the maximum work-order amounts charged.
WO-$/Area

Bldg | WO Type
--- | ---
RSC | CONS
GRIS | CONS
HARR | CONS
BIND | CONS
UPOF | CONS
AERO | CONS
PRCE | CONS
WTHR | CONS
ELLT | CONS
LILY | CONS
HERL | CONS
RPH | CONS
BRWN | CONS
CIVL | CONS
KRAN | CONS
MJIS | CONS
FRNY | CONS
MRRT | CONS
PUSH | CONS
HCK | CONS
BCHM | CONS
HOVD | CONS
YONG | CONS
STEW | CONS
KNOY | CONS
MATH | CONS
ME | CONS
PSYC | CONS

Figure 7-8. Clustering of buildings vs. unit WO amount charged.
### WO/PM Activity/year/Building

<table>
<thead>
<tr>
<th>WO Type</th>
<th>PM Activity Type</th>
<th>Bldg</th>
<th>WO Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONS</td>
<td>019</td>
<td></td>
<td>WO 11510-ARRA - Test fire alarm system.</td>
</tr>
<tr>
<td>BRWN</td>
<td></td>
<td></td>
<td>close per PM Jim Naville</td>
</tr>
<tr>
<td>CIVL</td>
<td></td>
<td></td>
<td>10877-Shut system down for the contract</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11052 - Shut-down fire sprinkler/fire alarm</td>
</tr>
<tr>
<td>ELRT</td>
<td></td>
<td></td>
<td>11400 - Fire alarm shutdowns, tie-ins an</td>
</tr>
<tr>
<td>FRNY</td>
<td></td>
<td></td>
<td>10895-ABATMENT, PERMITS, ALARM BAGG</td>
</tr>
<tr>
<td>FSTW</td>
<td></td>
<td></td>
<td>11076-Provide support to contractor for</td>
</tr>
<tr>
<td>GRLS</td>
<td></td>
<td></td>
<td>11366-ARRA-Support construction project</td>
</tr>
<tr>
<td>HAMP</td>
<td></td>
<td></td>
<td>12676-Provide zone support for project.</td>
</tr>
<tr>
<td>HARR</td>
<td></td>
<td></td>
<td>9354-Fire Safety concerns at Harrison Hall</td>
</tr>
<tr>
<td>IAF</td>
<td></td>
<td></td>
<td>9354-Fire safety concerns. Provide Hot work permits as needed. Coo</td>
</tr>
<tr>
<td>KRAM</td>
<td></td>
<td></td>
<td>Fire protection related Misc. work at KR</td>
</tr>
<tr>
<td>LAF</td>
<td></td>
<td></td>
<td>Provide testing services and furnish and</td>
</tr>
<tr>
<td>LILY</td>
<td></td>
<td></td>
<td>10951-Fire Safety Concerns. Location: Gr</td>
</tr>
<tr>
<td>LILY</td>
<td></td>
<td></td>
<td>9651 - Fire Safety Concerns</td>
</tr>
<tr>
<td>LILY</td>
<td></td>
<td></td>
<td>Fire alarm shut downs and reworking ecls</td>
</tr>
<tr>
<td>MACK</td>
<td></td>
<td></td>
<td>9384-DR4 Phase IV Hot Work Permits/FES</td>
</tr>
<tr>
<td>ME</td>
<td></td>
<td></td>
<td>Post 4/1/12 FES for Gaylor FA re-work</td>
</tr>
<tr>
<td>MJS</td>
<td></td>
<td></td>
<td>Remove smoke detectors and pull stations</td>
</tr>
<tr>
<td>PUFF</td>
<td></td>
<td></td>
<td>hot work permits, FA devices (remove, re</td>
</tr>
<tr>
<td>RPHN</td>
<td></td>
<td></td>
<td>11316 - All fire alarm issues.</td>
</tr>
<tr>
<td>RSC</td>
<td></td>
<td></td>
<td>Provide Fire Alarm Inspection and testing</td>
</tr>
<tr>
<td>SHRVC</td>
<td></td>
<td></td>
<td>remove, replace, test Fire Equip Service</td>
</tr>
<tr>
<td>STEX</td>
<td></td>
<td></td>
<td>For Hot Works and Testing</td>
</tr>
<tr>
<td>WIND</td>
<td></td>
<td></td>
<td>7026 - Fire Safety Concerns.</td>
</tr>
<tr>
<td>WTHR</td>
<td></td>
<td></td>
<td>9817 - Abatement is required- assist contr</td>
</tr>
<tr>
<td>YNGN</td>
<td></td>
<td></td>
<td>11483- Provide support services as needed</td>
</tr>
</tbody>
</table>

Sum of WO Amount Charged for each WO Description broken down by WO Type, PM Activity Type and Bldg. Color shows details about Clusters (2). The data is filtered on Year, which has multiple members selected. The view is filtered on WO Type, PM Activity Type and Clusters (2). The WO Type filter keeps CONS. The PM Activity Type filter keeps 019. The Clusters (2) filter keeps Cluster 2 and Cluster 3.

![Figure 7-9. Clustering of WO description vs. WO amount charged.](image)

---

**Figure 7-9.** Clustering of WO description vs. WO amount charged.
**Final Conclusion:**

Based on task 1e of Table 7-1, the scope of the analytics project is limited to understanding the mix of expenses at the most granular level. It is noted from Figures 7-8 and 7-9 that the analysis done in task 5a of Table 7-3 does not fulfill the project’s scope. Therefore, the decision maker should consider the recommendations provided in task 6a of the protocol, as shown in Table 7-3. The scope defined in task 1e of Table 7-1 confirms that there is no need to dig deeper. After discussing the analysis with the decision maker, the following recommendations, as defined in task 6a of Table 7-3, were provided:

1. Most of the buildings in clusters 1 and 2 (high-cost clusters), as shown in Figure 7-8, are laboratory buildings. Laboratory buildings may have higher fire maintenance cost. But with proper ventilation design, exhaust systems installation, and careful handling of hazardous elements, this cost can be reduced. The analysis done so far can help the Fire Engineering and Maintenance Department in getting a general idea about their expenses in laboratory buildings. Each year, they can add an additional amount of money into the budget calculation for these laboratory buildings, which will help them reduce the variability defined in task 1a of Table 7-1.

2. The WO description shown in Figure 7-9 does not have consistent task descriptions. Therefore, it is not beneficial to analyze the data at the activity level. The data management recommendation for items such as proper labeling of WO descriptions is required at the organizational level so that everyone will have the same understanding of activities or WO descriptions. Standardizing task descriptions will not have immediate benefits for the department. But in the long run, it will provide a transparent platform to all employees inside the organization and will facilitate a data-driven culture. Analytics is not only about increasing profile or reducing cost. Many times, an analytics project fails because of a lack of transparency inside an organization, so a decision maker should strive to build an analytical culture inside the organization. In this case, standardizing work-order descriptions is required because multiple people are updating the data. Also, standardizing work-order descriptions will provide a standard approach for everyone inside the organization to follow and bring them to the same page.
7.2 Case Study 2: Housing Market Outlook and Business Expansion Locations

The housing construction industry is facing various challenges because of the seasonality issue. Forecasting is generally used to project trends and gain knowledge using historical correlations, but companies in the housing construction industry are not able to forecast their data correctly (Laursen and Thorlund 2016). Some studies have been done using time-series modeling in the construction industry (Naderpajouh et al. 2015; Oyedele et al. 2013; Ashuri and Lu 2010; Hwang and Liu 2010; Nielsen 2005). However, there is a lack of research on predicting the housing construction market outlook. Sustained demographic unawareness in the industry has made it difficult for real estate development companies to expand their businesses to other regions. Thus, the inputs of this case study have been taken from a real estate developer in Florida. This company works on land development in the real estate sector. Given the $1 trillion infrastructure investment plan by President Donald Trump, the company believes that now is the time to expand its business to other regions. Infrastructure investment is known to have a positive impact on housing industry growth. The company’s main business is buying land and developing it based on an understanding of the market. The company is already in the land development business; therefore, the executives want to leverage their experience and become a leader in the land development sector. They are looking for optimal regions where for their business expansion. The target market for the company is single-family houses, and they are not interested in venturing into other areas such multifamily houses, commercial buildings, or the industrial sector because of their unfamiliarity with these areas. Still, expanding to other regions or cities is their business growth strategy. The protocol was applied, as seen in Table 7-4, to help them understand the current housing market outlook and find the best locations for their business expansion.
Table 7-4: *Implementation of the Protocol to Determine the Housing Market Outlook and Assess Suitable Regions for Business Expansion (Part 1)*

<table>
<thead>
<tr>
<th>Task #</th>
<th>Task Name</th>
<th>Description of the Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>Conceptualization: Describe the current business situation</td>
<td>Key Players: The decision maker and analyst</td>
</tr>
<tr>
<td>1a</td>
<td>Problem Statement</td>
<td>Describe the problem in your own words.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The company wants to grow in other geographical areas; however, it is difficult for them to figure out suitable regions for business expansion.</td>
</tr>
<tr>
<td>1b</td>
<td>Business Implications</td>
<td>How is it affecting your business?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Business expansion requires a significant amount of investment and involves huge risks. Therefore, it is important to understand the market and their customer profile so that they can create a competitive advantage in the market. Without comprehending these factors, the company cannot provide innovative solutions to their customers. This company has a good amount of resources that can be used to provide innovative solutions; however, they currently operate in only one location, which is hampering their growth.</td>
</tr>
<tr>
<td>1c</td>
<td>Current Solutions</td>
<td>What existing practices will you follow to resolve these problems?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The company is maintaining its competitive advantage by providing innovative amenities. Write the advantages and disadvantages of each practice.</td>
</tr>
</tbody>
</table>
- High quality amenities attract more customers. However, because the company is only operating in one city, it is serving a small population. Their experience in providing high-quality amenities can give them a competitive advantage in other geographical locations too.
- They are unable to be industry leaders because of their limited knowledge about the demographics of other locations.

| 1d | Entities Involved | Define the project stakeholders of the analytics project.  
|    |                  | • The company management, which is interested in expanding the business to other locations.  
|    |                  | Who are the end users of the analytics project?  
|    |                  | • The management team, which will be using this analysis to design the business expansion strategy.  

| 1e | Scope of the Data Analytics Project | Define the goal of the analytics project.  
|    |                                   | • What are the suitable regions or best cities (if possible, then counties) for business expansion?  
|    |                                   | • What is the housing sales trend in different regions?  
|    |                                   | • What is the trend for permits issued in different regions?  
|    |                                   | • How many permits will be issued in the region best suited for single-family houses?  
|    |                                   | (Their target market is single-family houses.)  

| 1f | Data Available | What data sets are available for analysis?  
|    |                | • Permits issued by region and housing type (Source: US Census Bureau)  
|    |                | • Houses sold by region |
Describe the important variables for analysis.

- Single-family for housing type
- Household growth rate and median home value

### Task 2: Design: Understanding the problem

<table>
<thead>
<tr>
<th>Key Players: The analyst</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design:</strong> Exploratory Analysis</td>
</tr>
</tbody>
</table>

Explore the available data and understand what information can be extracted.

- Explore the historical data of permits issued, construction starts, and housing sales to comprehend the opportunity for growing the business.

(During the analysis, it was found that there is a strong correlation between the permits issued and construction starts as shown in Figure 7-10. Construction starts depend on the permits issued. Therefore, the construction starts variable can be dropped from the analysis.)
- Explore other publicly available data sets such as household growth rate, median home value, crime index, and unemployment rate.
- During the exploratory analysis, it was found that the permits issued data is available monthly while construction starts, and houses sold data is available quarterly.

<table>
<thead>
<tr>
<th>2b</th>
<th>Problem Structure</th>
<th>Translate the business questions into analytical questions.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• What is the trend for houses sold by regions?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• What is the trend for permits issued for single-family houses?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Find suitable regions or counties for business expansion.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Predict the number of permits issued in the region best suited for single-family houses.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Find best-suited cities or counties for business expansion.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2c</th>
<th>Analyses Overview</th>
<th>Determine what types of analyses are required.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Use time-series modeling to determine the trend and prediction for the following items:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Housing permits issued for single-family houses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Houses sold in the single-family houses category</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 3</th>
<th>Development-Understanding the data</th>
<th>Key Players: The analyst, data engineers, and decision maker</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a</td>
<td>Data Quality</td>
<td>Identify the quality of the existing data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Completeness? - Yes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(There are no null values.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Consistent? – No.</td>
</tr>
</tbody>
</table>
(Permits issued data is available at the granular level [i.e. it is available for both single-family and multifamily houses]. However, the houses sold data is available only at the aggregated level. The developer is only interested in single-family houses. Therefore, it is useless to have aggregated data. Permits issued data is available monthly; therefore, it is aggregated and presented quarterly to match the format of construction starts and houses sold data.)

- Accurate? – Yes.
  (All values in the data are accurately described.)

| 3b | Data Structure Needed | Decide on the arrangement of your data.  
Data variables needed  
- Permits issued: As many as possible  
- Housing type: [single family]  
Data variables available  
- Permits issued: [2013, …, 2016]  
- Housing type: [single family, multifamily]  
Additional data strings needed  
- Single-family houses sold: [2013, …, 2016] |
| 3c | Data Sources | Identify the sources of additional data to be collected.  
- US Census Bureau.  
  (There is a need for an additional data string as defined in section 3b.) |
| 3d | Data Collection | Identify the constraints to capture the data or final data limitations.  
- The data is available at the aggregated level, and there is no other way to get the houses sold data for single-family houses. |
| 3e | Supplementary Information | Decide whether additional information is required to address queries/questions.  
• We need to ask the decision makers if they have access to the single-family houses sold data. In this case, they do not have any data. Therefore, we should drop the idea of predicting single-family houses sold. |
| 3f | Data Cleaning | • Check header.  
(Column headers perfectly labelled, there are no problems with the data type, and categories are properly labelled.)  
• Check data types.  
(All the data types are perfectly defined.)  
• Check category labelling.  
(All the categories in the columns are labelled properly.) |
| Task 4 | Refinement- Data refinement | Key Players: The analyst and data engineers |
| 4a | Slicing and Dicing (Data selection) | Ask business questions to identify important data attributes.  
• Housing type: Select [single family] based on task 3b of Table 7-4. |
| 4b | Splicing (Connecting different data sets) | What are the different data sources required to address the business problem?  
• None.  
(Initially, we were only interested in finding suitable regions that had the highest growth in permits issued. We have all the available data, so there is no need to combine any additional data sources.) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>How should we merge or combine these data sources?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• Not required.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(We do not need to combine any other data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sources. Therefore, this step is not needed.)</td>
</tr>
<tr>
<td>4c</td>
<td>Layering (Mapping data)</td>
<td>Identify different layers of information to find additional information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• None.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(In this case, the layering process would have</td>
</tr>
<tr>
<td></td>
<td></td>
<td>been implemented if both permits issued and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>houses sold data were available. However, we</td>
</tr>
<tr>
<td></td>
<td></td>
<td>have data on permits issued but not on houses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sold for single-family houses. Thus, there is no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>need for layering.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 5</th>
<th>Analyses- Data analyses and modeling</th>
<th>Key Players: The analyst</th>
</tr>
</thead>
<tbody>
<tr>
<td>5a</td>
<td>Data Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>What type of analysis (descriptive, diagnostic,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>predictive, or prescriptive) or methods are suitable to</td>
<td></td>
</tr>
<tr>
<td></td>
<td>address the business problem?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Use the Holt-Winters forecasting model (Winters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1960) to determine the trend for housing permits</td>
<td></td>
</tr>
<tr>
<td></td>
<td>issued for single-family houses.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Use the Holt-Winters forecasting model (Winters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1960) to predict the number of permits issued in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the region best suited for single-family houses.</td>
<td></td>
</tr>
</tbody>
</table>

Exploratory analysis was done for the construction starts and permits issued. It was found that there exists a strong correlation between both the variables shown in Figure 7-10. Construction starts is dependent on permits issued. Therefore, the construction starts variable was dropped from the analysis.
### Figure 7.10. Exploratory analysis of construction starts and permits issued.

Exponential smoothing models and Autoregressive integrated moving average models are two important time-series forecasting models used in the construction industry (Ashuri and Lu 2010; Hwang and Liu 2010; Touran and Lopez 2006; Ng et al. 2004; Hua and Pin 2000). The construction industry is strongly affected by short-term business cycles, which means that future demands are highly dependent on recent demands and that old demands do not have much impact on future demands (Goh 1998). Therefore, exponential smoothing models are more useful for modeling the housing market outlook because recent data has a bigger influence on the forecast. ARIMA models are not useful because they require a minimum of 12 years of quarterly data for forecasting (Office for National Statistics 2008). In the Holt-Winters forecasting model, which is an exponential smoothing model, forecast values are decided based on level, trend, and seasonality as described below.

1. Level: Reveals the importance of more recent data points
2. Trend: Reveals the importance of more recent changes in trends
3. Seasonality: Reveals the importance of more recent seasonal components

<table>
<thead>
<tr>
<th>Time</th>
<th>Northeast</th>
<th></th>
<th>Midwest</th>
<th></th>
<th>South</th>
<th></th>
<th>West</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Permits (in 1000s)</td>
<td>Construction starts (in 1000s)</td>
<td>Permits (in 1000s)</td>
<td>Construction starts (in 1000s)</td>
<td>Permits (in 1000s)</td>
<td>Construction starts (in 1000s)</td>
<td>Permits (in 1000s)</td>
<td>Construction starts (in 1000s)</td>
</tr>
<tr>
<td>Q1-2013</td>
<td>10</td>
<td>10</td>
<td>17</td>
<td>16</td>
<td>77</td>
<td>80</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Q2-2013</td>
<td>15</td>
<td>16</td>
<td>31</td>
<td>30</td>
<td>94</td>
<td>89</td>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td>Q3-2013</td>
<td>15</td>
<td>15</td>
<td>31</td>
<td>31</td>
<td>86</td>
<td>83</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Q4-2013</td>
<td>13</td>
<td>14</td>
<td>23</td>
<td>25</td>
<td>72</td>
<td>74</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>Q1-2014</td>
<td>11</td>
<td>9</td>
<td>17</td>
<td>14</td>
<td>80</td>
<td>79</td>
<td>31</td>
<td>32</td>
</tr>
<tr>
<td>Q2-2014</td>
<td>15</td>
<td>15</td>
<td>31</td>
<td>34</td>
<td>95</td>
<td>91</td>
<td>40</td>
<td>42</td>
</tr>
<tr>
<td>Q3-2014</td>
<td>15</td>
<td>14</td>
<td>30</td>
<td>32</td>
<td>93</td>
<td>92</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td>Q4-2014</td>
<td>13</td>
<td>13</td>
<td>23</td>
<td>25</td>
<td>80</td>
<td>83</td>
<td>31</td>
<td>32</td>
</tr>
<tr>
<td>Q1-2015</td>
<td>9</td>
<td>8</td>
<td>17</td>
<td>14</td>
<td>87</td>
<td>84</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>Q2-2015</td>
<td>15</td>
<td>15</td>
<td>33</td>
<td>34</td>
<td>105</td>
<td>107</td>
<td>47</td>
<td>50</td>
</tr>
<tr>
<td>Q3-2015</td>
<td>16</td>
<td>17</td>
<td>31</td>
<td>34</td>
<td>99</td>
<td>108</td>
<td>43</td>
<td>44</td>
</tr>
<tr>
<td>Q4-2015</td>
<td>13</td>
<td>15</td>
<td>24</td>
<td>26</td>
<td>88</td>
<td>88</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>Q1-2016</td>
<td>11</td>
<td>11</td>
<td>21</td>
<td>21</td>
<td>96</td>
<td>98</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Q2-2016</td>
<td>16</td>
<td>18</td>
<td>34</td>
<td>36</td>
<td>112</td>
<td>117</td>
<td>52</td>
<td>47</td>
</tr>
<tr>
<td>Q3-2016</td>
<td>14</td>
<td>16</td>
<td>32</td>
<td>33</td>
<td>105</td>
<td>109</td>
<td>45</td>
<td>49</td>
</tr>
<tr>
<td>Q4-2016</td>
<td>13</td>
<td>15</td>
<td>25</td>
<td>31</td>
<td>93</td>
<td>98</td>
<td>42</td>
<td>44</td>
</tr>
</tbody>
</table>

| Correlation | 0.94 | 0.96 | 0.96 | 0.95 |
For this case study, Holt-Winters forecasting methods will produce more robust results since they require less information (Winters 1960). In task 5a of Table 7-4, the Holt-Winters forecasting model was applied on permits issued for single-family houses data. The model is appropriate for the construction industry because it accounts for seasonality. The forecasting model is obtained by minimizing the mean absolute percentage error and varying smoothing parameters of the model: alpha, beta, and gamma. The smoothing parameters alpha, beta, and gamma correspond to level, trend, and seasonality as seen below.

1. Alpha: 0.1
2. Beta: 0.1
3. Gamma: 0.1

Based on the analysis, the southern region of the United States has the largest growth in number of permits issued for single-family houses as shown in Figure 7-11.

<table>
<thead>
<tr>
<th>Permit Issued (in Thousands)</th>
<th>Northeast</th>
<th>Midwest</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>13.29</td>
<td>24.00</td>
<td>75.99</td>
<td>29.10</td>
</tr>
<tr>
<td>Trend</td>
<td>0.01</td>
<td>0.25</td>
<td>1.78</td>
<td>1.03</td>
</tr>
</tbody>
</table>

*Figure 7-11. Level and trend for housing permits for single-family houses.*

Here, level represents the average number of permits issued in thousands per quarter for single-family houses, and trend stands for the additional permits issued in thousands per quarter for single-family houses without considering the seasonality effect. There are two types of growth in the construction industry: market growth and growth because of seasonality. In this case study, the decision makers are not interested in growth due to seasonality. It should be noted that the level and trend shown in Figure 7-11 do not represent the growth due to seasonality. Based on the level and trend shown in Figure 7-11, the southern region of the United States is the best-suited market to enter.

Finally, permits issued have been predicted for the next four quarters using the Holt-Winters forecasting model developed earlier. Historical values and forecast values for the model are shown in Figure 7-12.
Intermediate Conclusion 1:

Based on the results achieved in task 5a, the analysis does not fulfill the requirement of the data analytics project as defined in task 1e. However, it was found that the southern region of the United States is the one best suited for business expansion. The following states are included in this southern region: Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia.

To understand the bigger picture of the housing market in the South, decision makers could refer to Figure 7-12, which explains the future trend for permits issued for single-family houses. However, the decision makers were interested in drilling down and learning more about the best-suited cities or counties for business expansion. Now, based on the protocol defined in Table 7-4, we have a good background on the problem. Therefore, we do not need to reapply the protocol from Phase 1. However, we are still not sure what information and which analytical models are
required for the analysis. Therefore, we cannot fulfill the expected outcome of the Phase 2. Therefore, we need to apply the protocol again beginning at Phase 2 as per Table 7-5.

Table 7-5: Implementation of the Protocol to Determine the Housing Market Outlook and Assess Suitable Locations for Business Expansion (Part 2)

<table>
<thead>
<tr>
<th>Task #</th>
<th>Task Name</th>
<th>Description of the Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 2</td>
<td>Design: Understanding the problem</td>
<td>Key Players: The analyst</td>
</tr>
</tbody>
</table>
| 2a     | Exploratory Analysis | Explore the available data and understand what information can be extracted.  
- Explore publicly available data sets such as household growth rate, median home value, crime index, and unemployment rate.  
(The household growth rate should cover a long period, and data on other variables should be the latest available.) |
| 2b     | Problem Structure | Translate the business questions into analytical questions.  
- Find best-suited cities or counties for business expansion based on the data mentioned in task 2a. |
| 2c     | Analyses Overview | Determine what types of analyses are required.  
- No analysis or modeling is required to identify the best cities because this can be done using the layering technique as per task 4c of this table. |
| Task 3 | Development-Understanding the data | Key Players: The analyst, data engineers, and decision maker |
| 3a     | Data Quality | Identify the quality of the existing data.  
- Completeness? - Yes.  
(There are some null values in Nevada. However, it will not affect the process of understanding the data because we are only interested in the southern states.) |
| 3b | Data Structure Needed | Decide on the arrangement of your data.  
Data variables needed  
  - Average annual growth rate of households  
  - Median home value  
  - Total crime index  
  - Unemployment rate  
    (We need these variables for the south only.)  
Data variables available  
  - Average annual growth rate of households  
  - Median home value  
  - Total crime rate  
  - Unemployment rate  
    (We have these variables for all the regions of the United States, including the Northeast, Midwest, South, and West.)  
Additional data strings needed  
  - In this case, we do not need any additional data because we already have the required data. |
| 3c | Data Sources | Identify the sources of additional data to be collected.  
  - Not needed.  
    (There is no need for an additional data string as explained in section 3b.) |
| 3d | Data Collection | Identify the constraints to capture the data or final data limitations. |
• Required data is available from public sources. Therefore, there is no need for additional data collection, and there are no constraints on the data collection.

Supplementary Information

Decide on what additional information is required to address queries/questions.

• There is a need for additional information about each variable to select suitable cities or counties for business expansion. Therefore, the decision makers defined their requirements for each variable as seen below:
  1) Average annual growth rate of households: > 2%
  2) Median home value: > $150,000
  3) Total crime index: < 100 (Here 100 is the national average and 120 means the crime rate is 20% more than the national average.)
  4) Unemployment Rate: < 4.3% (Lower than the national average.)

Data Cleaning

• Check header.
  (Column headers are perfectly labelled, there are no problems with the data type, and categories are properly labelled.)
• Check data types.
  (All the data types are perfectly defined.)
• Check category labelling.
  (All the categories in all the columns are labelled properly.)

Task 4 Refinement- Data refinement

Key Players: The analysts and data engineers
| 4a       | Slicing and Dicing (Data selection) | Ask business questions to identify important data attributes.  
|          |                                  | - Regions: Select [South]  
|          |                                  | (We are only interested in the southern part of the United States. Therefore, we should slice the southern region out of total four regions as defined in task 3b of this table.) |

| 4b       | Splicing (Connecting different data sets) | What different data sources should be required to address the business problem?  
|          |                                  | - None.  
|          |                                  | (Although we have data from various sources, the data in this case is not interrelated.) |
|          |                                  | How should we merge or combine these data sources?  
|          |                                  | - Not required. |

| 4c       | Layering (Mapping data) | Identify different layers of information to find additional information.  
|          |                                  | - In this case, if we layer all the available layers as shown in Figures 7-13, 7-14, 7-15, and 7-16 with criteria defined in task 3e of this table, then we will be able to pull additional information from the result, as shown in Figure 7-17. The additional information is translated to the best counties for housing development. |

| Task 5   | Analyses- Data analyses and modeling | Key Players: The analyst |
| 5a       | Data Analysis | What types of analysis (descriptive, diagnostic, predictive, or prescriptive) and methods are suitable to address the business problem?  
|          |                                  | - After the layering process, a report with the counties best suited for business expansion is |
prepared. This report includes the average growth rate, median home value, total crime index, and unemployment rate for each county as shown in Figure 7-18 and can be prepared using descriptive analytics. Later, an analyst can provide recommendations on a business strategy to enter the new market. This is a part of prescriptive analytics because the analyst is providing an optimal solution or recommendations.

<table>
<thead>
<tr>
<th>Task 6</th>
<th>Outcome-Interpretations and outcomes</th>
<th>Key Players: The analyst, decision maker, and end users</th>
</tr>
</thead>
<tbody>
<tr>
<td>6a</td>
<td>Deployment</td>
<td>Report, dashboard, predictive model, or recommendations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Based on the report shown in Figure 7-18, the decision maker can choose the best county based on land availability or the company’s business strategy.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• After discussions with the decision maker, it was concluded that the company is interested in the counties that have high median home values. The decision maker thinks that the company is providing high-class amenities in their current market, so if they focus on delivering their amenities to other counties with high median home values, they may gain a competitive advantage in the industry.</td>
</tr>
</tbody>
</table>
Figure 7-13 covers the average household growth rate for the South. The represented growth rate is the average growth rate between the years 2010 and 2017. This growth rate, shown in Figure 7-13, varies from -2.31% to 4.55%. However, the decision maker was only interested in counties that have average household growth rates of more than 2%.

*Figure 7-13. Average household growth rates (2010-2017) at the county level.*
Figure 7-14 covers the median home value for the South. The represented median home value is for the year 2017. The median home value in Figure 7-14 varies from $36,000 to $1 million. However, the decision maker was only interested in counties with a median home value of more than $150,000.

*Figure 7-14. Median home values at the county level.*
Figure 7-15 covers the total crime index for the South. The represented total crime index value is for the year 2017. The total crime index value in Figure 7-15 varies from 11 to 268. In this case, the total crime index includes all types of crime, including murder, rape, robbery, assault, burglary, and motor vehicle theft. Here, 100 is the national average and 120 means the crime index is 20% more than the national average. However, the decision maker was only interested in counties where the total crime index was less than 100.

*Figure 7-15. Total crime indexes at the county level.*
Figure 7-16 covers the unemployment rate for the South. The represented unemployment rate is for the year 2017. The unemployment rate in Figure 7-16 varies from 0% to 25.3%. However, the decision maker was only interested in counties where the unemployment rate was less than the national unemployment rate of 4.3%.

Figures 7-13, 7-14, 7-15, and 7-16 represent the average annual household growth rate, median home value, total crime index, and unemployment rate, respectively. However, these individual pieces of information are not very to real estate or land developers since they are interested in buying properties that fulfill all the criteria as a whole. The criteria defined by the decision maker in this case was the following:

1. Average annual growth rate of households: > 2%
2. Median home value: > $150,000
3. Total crime index: < 100
4. Unemployment rate: < 4.3%
Once the layering process was implemented as described in task 4c of Table 7-5, we obtained information on the best-suited counties for housing development as highlighted in Figure 7-17.

![Map of United States showing best suited counties for housing development after the layering process.](image)

*Figure 7-17. Best suited counties for housing development after the layering process.*

Based on intermediate conclusion 1, we are only interested in the South. As per task 6a of Table 7-5, the decision maker is looking for a report so that the company can make data-driven decisions. Therefore, Figure 7-18 presents all of the counties that fulfill their criteria.
### Table 7-4: Growth Rate and Housing Indicators for Selected Counties in the South

<table>
<thead>
<tr>
<th>Counties</th>
<th>2010-2017 Growth Rate: Households</th>
<th>2017 Median Home Value</th>
<th>2017 Total Crime Index</th>
<th>2017 Unemployment Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benton County, AR</td>
<td>2.2%</td>
<td>$173,098</td>
<td>84</td>
<td>2%</td>
</tr>
<tr>
<td>Canadian County, OK</td>
<td>2.56%</td>
<td>$164,106</td>
<td>92</td>
<td>4.3%</td>
</tr>
<tr>
<td>Collin County, TX</td>
<td>2.55%</td>
<td>$248,837</td>
<td>64</td>
<td>3.6%</td>
</tr>
<tr>
<td>Columbia County, GA</td>
<td>2.99%</td>
<td>$188,722</td>
<td>66</td>
<td>4.2%</td>
</tr>
<tr>
<td>Comal County, TX</td>
<td>3.19%</td>
<td>$268,029</td>
<td>97</td>
<td>4.1%</td>
</tr>
<tr>
<td>Denton County, TX</td>
<td>2.7%</td>
<td>$232,386</td>
<td>71</td>
<td>4.1%</td>
</tr>
<tr>
<td>Effingham County, GA</td>
<td>2.05%</td>
<td>$172,321</td>
<td>61</td>
<td>4%</td>
</tr>
<tr>
<td>Forsyth County, GA</td>
<td>3.59%</td>
<td>$304,847</td>
<td>68</td>
<td>3.7%</td>
</tr>
<tr>
<td>Kendall County, TX</td>
<td>3.71%</td>
<td>$354,545</td>
<td>64</td>
<td>2.6%</td>
</tr>
<tr>
<td>Loudoun County, VA</td>
<td>3.07%</td>
<td>$501,962</td>
<td>55</td>
<td>2.3%</td>
</tr>
<tr>
<td>McClain County, OK</td>
<td>2.04%</td>
<td>$168,625</td>
<td>79</td>
<td>2.9%</td>
</tr>
<tr>
<td>Rockwall County, TX</td>
<td>2.53%</td>
<td>$233,960</td>
<td>57</td>
<td>4.3%</td>
</tr>
<tr>
<td>St. Johns County, FL</td>
<td>3%</td>
<td>$294,971</td>
<td>84</td>
<td>4.3%</td>
</tr>
<tr>
<td>Williamson County, TN</td>
<td>2.66%</td>
<td>$373,111</td>
<td>37</td>
<td>2.9%</td>
</tr>
<tr>
<td>Wilson County, TN</td>
<td>2.39%</td>
<td>$224,150</td>
<td>79</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

**Figure 7-18.** Report on the best-suited counties for housing development in the South.

**Final Conclusion:**

Based on task 1e of Table 7-4, the scope of the analytics project is limited to finding suitable counties for real estate development. Figure 7-11 shows the trend for new housing permits, and Figure 7-12 shows the prediction for housing permits in the future in the South. However, this information was not sufficient for the decision maker. The company was interested in drilling down the information and finding the best cities or counties for its business expansion. Figure 7-18 presents the best-suited counties to expand the business in. However, which county is ultimately chosen from the proposed counties in Figure 7-18 should be decided based on the company’s business strategy. Figure 7-18 represents suitable counties based only on data. In the real world, business expansion will likely depend on other factors as well. The construction industry is continually facing challenges, so it is important to consider social, economic, and political factors when expanding a business. Therefore, the decision maker should consider the factors listed below before entering the new market:

1. the cost of entering the new region,
2. differences in construction techniques compared to those in their current business locations,
3. their familiarity with the new region or county,
4. the availability of land, and
5. strategic investment locations.
8. CONCLUSIONS

8.1 Summary

There is an abundant amount of data present in today’s business environment. Fast decision-making has become important, so relying only on data management approaches has become redundant. By using the proposed protocol, organizations can focus on specific subject areas. This protocol can also help them make the best use of any available data. The protocol will guide organizations in the construction industry on how to implement analytics. Organizations can understand the value chain of their processes and study an integrated view of their operations using the protocol. Moreover, it will help them develop their business strategies and transform their businesses using actionable insights.

The traditional method of data collection and analysis often requires more resources than analytics and additional work. Using this protocol, organizations can eliminate the need for intensive data collection, eliminate data redundancy, minimize data infrastructure, and improve business processes. Understanding the structure of this protocol enables executives to design, produce, deploy, and operate analytical thinking within their organizations to deliver high-value services to their customers or clients.

Because construction projects are unique in nature, the underlying problems they face are also distinctive. The proposed protocol supports effective data-driven decision-making. It enables an analyst to reuse knowledge and makes it easier to solve similar problems in the future. Analysts can use this protocol and their experiences to understand a problem and design the analytical solutions themselves.

It must be acknowledged that SMEs perform the majority of the work in the construction industry. Since the amount of data available is increasing rapidly, analytics has become a buzzword everywhere. This has become a problem for SMEs because they have various constraints in their business operations (Robert et al. 2006). Implementing analytics is an additional challenge for them. On the other hand, big organizations are already leveraging analytics in various fields and increasing their competition with SMEs. Coleman et al. (2016) explored some of the challenges SMEs face when implementing analytics:
1. Lack of understanding: Because of the initial investment, SMEs are not sure whether data analytics can really help them. They are also not sure whether they have the right data for analysis. This protocol can serve as a guideline to all the SMEs in the construction industry. Using this protocol, SMEs can identify their important data, which in turn will help them create value in their business processes and make fact-based decisions.

2. Domain specialist: Many SMEs in the construction industry operate by working on specific jobs such as maintenance, lighting, plumbing, or concreting. Due to the narrow scope of their work, they are still not sure how analytics can help them in their field. This protocol has been purposely designed in a general manner so that it can be applied in any field in the construction industry.

3. Lack of business cases and management models: Although there are some examples to study, most analytics business cases do not apply to SMEs. This protocol and the case studies presented in Chapter 7 serve as relevant examples of analytics.

The proposed protocol stands out because it paves a path for SMEs to implement analytics within their own organizations. This protocol facilitates information accessibility for organizations and can contribute to making them more competitive in their domains. Finally, implementing this protocol will likely reduce various forms of risk found in construction projects in addition to those risks found in an organization. The following are some of common risks in the construction industry:

1. business risks,
2. market risks,
3. system integration risks,
4. decision-making risks, and
5. analytics and technology implementation risks.

Decision makers and project execution teams develop goals, objectives, and strategies during the front-end planning process to minimize construction project risks. In this vein, analytics can be closely linked to the front-end planning process of a project in order to minimize the abovementioned. Construction organizations require sensitive information at a fast pace because
of the dynamic nature of construction projects (Golparvar-Fard et al. 2013; Jussila et al. 2012). This protocol enables them to gather information quickly using a consistent approach. Finally, this protocol can be implemented on the process level, project level, or company level.

Effective execution of processes and methods enhances project performance and can lead to the establishment of best practices. This protocol can make it easier for organizations to overcome their data analytics challenges and facilitate decision-making processes. It can be applied across any organization or project in the construction industry to achieve sophisticated business results. Moreover, the protocol can serve as a best practice for analytics in the construction industry because it provides a consistent application of analytics.

Currently, construction organizations are capturing different types of data such as financial data, documents, schedule elements, and weather data. This data is often used in multiple ways for tasks such as estimation, costing, planning, scheduling, project controls, materials management, and procurement. The explosion of data has led to the creation of the field known as informatics, which focuses on data collection, scanning, and data preparation. However, the construction industry is still facing various challenges with making robust decisions using data. Since many other industries are successfully using analytics to improve their decision-making, we believe that implementing analytics will benefit the construction industry as well. Analytics focuses on getting useful information and manipulating it to make sound decisions. However, decision makers in the construction industry are still not sure how to use data effectively because data for a given project in the industry is collected by multiple entities and appears in numerous formats (Karan et al. 2015; Hammad et al. 2014; Azhar 2005). This protocol has been created to provide leaders in the industry with a collaborative approach to making robust decisions.

For example, in a construction company, the estimating team may make bids for the project while the project management team is capturing information on the production rate in the field. In this case, both the teams require data from each other. The estimating team might be interested in getting the actual production rate on the field so that they can prepare robust estimates for future bids. Meanwhile, the project management team might be interested in obtaining the estimated bid data for the project so that they can compare their production rate to data and get a clearer picture about the profit or loss on a job. This example demonstrates that different types of data are captured or generated by various entities in the construction industry.
Every entity is busy doing its day-to-day job, but in order to maintain a competitive advantage in the industry, an analyst should strive to understand business problems and seek solutions for them using the proposed protocol. Because decision makers in the construction industry primarily rely on their instincts or experiences instead of looking at data, the protocol can improve businesses process by providing concrete information to make sound choices with. This protocol integrates the knowledge of decision makers with useful information derived from data. Phase 1 of the protocol gives the analyst an opportunity to comprehend a problem by interacting with decision makers and looking at the issue from their perspective. Phase 6 of the protocol makes sure that the analyst communicates any useful information to decision makers so that an analytics project will see successful implementation. Both Phases 1 and 6 draw on the experiential knowledge of decision makers to make the most of any derived data. However, the success of any protocol depends on its users and the active involvement of its stakeholders. In the same way, the successful implementation of any analytics project depends on active participation from analysts and user involvement.

The flexible and interactive approach to data refinement explained in Section 4.4.4 allows users to get the right data and information for analytics. In order to survive in today’s competitive market, it is essential that organizations innovate and expand their current capabilities. There are many factors that can affect the performance of an organization, and market pressure, global reach, and organizational size are just a few of them (Zeid 2014). Therefore, organizations should create their business strategies based on these factors. The protocol takes an integrated approach that enables organizations to gain a competitive advantage in the market by leveraging information. Applying this protocol and developing an analytical culture can help organizations:

1. find critical areas of the business to consider,
2. build robust data management abilities, and
3. obtain measurable outcomes for the business.

8.2 Research Contributions

Fast decision-making is essential in the construction industry because of its dynamic nature (Vasenev et al. 2014; Ren et al. 2013). Thus, it is unsurprisingly that the industry’s inability to efficiently retrieve information is a big problem standing in the way of its growth (Chettupuzha et al. 2015; Karan et al. 2015; Skibniewski and Golparvar-Fard 2015). This protocol can provide
information in an efficient manner with the help of data refinement techniques. Because multiple entities and stakeholders are involved in a single construction project, the decisions made by a single party will affect other parties associated with the project (Ren et al. 2013). Therefore, this protocol enables an analyst and decision makers to identify the relevant parties in the decision-making process by asking some fundamental questions. Identifying project parties makes it easier to keep everybody in the loop, receive feedback from people, and obtain information from different sources.

To implement analytics in any organization, it is essential to first build an analytics culture. This type of culture can be built by encouraging a systematic approach to processes and keeping everybody on the same platform (Coleman et al. 2016; Holsapple et al. 2014). The proposed protocol makes it possible to ask robust questions related to data analytics so that sound decisions can be made with any resulting answers. Most construction organizations today lack people who can speak both technical and business language, which has created walls between different departments in a given organization. This protocol can act as a common language to bridge this gap and break down the barriers between different information silos within an organization. Moreover, the protocol can be considered as a rule of thumb to guide organizations in reducing time spent on decision-making and finding the best possible answers. Companies using this protocol should keep the following statement in mind: “Think big, start small, and deliver fast.” (Laursen and Thorlund 2016, p. 166) The protocol can help organizations because it does all of the following:

1. explains how to leverage available data,
2. emphasizes solving business problems and creating value for the organization,
3. helps analysts and decision makers ask robust questions,
4. fills the data literacy gap in organizations by acting as a common language,
5. provides a structured approach to build an analytical culture inside organizations, and
6. remains flexible enough to be implemented on an organizational, project, or process level.

Finally, this protocol contributes to the existing body of knowledge by proposing a systematic way to implement analytics. Slicing, dicing and splicing techniques have been implemented before, but to the best of our knowledge, nobody has discussed how data can be
represented in layers and how extra information can be obtained using layering. This research presents the implementation of the layering process in analytics, and the protocol uses available data to create vectors of important data.

### 8.3 Limitations of the Research

The splicing technique defined in the data refinement process has been explained using a relational database. While it is true that splicing focuses on connecting different data sets using a common variable, other types of databases exist, including object relational databases, NoSQL databases, and multimedia databases. It is important to note that all these different types of databases have unique characteristics. For example, it is not necessary to define a schema in a NoSQL database. Hence, it is essential for an analyst to be aware of the different types of databases because construction companies collect diverse sets of data such as videos, images, and location-based data, and it is not possible to store all types of data in a relational database. This thesis paves the way for other researchers to implement the splicing technique. In the future, researchers can extend it to other types of databases as well.

### 8.4 Recommendations for Future Research on Analytics

The key to implementing analytics in any organization is to identify potential areas for its use and then ask challenging questions (Coleman et al. 2016). This research work has already made it possible to more readily implement analytics by creating an analytics protocol, thus helping analysts ask stimulating questions. The protocol builds an analytical road map for organizations to follow. However, it is necessary to also explore possible situations where analytics can help industry leaders, decision makers, analysts, and researchers implement the protocol. Although two case studies were presented in Chapter 7, more case studies from the construction industry are needed for the successful implementation of analytics. This protocol can be used as a tool in future case studies as well.

Martínez-Rojas et al. (2015) say that construction processes have lots of uncertainties that can come from a variety of sources. The sources they identify include the following:

1. uncertainty of labor productivity due to task complexity, location, and dynamic team nature;
2. stock-out of resources or materials;
3. material arriving on the site behind schedule;
4. adverse weather conditions; and
5. the absence of workers.

Effective decision-making requires performance management visibility, which can help an organization improve cost control, risk management, and the feasibility of its projects. Innovation can differentiate an organization from its competitors and help it achieve a competitive advantage. Stubbs (2013) states that survival analysis, a technique used in the medical field to understand the impact of different treatments in order to delay death, has been used in the telecommunications field to predict churn time. This is a good example of innovation, and this technique could also be useful to the construction industry when it comes to predicting when equipment may break down.

An operational excellence approach is required in capital-heavy businesses that have high initial investments in order to see good returns on investment (Laursen and Thorlund 2016). We believe that analytics can improve construction operations since the efficient delivery of projects is an influential factor in gaining a competitive advantage in the industry. Hence, construction companies should invest in technology and analytics to improve their operations and business processes. Examples of operational excellence include:

1. cost control,
2. retaining market share in the declining market,
3. improving internal processes,
4. identifying where more resources are required to increase organizational competency,
5. bidding more competitively,
6. increasing safety to strengthen the value of an organization,
7. improving quality, and
8. using innovative materials and techniques to save time.

Various companies are already using analytics for their planning and the optimization of their project workforce (Watson 2011). Using simulation models, companies can determine the supply and demand for specific types of workers, and as conditions change, the companies revisit their hiring and retention plans. The construction industry is a labor-intensive industry that needs
multiple workers with specific skills in fields such as carpentry, electrical work, mechanical work, and masonry. Moreover, it is a seasonal industry, so people are regularly hired and laid off. Therefore, analytics has the potential to create value in the construction industry by putting existing data to good use. Researchers in construction analytics have many different topics they can choose to study.
REFERENCES


Marr, B. (2016). *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. John Wiley & Sons, Ltd, Chichester, UK.


Shahandashti, S. M., Razavi, S. N. S. N., Soibelman, L., Berges, M., Caldas, C. H., Brilakis, I.,
Approaches and Applications for Construction Engineering.” Journal of Construction
Engineering and Management, 137(10), 863–869.

Business Process.” Proceedings - 17th IEEE Conference on Business Informatics, CBI 2015,
2, 49–54.

Perspectives.” Proceedings of the ITI 2013 35th International Conference on
INFORMATION TECHNOLOGY INTERFACES, University Computing Centre - SRCE,
Zagreb, 3–8.

completion time for Malaysian construction projects.” Built Environment Project and Asset
Management, 5(1), 52–68.

Engineering.” (December).


Enterprise Resource Planning Systems in Engineering Construction Firms.” Journal of
Construction Engineering and Management, 135(10), 965–978.


Walker, D. H. T. (2016). “Reflecting on 10 years of focus on innovation, organisational learning and knowledge management literature in a construction project management context.” *Construction Innovation*, 16(2), 114–126.


