Univerza *v Ljubljani*
Fakulteta *za gradbeništvo in geodezijo*

JOÃO BRÍGIDO RIBEIRO GONÇALVES VIEIRA

BUSINESS INTELLIGENCE DATA CENTER FOR THE CONSTRUCTION INDUSTRY

PODATKOVNI CENTER ZA POSLOVNO INTELIGENCO V GRADBENIŠTVU

BIM A+ European Master in Building Information Modelling

Master thesis No.:

Supervisor:
Assist. Prof. Tomo Cerovšek, Ph.D.

Ljubljana, 2020

Co-funded by the Erasmus+ Programme of the European Union

ERRATA

Page	Line	Error	Correction

»This page is intentionally blank«

BIBLIOGRAFSKO – DOKUMENTACIJSKA STRAN IN IZVLEČEK

UDK:	**004.9:624(043.3)**
Avtor:	**João Brígido Ribeiro Gonçalves Vieira**
Mentor:	**doc. dr. Tomo Cerovšek**
Somentor:	
Naslov:	**Podatkovni center za poslovno inteligenco v gradbeništvu**
Tip dokumenta:	**magistrsko delo**
Obseg in oprema:	**43 str., 9 slik, 7 preglednic**
Ključne besede:	**poslovna inteligenca, podatkovna skladišča, BIM, optimazacija terminskih planov, vizualizacija podatkov, arhitektura in inženirstvo**

Izvleček:

Arhitektura, inženirstvo in gradbeništvo v zadnjih desetljih beležijo nizko rast produktivnosti. Z eksponentno rastjo količine podatkov želi gradbeni sektor vzpostaviti stik z boljšim upravljanjem projektnih informacij, kar izpostavlja in udejanja informacijsko modeliranje zgradb (angl. building information model - BIM).

Obstaja širok nabor metodologij in protokolov za upravljanje s podatki, ki jih lahko obravnavamo pod pojmom poslovne inteligence. Uporabnost in razvoj tega področja v gradbeništvu je še vedno v začetnih fazah. Pristopi poslovne inteligence so bili zasnovani za druge industrije, njihove značilnosti in izzive. Uporaba poslovne inteligence v točno določeni industriji, kot je gradbeništvo, zato ahteva ustrezne prilagoditve, še posebej v primeru če poslovni inteligenci temelji na pristopu BIM.

Pričujoče delo obravnava sinergije, ki jih je mogoče doseči pri uporabi poslovne inteligence (angl. Business intelligence - BI) in uporabi podatkov iz modelov BIM, pri čemer želimo zatgotoviti prilagojene rezultate glede na posebnosti gradbene industrije. Študija vključuje pregled literature, na osnovi katere je predlagana zasnova podatkovnega centra za poslovno inteligenco v gradbeni industriji. Temeljit pregled rešitev vodi do konceptov, ki nudijo okvir za vspostavitev procedur za boljše upravljanje projektnih informacij v gradbeništvu.

»This page is intentionally blank«

BIBLIOGRAPHIC– DOKUMENTALISTIC INFORMATION AND ABSTRACT

UDC:	**004.9:624(043.3)**
Author:	**João Brígido Ribeiro Gonçalves Vieira**
Supervisor:	**Assist. Prof. Tomo Cerovšek, Ph.D.**
Cosupervisor:	
Title:	**Business Intelligence Data Center for the Construction Industry**
Document type:	**Master Thesis**
Scope and tools:	**43 p., 9 fig., 7 tab.**
Keywords:	**Business Intelligence, Data Warehouse, BIM, Schedule optimization, Data Visualization, AEC Industry**

Abstract:

The Architecture, Engineering, and Construction Industry have seen a relatively low increase in the efficiency rates over the recent decades, and the data management is still trying to catch up with the exponential data growth in the sector, both emphasized and embodied by BIM. There is a plethora of methodologies and routines for data management kept under the Business Intelligence umbrella; however, the usability of these tools in the AEC Industry is under development. These concepts were design for other sectors, their characteristics, and challenges, to adopt such tools require tailoring and should embrace the leading trend in the industry: BIM. This study looks at the synergy that is possible due to the overlapping benefits in applying established BI's technologies and BIM data while assuring a tailored result by putting the peculiarities of the business in the spotlight while building around it. The study is a Literature review that aims to develop a routine for implementing a Business Intelligence Data Center for the Construction Industry. It looks at past research studies that have investigated similar issues for other industries. The sharp look at these studies shows how these two concepts can be combined and, then, proposes a data management routine for more integrated information flow in construction.

»This page is intentionally blank«

ACKNOWLEDGMENTS

I thank my parents, who have given me the education to participate in this master and who provided direction to see through uncertainties. I also thank RSC for the inspiration, without who I would never engage such an ambitious journey. My heartfelt appreciation also goes out to my family and friends who stood by me through all this time.

»This page is intentionally blank«

TABLE OF CONTENTS

INDEX OF FIGURES

1 INTRODUCTION

Currently, the AEC Industry is trying to catch up with the digital revolution. There is plenty of room for improving, according to the Committee for European Construction Equipment (CECE) the civil engineering and construction sector is among the world's least digitized. The gains of documents and processes digitalization have proven to be worth the investment by increasing productivity in other markets. While the total economy productivity increased by almost 25% since the '90s, the AEC industry is struggling to reach double digits [1].

The challenges of digitalization for this industry are mainly: lack of replication, decentralization, transience (constant change), and fragmentation [2]. These particularities of the sector demanded specific tools and methodologies which are fully developed nowadays. Such solutions are attempting to bridge the gap between industries aiming to leverage the AEC Industry productivity to the same level as the entire economy; by doing so, a value estimated at $1.6 trillion could be generated worldwide.

Digitalization challenges common to all industries include unclear definitions of what digital means, a vague idea about what the transformation should accomplish, and poor integration of digital tools with business processes [3]. A survey in the UK [4] about the reasons for not implementing BIM listed lack of expertise and cultural resistance at the top.

These defying factors can be found during the whole building life cycle. Still, working on-site makes them even harder to handle at the construction phase, which is correlated to the Building Information Modelling (BIM) adoption gap between the design segment and the construction [4]. The endorsement is the majority of research on the implementation and use of BIM focus on building design and pre-construction planning[5]. Besides, the fact that there is a plethora of detailed tutorials and workflows available for creating and handling information during the design phase, but when it comes to construction, the material available gets vague.

A significant disincentive is the way information is handover to construction. A survey performed in the UK showed that, more often than not, only 2D documents are handed to construction. Also, the rate of design and pre-construction teams that use BIM but do not supply it to the construction is around 40% [4]. Static information is bound to lose its integrity by getting obsolete as the design develops from as-design to as-built, so even if the design passes on the models, the construction company must continue its development.

Which leads to the proposed problem: how to empower data-driven culture on construction to optimize cost estimation and schedules?

For non-design disciplines stakeholders such as contractors and project managers, BIM is more like an intelligent Database Management System with the advantages of visualization and navigation [6]. While elaborating a solution, it is crucial to observe one shared achievement of engineering and construction companies that successfully implemented digital technologies: linking different projects to unlock impact across the enterprise [2].

Fundamentally, adequate data storage allows five basic operations: collecting data, cleaning & transforming data, storing data, maintaining speed & performance, accessing & analyzing data. Since data storage is a thesis on its own, this study will go as deep as needed in each topic.

The current study will take into consideration all the challenges presented above to answer the proposed problem. The creation of a Business Intelligence data center for the Construction Industry will follow these steps: introducing a routine for construction progress measurement (collecting); handling interoperability (cleaning & transforming) while creating the structure for a dataset of historical information (storing); applying probability to actual production rate and price list to optimize the schedule (analyzing). Finally, since all this information must be made “human-readable,” this study will develop a visualization scheme to assist decision making to overcome construction’s setbacks (accessing).

1.1 Research objectives

The aim of the paper leads to the research question: how to empower data-driven culture on construction to optimize cost estimation and schedules. BIM has successfully inspired the data-asset feature in the design phase and is on the path to do the same on the whole building life cycle. However, the current approach for 4D and 5D is often based on closed BIM, which means the integration between software is limited, and to unlock the full potential of a tool, all data must be following its storage protocol. This study, therefore, explores how to bypass these interoperability constraints to get the most from the whole spectrum of construction data without software limitations and propose a routine for implementation. To answer the research question and to facilitate the study, the following objectives are set:

- Investigate data storage architecture and the benefits of applying it in historical data analyses;

- Propose a routine for digitalizing the Construction Progress Monitoring activity;

- Study the information required, how to harvest the data, and how to handle interoperability between BIM, Cost, Schedule, and Resources;

- Suggest ways to best visually display the information.

1.2 Methodology

The research carried out is a deductive study, reviewing the previous studies carried out by different scholars related to construction data management, that include or not BIM, and papers related to analyses of construction progress and how to display this information. The studies are from different resources that include; peer-reviewed journals, conference proceedings, past written master's thesis, reports, and textbooks. These sources were selected based on two main criteria, the first being that they contain the relevant information in the study area, and the second criteria is the studies where from highly ranked sources. Highly rated sources meaning that they were commonly cited in various sources. By the end of the study, the BEXEL Manager software was used as a benchmark to validate the proposed Business Intelligence routine

1.3 Scope and Limitations

The research is limited to the use of BIM in relational database management systems, data warehousing applications in the construction phase, and development of analyses and visualization schemas for the structured data.

A limitation of this study is that it will only look at previous studies for individual topics due to the unavailability of information about how different software architecture data. To overcome this obstacle, BEXEL Manager software was assessed and used as benchmark.

The study is also limited by short time meaning; therefore, priority is given to practical application at the expense of the depth of technical concepts enabling the routine proposal.

Another limitation is the case study; due high level of commitment required for implementing it and the extensive range of processes that would be affected, it was not possible to find a company willing to engage fully. To do it without total commitment would compromise the validation propose of the case study.

2 LITERATURE REVIEW

The present section addresses these four topics: **Error! Reference source not found.**, **Error! Reference source not found.**, Schedule Optimization, **Error! Reference source not found.**. The first two focus on finding solutions to problems related to structured data production, while the last two are about data consumption, visualization, and insight delivery. Most Business Intelligence resources came from solving similar problems in different sectors. The strategy to tailor BI to the features and solutions that make sense for the AEC Industry is to hold the challenges in the spotlight as a central point of research.

2.1 Historical Data Structure

When developing data storage architecture, one must start by listing the business needs, only with a clear definition of its demands, the solution can be obtained [7]. The development methodology consists of working backward, start from the request, and finish at selecting tools and technologies. This section will follow the same path: first, focus on the useful features of historical data; second, it will choose which business process must be engaged to allow these analyses and how to describe this event; third, what are the dimensions related to the selected business process; fourth, shine a light on different ways to structure the information to make it viable.

There are many possible approaches to the use of historical dataset in construction, presented below are the most advantageous features, in the author's view.

- Generation of key benchmark metrics;
- Cost trend calibration;
- Front-end or strategic level cost estimates and schedules;
- Forecasting critical path problems;
- Performance and productivity management.

One of the primary benefits of a historical construction dataset is to provide **key benchmark metrics** from similar buildings and even from the current construction. Having at hand the actual construction ratios queried in different ways allows the manager to estimate better the cost and duration of each phase of the project. These pieces of information can assist decision making in many ways; for instance, it can be used for reviewing estimates and even for appointing suppliers, subcontractors, and design solutions.

The collected data can also be used to **cost trend calibration**. When labor and material pricing is more volatile than usual, keeping track of the market variation during time can be crucial. The historical dataset can help one decide if these cost fluctuations are a long term trend or merely a short term spike, allowing to calibrate the estimation better.

Creating **front-end cost estimates and schedules** generated by production rates from the information from previous projects. Using the resources and productivity-related data generated by the system, it is possible to prepare precise conceptual-level estimates that will require a minimal amount of time or engineering input.

There are several methodologies for implementing a routine for **forecasting critical path problems,** which results in performance early warnings to the construction manager. It is possible to predict delays by keeping track of each crew's productivity rate and checking if the actual production is meeting the predicted in activities related to the critical path. This forecast provides the data the construction manager requires to adapt on the fly.

Improving **performance and productivity management** promotes the HR department from being a support function to a strategic partner of a company. It can generate outcomes such as rewards, promotions, succession planning, and contracting models. Although the implementation of such policies in the construction sector is challenging [8], having a single source of truth for these metrics provides the means to do so. As stated earlier, the culture of resistance to changes is at the top of the reasons for failure in the age of digital construction, and one of many uses of rewarding performance is behavior change.

This feature encourages digital implementation because it provides short term outcomes: it measures its own results and creates awareness across business sectors, while giving strategic-level information for the executive board, such as the data for elaborating an Integrated Project Delivery contract based on performance.

After exploring the business needs, it is time to select the business process that must be digitalized to able all the previous analyses. To have structured data of what, when, who, where, how, and so forward for each construction task would allow real-time management and provide valuable business intelligence. Construction Progress Monitoring (CPM) is the chosen business process due to the multiple dimensions it combines, and every single task is the event in which the process will be modeled.

The information required to perform those analyses belong to four dimensions: Cost, Schedule, Resources, and BIM (referring to the actual model). The data will most likely be spread across different sources, and it is expected well-structured data stored in several data frames (tables), they may be interconnected or not. It is safe to assume similar conditions to cost (invoice, transaction, payroll, etc.), resources (inventory, equipment, employees, subcontractors, etc.), schedule (start and end date of each task, late start, task dependency, etc.) and BIM (construction elements, quantities takeoff, etc.). They can be stored differently from company to company; therefore, proposing a general procedure or tool for extracting it does not make sense since it must be developed on demand.

When it comes to extracting information for the historical dataset, it is crucial to have in mind what is the desired information, where it is stored, and if it is not digital yet, how to make it. To accomplish this task, one must gather knowledge of the company's files structure, performing interviews with different sectors may be required. The next topic is entirely dedicated to it.

Now, with the selected business process, dimensions, sources, and event organization, it is time to propose the architecture for data storage. The most usual options are shortly defined below, and Table 1 summarizes it.

Relational Database Management System (RDMS): this type of database stores and provides access to data points that are related to one another by keys. An RDMS must be managed in a secure, rule-based, consistent way. As the complexity increases (the number of interconnected tables), they tend to be an unsatisfactory solution for analyzing historical data. Friction will compromise its performance when dealing with large queries on complex databases; the friction created by its structure and the path between data frames will slow the process.

Data Warehouse: it stores structured, curated data pulled from separate intermediary databases. That makes it easy for analysts from different parts of an organization to access and analyze for their respective purposes. Data Warehouses retrieve and store data to have all the relevant information at the same data frame, optimizing storage, and analyses time performance.

Data Lake: it also pulls in data from multiple sources, but generally contains a much wider array of data types, including unstructured data and data from sources outside the organization. Whereas Data Warehouses are designed to be a central data repository for known and specific purposes, data lakes are intended to contain data that might not be useful at the moment but could play into some potential future analysis: the more, the better. They are generally most useful for more exploratory studies undertaken by data scientists and researchers.

The main difference between a Data Warehouse vs. data lake vs. RDMS is that the first is built to hold structured data from multiple sources, the last is used to store and organize structured data from a single source, such as a transactional system. Data lakes differ from both in that they store unstructured, semi-structured, and structured data. [9]

Table 1 - Data Storage Characteristics[9]

CHARACTERISTIC	RDMS	DATA WAREHOUSE	DATA LAKE
Data types	Structured, numerical data, text and dates organized in a relational model	Relational data from transactional systems, operational databases, and applications	Structured and unstructured data from sensors, websites, mobile apps, etc.
Purpose	Transaction processing	Data stored for business intelligence, batch reporting, and data visualization	Big data analytics, machine learning, predictive analysis, and data discovery
Data capture	Data captured from a single source, such as a transaction system	Data captured from multiple relational sources	Data captured from multiple sources that contain various form of data
Data normalization	Uses normalized, static schemas	Denormalized schemas, schema-on-write	Denormalized, schema-on-read
Benefits	Provides consistent data for critical business applications	Historical data from many sources stored in one place, data is classified with the user in mind for accessibility	Data in its native format from diverse device sources gives data scientists flexibility in analysis and model development
Data quality	Data is organized and consistent	Curated data that is centralized and ready for use in BI and analytics	Raw data that may or may not be curated for use

From this explanation of historical data features and structures, it is clear the value of information and why it should be treated as an asset. This topic did not propose new types of performance indicators, measuring information or innovative solutions, leading companies already have almost all the data discussed. The question was what is the best way to organize information to get the most out of it and using RDMS to connect the four dimensions (cost, schedule, resources, and BIM) and storing the information in a Data Warehouse is the answer.

The use of RDMS is interesting because the four dimensions are already digital in most construction companies, so harvesting it from the original format can be automated, and one can create the link between dimensions with foreign keys. The Data Warehouse is proposed as a solution because of its optimum performance for accessing data. Another advantage of this information flow is the division of information into the back end and the front end, given that the training to reach a certain level of expertise demanded to analyze the front end is basic for engineers and interns. The combination of these two solutions is capable of providing structured historical data. Figure 1 illustrates the proposed information flow.

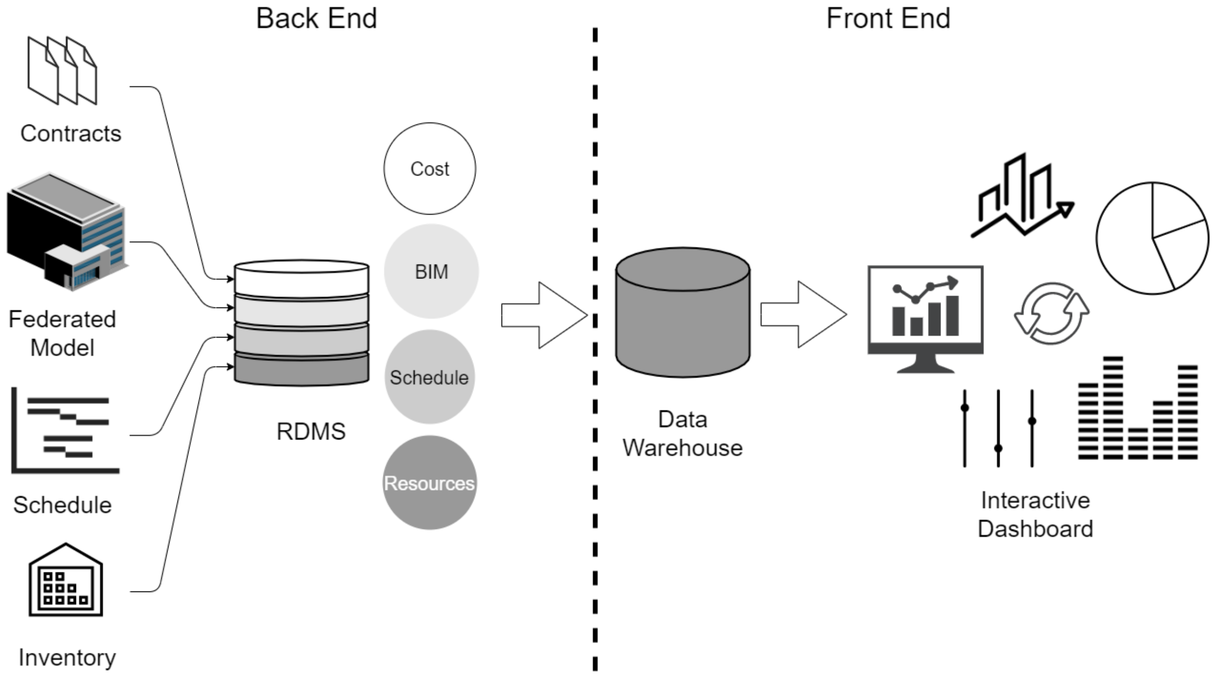


Figure 1 Information Flow

2.2 Construction Progress Monitoring

Defined in the previous section, CPM is the process that must be digitalized. This section will now approach the following questions: what information must be stored to enable the proposed analyses, how the dimensions (Cost, Schedule, Resources, and BIM) can be linked, and how to capture this information and store it in the desired structure.

To choose which pieces of information must be captured in the monitoring process, it is crucial to start from the business needs by reviewing the intended analyses and go through the formulas; by doing so, the desired data will be precise. Table 2-Variables, Data, and Dimensions and Table 3-Analysis' Formulas summarizes the outcomes of this step. Once the data is selected and with comprehension of the company's files structure, mapping the source of each data is elementary.

Best practice says one must try to make the Data Warehouse "query-proof," this is achievable by storing the maximum granularity the business offers so that it can perform well under unforeseen analyses [7]. For instance, when storing costs, instead of storing the total direct cost, collect the cost of each resource even if it seems unnecessary at the moment. The Data Warehouse is capable of managing GB's of information, so splitting one column into two or three will not compromise its performance, go for the smallest breakdown structure.

To define the data to be captured for the analyses intended, this section will follow the same order presented previously:

- Generation of key benchmark metrics;
- Cost trend calibration;
- Front-end or strategic level cost estimates and schedules;
- Forecasting critical path problems;
- Performance and productivity management.

The **key benchmark metrics** for monitoring construction progress were selected from established methodologies in construction: Earned Value Management (EVM) and Last Planner System (LPS). These methodologies were chosen because both are task-based and uses data from the four dimensions. The proposed data management routine does not mandate the adoption of any of these management systems but uses the metric developed by them to generate benchmarks. These metrics will help to monitor the current construction and provide parameters for future projects.

EVM applied to construction provides early warnings of performance problems based on variables that can be captured from the four dimensions. It is used on the cost and schedule control by checking the planned cost and dates against the actual value of them. Called "Data Points", there are plenty for metrics involved in this methodology, they are divided into Primary Data Points and Derived Data Points. They are presented in Table 2 and Table 3. Cost and schedule controlling metrics from this methodology are variance and performance index.

LPS is also a task-based system, and it provides an indicator for how well the planning system is working, presented as the percentage of planned tasks that were completed as scheduled, the metric is named Percent Plan Complete.

Also, a strategic-level analysis was included: gross value added per hour worked. It can only be performed after the asset is liquidated and consists of subtracting the total material and equipment cost from the total revenue and divide this value by the sum of working hours invested. The new information for this analysis is the total revenue, which is not usually openly shared. Therefore, the Data Warehouse will able this calculation by providing the complete material and equipment cost and the sum of working hours, but not the actual gross value added.

For **cost trend calibration**, the metric must show the trend in the cost through time, and it can also show the variation of the actual cost based on the planned cost. The data captured for the previous analyses are sufficient for creating this index if the granularity is downgraded from "total cost per task" to "cost per resource per task" and if the "resource quantity per task" is also captured. The same logic can be applied to material and labor, creating a productivity ratio trend curve.

Engineering and construction companies usually have a total cost number related to the square meter used for **front-end cost estimates,** but to have a similar one for the duration demands more advanced data analyses. With the data collected for the previous analyses, it is possible to enhance this ratio adding the location (story) and its area to each task. The usual formula for it uses the primary unit (e. g. square meter) and the cost, and it is created by interpolating data points of area and cost to build a curve. The granularity captured by the present study allows a more accurate correlation using different filters, for instance, creating one curve per story with multiple points to analyze how the height affects the performance. The same logic can be applied to task duration and productivity.

Both EVM and LPS have indicators to **forecast issues**. The first mainly apply previous productivities ratios to upcoming activities and check if the performance will deliver as planned. These Derived Data Points are: Cost Estimate at Completion (EAC); Cost Variance at Completion (VAC); Estimate to Complete (ETC); To-Complete Performance Index (TCPI).

LPS forecasting is tied to its management methodology, but it can provide insight to the manager even if the LPS is not fully adopted. The forecast metric is based on the critical path and dependency between tasks and not on the productivity ratio. Tasks Made Ready is a measurement of the percent of tasks cleared from dependencies and ready to begin as planned. Table 3 shows the formulas for the selected metrics of both systems.

Productivity is a well-defined concept, and it is widely stated as the ratio of output to input. These ratios can be configured in partial productivities measures using different dimensions as input and outputs. For instance, productivity for roofing can be calculated as $\frac{area}{material\ quantity}$, $\frac{total\ cost}{labor\ hour}$ or $\frac{material\ quantity}{labor\ hour}$ and they are all correct and covered by the data collected for previous analyses.

Performance is a more subjective metric because it relies on the evaluation of how well the task was executed, with the data captured so far, the performance will be resumed to on time and within budget. The cost performance is already divided into each resource and the extraordinary tag can be assigned to each contract. The granularity to the cost metric is appropriate; however, the delay metric requires some improvement. To measure delay one must follow the chronologic of the project and before assessing one task the schedule must be updated with all the previous delays [10], this requires the scheduling tool in use to keep the as-planned schedule always up to date, most software used for it have such functionality.

The analyses to be performed on the delays are duration and late start. The duration analysis demands the user to provide data if the activity has stopped between its start & end date and if so, for why and how long.it was on hold. The late start analysis requires data about when the task should have started, when it was possible to start and when it actually started. The planned start and the actual start are already in the system; the data about when it was possible to start is not a raw data; it is a processed one, meaning it is retrievable from the current data frames. It only takes more processes to capture, the query must be about which of its previous dependable tasks finished last and the date it ended. This means that even though the RDMS will not contain a field for "PossibleStart" for this specific information, it is retrievable through the link of primary and foreign keys. On the other hand, the Data Warehouse will contain a specific column for it, this is the way to speeds up queries, reducing friction to increase processing speed.

It is possible to add one quality indicator to each task. Adding quality indicators is only recommended to companies in which it is already established to limit changes in the current workflow.

Table 2 summarizes the relationship between the methodology's variables, data to be collected, and which dimension is involved in capturing it. Table 3 displays the analyses to be performed with the formulas and variables used to each one [11].

Table 2-Variables, Data, and Dimensions

Variables	Data	Dimensions
Planned Value per Task (PV) Authorized budget for a planned task	PlannedValueMaterial PlannedValueEquipment PlannedValueCrew	Schedule, Cost, Resource Schedule, Cost, Resource Schedule, Cost, Resource
Earned Value per Task (EV) Authorized budget for the partially completed task and which percentage is completed	ActualValueMaterial ActualValueEquipment ActualValueCrew CompletenessPercentage	Schedule, Cost, Resource Schedule, Cost, Resource Schedule, Cost Schedule
Actual Cost per Task (AC) Budget for the completed task, the final value of EV when the task is completed	ActualStart ActualEnd CostType	Schedule Schedule Schedule
Percent Plan Complete (PPC)	PlannedStart PlannedEnd	Schedule Schedule
Tasks Made Ready (TMR)	Depency	Schedule
Material and Labor Cost Index	PlannedQuantityMaterial PlannedQuantityLabor PlannedCrew ActualQuantityMaterial ActualQuantityLabor ActualCrew	Schedule, BIM Schedule, BIM, Resource Schedule Schedule, Resource Schedule Schedule
Delay Analyses Duration and late start metric	HasStopped StopMotive StopDuration	Schedule Schedule Schedule

Table 3-Analysis' Formulas

Analysis	Formula
Schedule Variance (SV)	$EV - PV$
Cost Variance (CV)	$EV - AC$
Schedule Performance Index (SPI)	$\frac{EV}{PV}$
Cost Performance Index (CPI)	$\frac{EV}{AC}$
Cost Budget at Completion (BAC)	$\sum_{present}^{end} \mathrm{PV}$
Duration Budget at Completion (BAC(dur))	$\sum_{start}^{end} (\mathrm{CriticalTaskEnd} - \mathrm{CriticalTaskStart})$
Cost Estimate at Completion (EAC)	$\sum_{start}^{present} \mathrm{EV} + Task(Type1) \sum_{present}^{end} \frac{\mathrm{PV}}{\mathrm{CPI}} + Task(Type2) \sum_{present}^{end} \mathrm{PV}$
Cost Variance at Completion (VAC)	$BAC - EAC$
Estimate to Complete (ETC)	$Task(Type1) \sum_{present}^{end} \frac{\mathrm{PV}}{\mathrm{CPI}} + Task(Type2) \sum_{present}^{end} \mathrm{PV}$
To-Complete Performance Index (TCPI)	$\frac{ETC}{(BAC - \sum_{start}^{present} \mathrm{EV})}$
Percent Plan Complete (PPC)	$\frac{Number\ of\ Task(PlannedEnd \geq ActualEnd)}{Number\ of\ Task(PlannedEnd \geq PresentDate)}$
Tasks Made Ready (TMR)	$\frac{Number\ of\ Task_{(Clear\ of\ dependency)}^{(Present<PlannedStart<GivenDay)}}{Number\ of\ Task(Prensent < PlannedStart < GivenDay)}$

Type1: when previous variances are expected to continue at a similar rate during the rest of the construction. Type2: when the previous variance was caused by an extraordinary event that will most likely not repeat itself.

As established previously, the study will use an RDMS to link the dimensions of cost, schedule, resources, and BIM. Structuring how they are connected is the actual process of creating the RDMS.

For the intended use, it will be enough to normalize it up to the third normal form (3NF). To reach this 3NF level, the database must comply with the following rules:

- Each table cell shall contain a single value;

- Each record shall be unique;
- Each table shall contain one single column for the primary key;
- Tables shall not have transitive functional dependencies.

The way to link dimensions to each other is to establish compatible breakdown structures, for instance, if the schedule model is task-oriented and the BIM model is object-oriented, all objects in the BIM model must have one property relating to a task id, or there must be an intermediate table linking these dimensions, it is up to how the relationship between dimensions and tables: many to many or many to one. Or the other way around, all tasks must have information about the object related to it. The RDMS will not depend on the company's file structure, so it is reasonable to create an architecture for all cases. RDMS will be further developed during the routine proposal; however, Figure 2 presents the final Entity-Relationship Diagram.

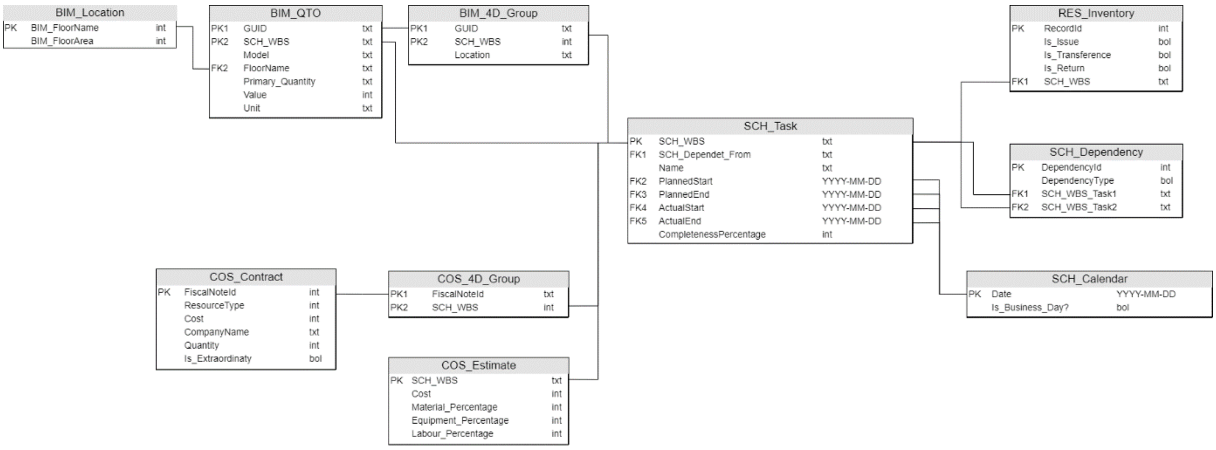


Figure 2 Entity-Relationship Diagram

Finally, it is time to populate RDMS with data from the four dimensions. To capture data from the mapped sources, the information from the four dimensions must be digital and interoperable.

Interoperability is the "ability of two or more systems or components to exchange information and to use the information that has been exchanged" [12]. For the intended use, interoperability means to be able to feed the database system using files as sources without direct human input. For the proposed workflow, it is acceptable to require exporting from the original software to another format. For the proposed routine, interoperability will be limited to the exporting data frames to be loaded to RDMS; this will be thoroughly explored during the routine proposal.

After populating the RDMS, the data must be transferred to the Data Warehouse, to do so one must use the Structured Query Language (SQL) to select the desired information to be added to the Data Warehouse. SQL is similar to a programming language; it was designed to manage data stored in relational databases. SQL operates through simple, declarative statements. This language helps in keeping data accurate and secure. Also, it maintains the integrity of databases, regardless of size.

The digitalization of CPM seems overwhelming at first, but its implementation is divided into the work of each dimension sector, which makes it straightforward and not that laborious. The BIM dimension will have to find a way to automate the extraction of the data frames required from itself and so on and so forward.

2.3 Schedule Optimization

An optimization algorithm is the final step of the scheduling system. It requires previous development, such as modeling the problem and selecting a solution approach to be used for optimizing schedules. All decisions shall suit most conditions and characteristics of the project, and the last one can only be as successful as the earlier phases.

Modeling the scheduling problem has been the focus of a large number of researches for the past decades. These studies are almost reaching a common agreement on how the mathematical model should be, however, the extension of the model depends on the field and its particularities [13]. Depending on the extension the schedule may include different ways of executing the same activity, causing changes in durations and resources (*'Multi-Mode Scheduling'*); splitting activities into stages and assigning resources to each phase (*'Pre-emption Modelling'*); modeling storage constraints for tasks' outputs (*'Cumulative Resources,'* e.g. precast elements).

There are numerous extensions, and the idea of the present study is to optimize the solution approach with historical data and not to improve the current modeling. The most common model used in the construction industry is a rather basic extension of the Critical Path Model. It demands three items: a list of all activities within the project, the duration of each activity, and the precedence relationship between them [14].

This study demanded no further development from the schedule dimension for the Construction Progress Monitoring and will remain the same for this section. This basic version of Critical Path has several shortfalls concerning over-simplifying the characteristics of real-life scheduling problems; however, it is the mainstream solution for the industry and, therefore, is the model chosen for the optimization.

In literature, the solution approaches can be classified under two main categories: static scheduling and dynamic scheduling. These can also be divided into two: deterministic or stochastic project scheduling. The focus of this section will be on a dynamic stochastic scheduling approach. Dynamic means the scheduling solution architecture involves several cycles of static scheduling to be executed based on specific timing or criteria. By definition, one cannot apply an algorithm on a static approach. Stochastic means it uses data and probabilistic analysis instead of mathematical formulas in the process to obtain outputs, such as task durations.

This method of scheduling focuses on generating an updated and realistic schedule by taking into consideration the risk factors that may have a positive or negative impact on the project. It captures risks and uncertainties associated with the project tasks and the overall project without actually going through each risk, but by analyzing the previous tasks patter.

The schedule will be updated periodically, consolidating the completed tasks and updating the estimation for the remaining tasks. The task's order will be preserved, and the dependency between tasks will able to advance or to delay the start date of its successors.

For the probabilistic analysis, the proposed workflow was inspired by the Programmable Evaluation and Review Technique (PERT). The estimation done in PERT is 3-points estimate: Pessimistic (P), Optimistic (O) & Most Likely (M) for each task. In PERT analysis, the duration of an activity is assumed to have a beta probability distribution. According to this distribution, the expected task duration and variance of the task completion is calculated using the formulas presented in Table 4.

Table 4 PERT Formulas

Expected Task Duration	$\frac{O+4M+P}{6}$
Variance	$\frac{P-O}{6}$

This section proposition is to reverse engineer PERT, instead of starting with the P, O & M values to build the beta distribution, it will begin with the distribution to obtain the P, O & M values, using the normal distribution instead of the beta It will use the EVM idea of applying previous productivities ratios to upcoming activities. Still, instead of using the average of the rate, it will use the normal distribution. The Central Limit Theorem dictates that as the number of events increases, the distribution of the sample mean or sum approaches the normal distribution, and it is suitable even if the original variables themselves are not normally distributed.

The workflow is creating the normal curve for the productivity ratio (cost and duration per primary unit) of each type of activity from the Data Warehouse, then calculating the ratio value for the probability of each estimation (P, O & M) from the curve for each task. Applying these values to the schedule and obtaining total value and variability of duration and cost of each estimation scenario.

The algorithm will be used to calculate the probability of completing the construction within the budget and on time. This can be accomplished by generating schedules with the same likelihood of cost & duration for all tasks and ranking the schedules according to the total cost & duration. The closest the total cost & duration are to the proposed budget and deadline, the higher the rank. This a simple task to most algorithms; therefore, the solution best embedded with the software will probably be the chosen one.

Breakdown structures are a hireable approach to manage a large amount of data that can be directly applied with BIM [15]. Developing an appropriate Work Breakdown Structure (WBS) is a significant step towards implementation, as any change in its structure will propagate from that moment on, but not backward. Due to it, the proposed WBS must be able to grow incrementally without losing the relationship with the data already collected and stored in the Data Warehouse. The granularity for representing the activities is crucial. It will be achieved by collecting data from different dimensions, as previously stated. Still, the WBS branch structure is what will allow it to grow and to search for data from a more general category until a robust dataset is consolidated.

The downside of this routine is that it demands a large number of samples to build a valid normal distribution. The time needed to capture enough data delays the implementation since collecting data from buildings before the CPM digitalization is not practical, using previous constructions to increase the historical asset is not an option. On the bright side, after collecting data from one building, it is probable to have enough data for implementing it since the tasks repeat themselves multiple times during the construction. The proposed methodology does not optimize the variable's values to reduce completion time or cost; it uses probabilistic analyses to provide two outputs: the chance of finishing the project within the time and cost constraints; and given the desired probability, what is the completion time and cost.

2.4 Information Visualization

Data visualization provides a powerful way to communicate data-driven findings, motivate analyses, or detect flaws. The fact that it can be difficult or impossible to notice an error just from the reported data frame of results makes data visualization particularly important. The goal is to develop a visualization so convincing that no follow-up analysis is required, to accomplish that, each set of data shall be treated individually and displayed with the most suitable visualization tool. The gathering of these visualizations tools (trend charts, histograms, Gantt charts, S-curve, etc.) on the same screen is called dashboard [16].

These real-time data visualizations enable the cross-contractor control war tower. War Room is a place dedicated to providing communication and collaboration space for a project team. The goal of this room is to ensure rigorous problem solving, visual management, and performance dialogues. The idea of a War Room is to gather all the vital information in one place, creating the environment to support sophisticated analysis and brainstorm solutions with all the people who are actively working on a particular project.

The War Room concept has been widely implemented in industries after WW II, the Toyota System Production embraced it, and it is currently adopted in the Lean methodology.

Evaluating current dashboards software for the construction industry, such as Power BI, Inetsoft, and Constructiononline, it was clear the struggle to handle data from different dimensions linked to BIM. When they do so, it is mostly related to metadata. Take Figure 3- Frequency Data Connector Dashboard and Figure 4- Reported Issues Dashboard, for instance. The first uses Power BI software to display data about the frequency of users on the platform. The last is the dashboard in BIM360 that gathers data from reported issues related to construction, but none of them provides actual construction progress information.

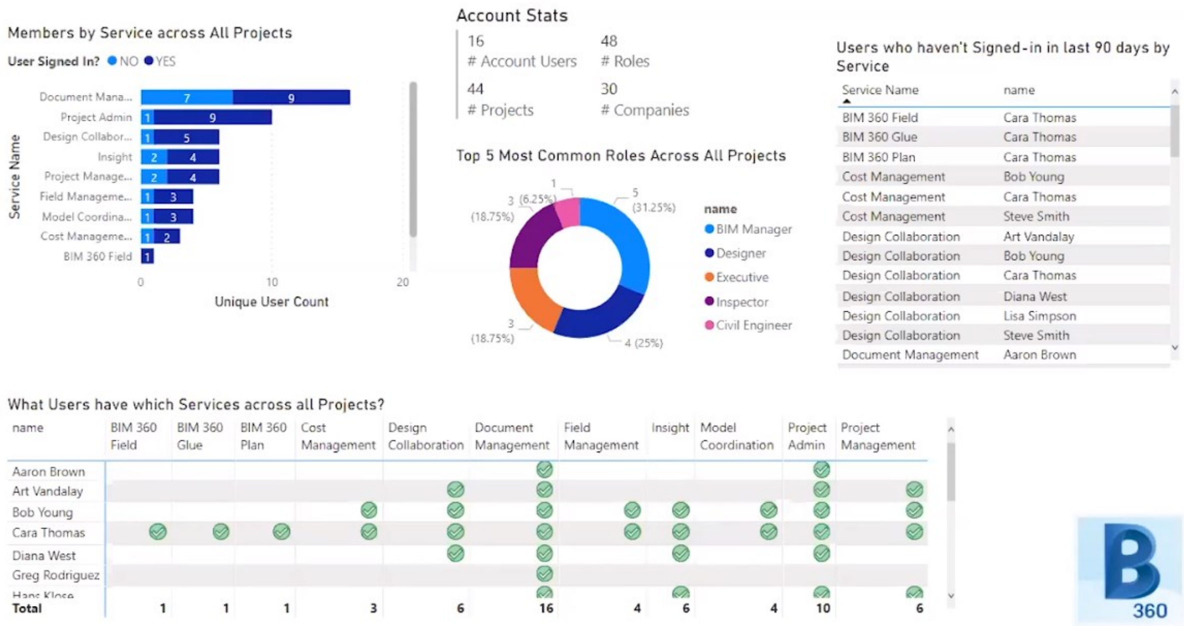


Figure 3- Frequency Data Connector Dashboard

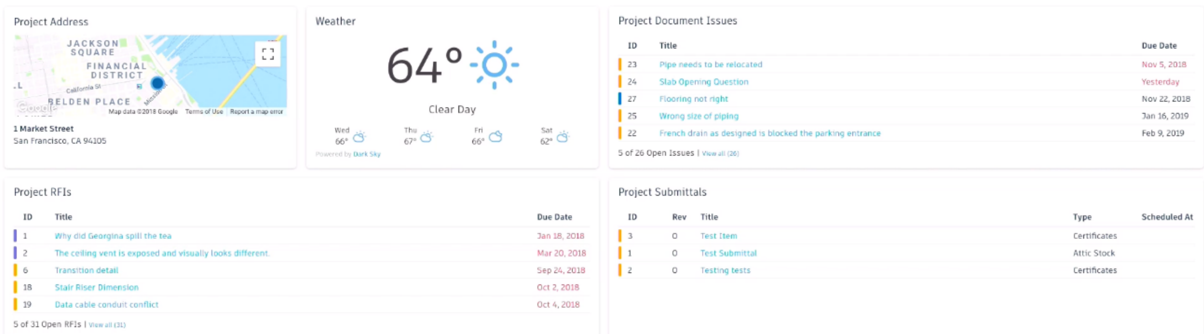


Figure 4- Reported Issues Dashboard

While selecting the visualization software, one must know what analyses will be performed, the database format, and the operator technical skill. It is vital to remember a significant advantage of the Data Warehouse information flow: to split it into the back and front end, allowing naïve programmers to perform data analysis. The perfect tool would not elevate the computer skills requirements while being able to perform all the desired studies. There are two types of solutions: visual coding software such as Power BI and hard coding like R. In general, the former is preferred, but there are times when the latter is the only solution for a problem.

This section proposes two dashboards, the Construction Progress Dashboard and the Activity Dashboard. The first one delivers a clear overview of the schedule, cost, and elements showing real-time construction progress while the second one used for individual activities analysis.

Presented below is Figure 5- Construction Progress Dashboard concept; each element is tagged and briefly explained.

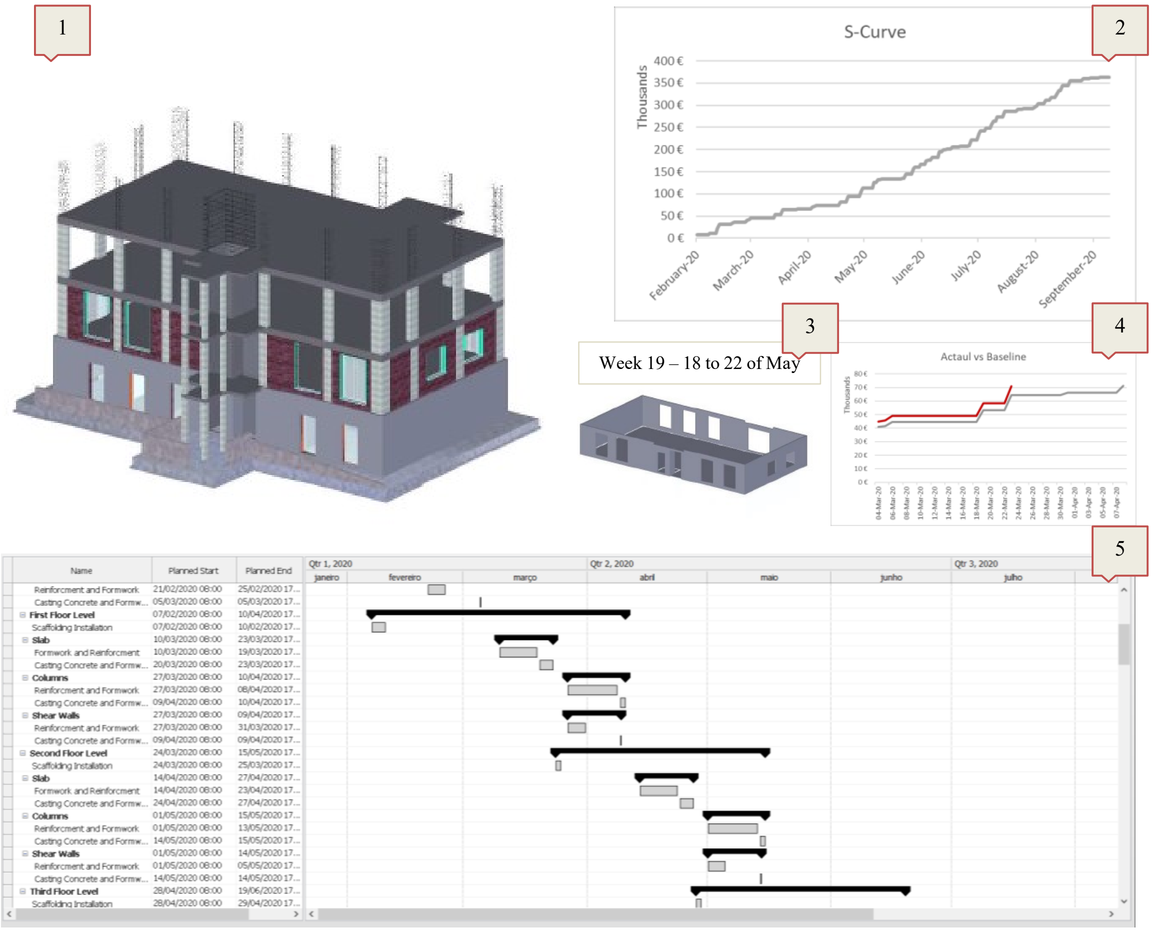


Figure 5- Construction Progress Dashboard

1- Construction Elements Completed: Model View Definition (MVD) using the actual end date as a filter;

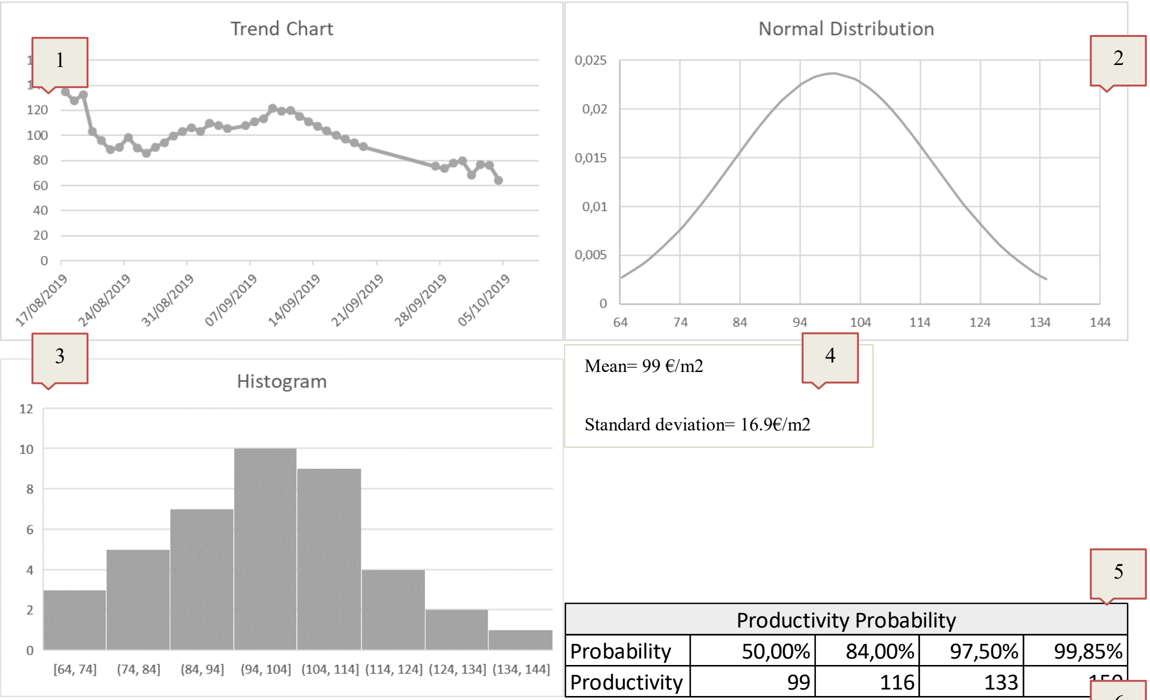
2- S-Curve: graphical representation of cumulative baseline cost of the whole project;

3- Construction Elements to be Completed: MVD using the planned end date and the week period as a filter;

4- Plan vs. Actual S-Curve: zoom in and compare the cumulative cost of the baseline with the actual cost;

5- Gantt Chart: a type of bar chart that illustrates the project schedule.

Figure 6- Activity Dashboard display relevant metric about individual activities, following the same procedure of the first dashboard, bellow is the concept of it.



Name	Planned Start	Planned End	Material Cost
Blockwork	21/04/2020 08:00	27/04/2020 17:00	5716,12
Blockwork	21/04/2020 08:00	04/05/2020 17:00	7619,63
Blockwork	26/05/2020 08:00	01/06/2020 17:00	991,67
Blockwork	26/05/2020 08:00	08/06/2020 17:00	7075,69
Blockwork	30/06/2020 08:00	06/07/2020 17:00	1013,22
Blockwork	30/06/2020 08:00	13/07/2020 17:00	7014,85
Blockwork	04/08/2020 08:00	05/08/2020 17:00	2077,99
Blockwork	04/08/2020 08:00	10/08/2020 17:00	1018
Blockwork	04/08/2020 08:00	17/08/2020 17:00	7166,95

Figure 6- Activity Dashboard

1- Productivity Trend Chart: shows trends in data over time;

2- Normal Distribution Curve: illustrates the probability distribution of samples;

3- Histogram: splits data into ranges and represents it as a bar chart, the propose of this chart is to validate the normal distribution, if format created by these two graphical tools are not similar, the normal distribution probability analysis is not valid;

4- Population Parameters: two numbers that describe the entire population, the Mean is the average, and the Standard Deviation is how much the members of the population differ from the mean value;

5- Productivity Probability: this table display the productivity for each probability, given the validation of the Normal Distribution;
6- Activity List: shows when the activity appears in the schedule and some data attached to it.

The two data visualization dashboards proposed will efficiently display fundamental performance analyses, making the most out of the historical data from the Data Warehouse. It will empower the data-driven culture of the company by allowing the manager to use cross-sector information, increasing the digital asset value by making it available, which means to be able to locate, retrieve, and use the desired data.

3 BACK END ROUTINE PROPOSAL

The routine proposal for the implementation of the Business Intelligence data center for the Construction Industry is divided into front and back end, this topic will cover the back end. Figure 7 enumerates the steps for implementing the proposed routine while following the Information Flow.

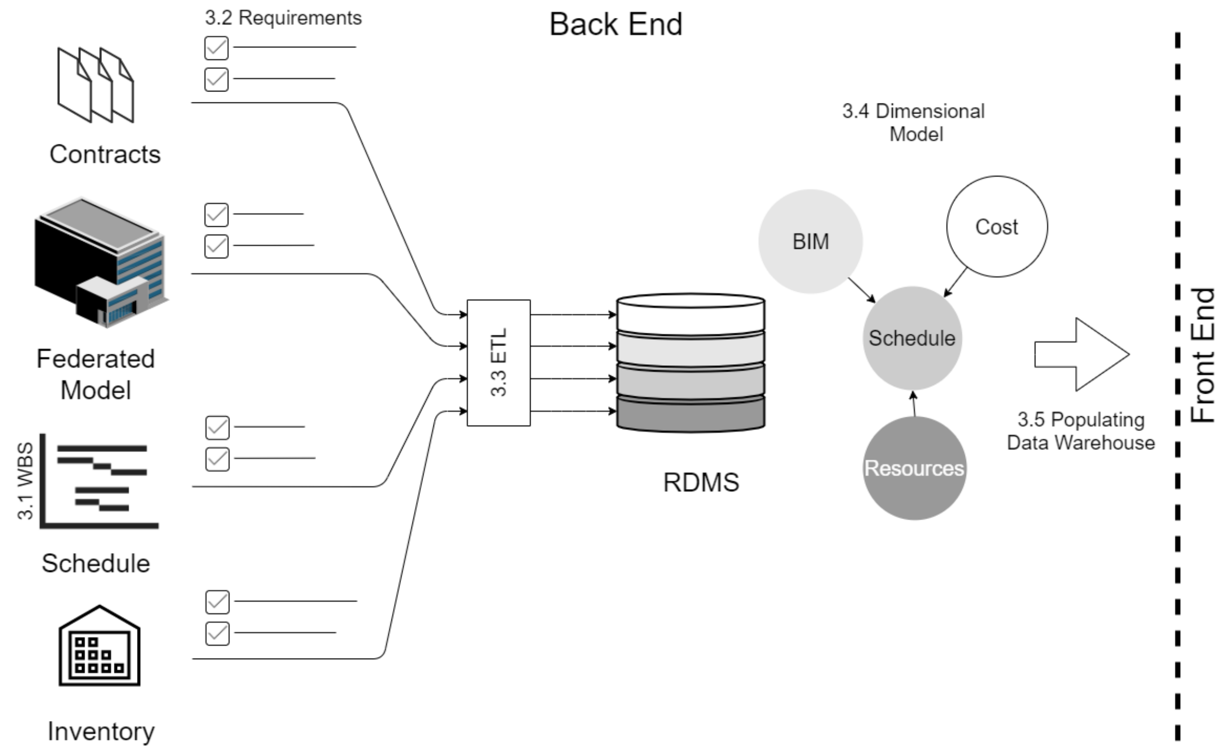


Figure 7 Information Flow Back End

The next topics are about building the data architecture: 3.1 3.1Work Breakdown Structure, 3.2 Implementation Requirements, 3.3 Extract, Transform and Load, 3.4 Dimensional Model, 3.5 Populating Data Warehouse.

3.1 Work Breakdown Structure

The WBS adopted for the Data Warehouse will be the schedule's WBS because the events selected for building the Data Warehouse are the tasks. However, some changes might be needed, and these are the two main reasons: linking it with other dimensions and allowing it to upgrade on the fly.

Every task must have a modeled element to be attached to in order to enable the link with the BIM dimension. The cost of every contract will be distributed into all tasks it is related to, according to the weighted average of the primary quantity. Therefore, the cost for the material resource will come from the elements quantities takeoff associated with each task, so one task without elements is a task without cost. There are many solutions for this issue; it can be to add elements to the model, clustering tasks together, or even replacing a task of a lag in the schedule.

As mentioned earlier, the WBS should have a branch structure that allows for the collection of missing data from the broadest category and a granularity that distinguishes only what should be distinguished through the productivity lens, but no more. Take, for instance, the activity of casting concrete; it is preferred to split it into ordinary & extraordinary elements or horizontal & vertical elements so that until the database has enough samples it can cluster them into these two categories. Also, when another element type is added, it can feed on more similar elements.

Also, the WBS must follow the same standard in different constructions, or else the Data Warehouse will not provide the expected result. With that in mind, adopting a third party established construction classification systems such as UNI Class or Omni Class instead of developing the WBS from scratch is advisable due to the following advantages: speed up the implementation; unlocks third party's benchmarks and collaboration; increases transparency and clarity of bidden; provides the possibility to further customization. It should be noted that third party institutions developed WBS, and most leading countries have its own, to select one is to choose the country as well, it is a strategic and political choice that should not be taken lightly.

3.2 Implementation Requirements

This section will list and explain all the digital documentation required to implement the proposed routine. Finding a project that already has the required document in a digital form entirely developed is not necessary, even though it is not that rare, each part can be created within its sector with the proper understanding. All data was resumed to data frames, table format; this was done to cover all software from each dimension since all of the existing ones will be able to export data frames.

Table 5 divides the documents into the four dimensions.

Table 5- Documentation Required

Cost	Schedule	Resources	BIM
Contract Description (what, how much, who); Planned Direct Cost per Task (divided by resource).	Schedule sheet. Working Calendar Dependency between tasks	Inventory issues, transferences, and returns.	Federated Model; Primary Quantity Takeoff Model; 4D Model.

From the Cost dimension, two data frames are needed. One listing every contract and showing the primary information about it under the following categories: contract's value, contracted company, contracted service (material, equipment, labor or subcontractor), contracted quantity, and if it is ordinary or not. The second is the planned cost per task; usually, this is done at the beginning as a front-end estimate and each resource percentage in the estimated value, the objective is to relate it with the actual cost, improving this estimation for the next projects.

The schedule dimension will provide the WBS id, tasks, planned start; planned end, actual start, actual end, progress, task location (story/room) in the same data frame. Another two tables are required, the calendar with the working days & holidays and the list of relationships (Finish-to-Start; Finish-to-Finish, Start-to-Finish, and Start-to-Start) between tasks. The data from this dimension is more comprehensible than the rest; therefore, no further explanation is needed.

The data originated from the Resource dimension is the record of the inventory issues, transferences, and returns; the goal is to be able to quantify the use of material resources. The transactions will belong to the same data frame with information under these categories: material, quantity, in or out.

The BIM dimension requires state of the art technology to implement, unlike the other dimensions so far, every tool and procedure was created in the previous century. The federated model is a summon of all distinct discipline models, creating a single model of the building. It is the only document required whose absence can make implementation impossible; however, it is not unprecedented for a construction company to build the model.

With the federated model, the primary quantity takeoff model shall be developed, providing a data frame for each task listing the elements GUID (unique identification), its primary quantity for that task, and the primary quantity unit. This data frame will be used to assign the number of resources to each element, through weighted average relating resources to the primary unit and, by the group of elements, assign the number of resources to each task.

Also, the 4D model must be produced from the federated model, there are many ways of doing it, but the outcome must be a data frame listing the elements GUID for each activity from the schedule dimension.

3.3 Extract, Transform and Load

The proposed routine for loading the information to RDMS can be done with human input, or it could be replaced for an ETL tool, the method and steps are the same regardless of the solution. ETL stands for Extract, Transform, Load. It is a data integration tool used for blending data from numerous sources and involves four steps.

First, one must locate the data. All the data frames explained in the previous section have a known path in the server, and the ETL will get the data from these frames but not from the actual source (e. g. BIM model). Therefore, it is advisable to have a script to update all the data frames from the source file and to run both in sequence. Companies will have the desired data stored across different software, so it has different formats and is stored in numerous sources. Therefore, during this stage, it is essential to define the required sources and format of data. Each source must be evaluated individually.

Secondly, the data is transformed into a suitable format; it can mean to adjust data to a standard or to split data from one column into two or more. Each dimension will have one data format standard, for instance, the schedule dimension: the information collected could be 12/may/2020, 05-12-2020, 12.05, and so on. After formatting, it should all look the same. This step shall not be dismissed regardless of how strict the company is on following standards; the RDMS only accepts specific data format, and to rely on the perfection of human input is foolish. Also, in the schedule dimension, one can wish to divide the WBS number into different columns, to increase the processing speed in the Data Warehouse, the one WBS string with the value "3.2.1.2" would be transformed into four columns.

The third step is to put the ETL tasks in the proper order, to create the routine for it. Which data frame should be transformed first, and so on. A primary concern of this step is the loading order since the RDMS only accepts a foreign key in one dimension once the primary key to which the foreign refers is already loaded.

The final step is the execution; to extract, transform, and load the data into the RDMS. The result of this step is a report which will have metadata about the procedure, such as duration, list of successful tasks, and a list of errors.

The ETL solution is the best by far, not only to prevent mistakes and to save time but also to save money. One employer to perform this task must have a holistic knowledge of the companies file structure, proficiency in all software used in each dimension, and will spend nearly 3h per update.

3.4 Dimensional Model

Linking the dimensions demands a dimensional model, the one suggested for the implementation uses the star-like structure and places the schedule dimension in the center and surrounds it with the other three composed by the data frames described previously. Setting the Schedule dimension is the center of the model is appropriate since the event selected for the Data Warehouse is the task from that dimension.

The WBS is already implemented in the Schedule at this point, so to link the Cost, Resources, and BIM dimensions, there are two ways, one for "one to many" and "one to one" relationship and another for "many to many." The first is highlighted in green in Figure 8, while the last is in red.

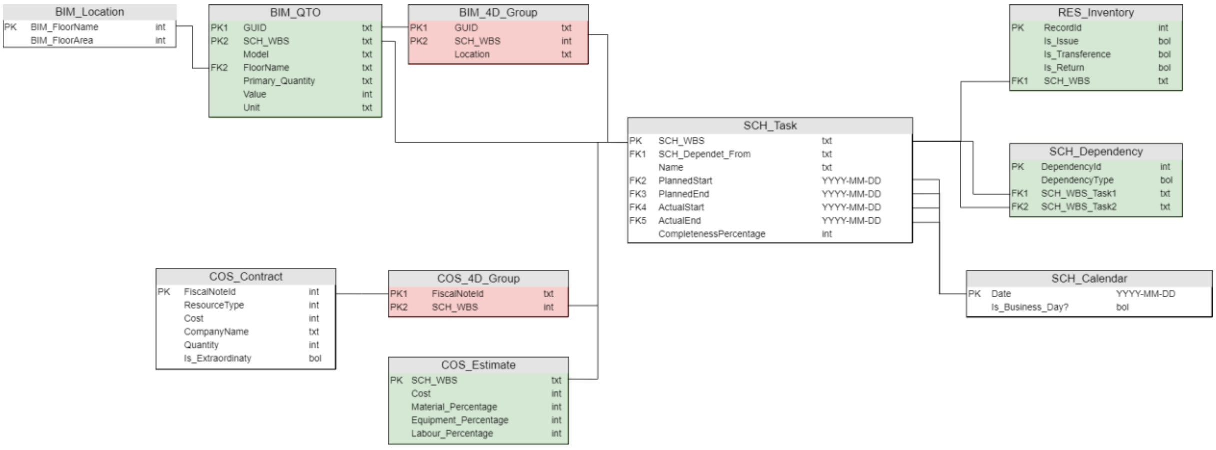


Figure 8 Entity-Relationship Diagram with Highlight

The green tables relate to the central one SCH_Task as one to many or one to one; the link is done by adding another column to data frames referring to each task that information belongs. For instance, COS_Estimate has the estimated cost of each task in the schedule and its composition, every task has only one estimation, and each estimation refers to one task only.

The red tables are added; they are frames used to link dimensions whenever the relationship is many to many. For instance, the COS_4D_Group table is connecting the COS_Constract with the central table, this means one contract can refer to more than one task, and one task is composed of more than one contract. As the activity of placing rebar in the third level slab, there are many contracts related to this activity, but many of these contracts also relate to placing rebar in other elements.

3.5 Populating Data Warehouse

To populate the Data Warehouse, one must perform one query in the RDMS, which results in one single table with all the data required, this is done by writing code in SQL. It is essential to know that SQL is not a programming language such as Python, it is software dependable. This means that even though the general commends are the same in most software, there are differences, and the code developed in one tool is not guaranteed to work on all RDMS software.

There are two types of data: raw and processed. The former is directly queried from its data frame (e. g. WBS, Name, and Planned Start), the latter is transformed and require more profound SQL knowledge (e. g. Actual Duration, Material1 Cost and Actual Primary Unit Cost). Some columns, despite appearing to be a redundancy of information, are not. The reason is not just to speed up processing for analysis, but to add information. For example, the Duration of the task is stored even though it can apparently be deducted from the Start Date and End Date, in which case the information added is how many working days there were in that period and if the task was on hold, this time is already subtracted to calculate the duration.

Similar to the Implementation Requirements, where it was listed all the data to be harvested, Table 6- Data Warehouse: Columns and Examples contains all the data to be framed for the Data Warehouse and provides two examples of value for each column. It will be the starting point for developing the SQL code.

Table 6- Data Warehouse: Columns and Examples

Column	Example 1	Example 2
Building Code+Date	1001-2020-05-20	1001-2020-05-20
WBS	2.2.1.1	2.2.2.2
Name	Structural work. Formwork. Horizontal element. Ordinary	Structural work. Concrete casting. Vertical element. Extraordinary
Possible Start	2020-05-21	2020-04-02
Actual Start	NULL	2020-04-02
Actual End	NULL	2020-05-04
Actual Duration	NULL	33
Planned Start	2020-05-21	2020-04-02
Planned End	2020-05-28	2020-04-28
Planned Duration	8	27
Stop Motive	Late arrival of Formwork sheet 200x100	NULL
Attached BIM Set	FormworkHorizontalLevel-3JointA	ConcreteCastingVerticalFacadeJointB
Level	-3	NULL
Area	2235,00	NULL
Joint	A	B
Provided Progress %	NULL	100,00%
Extraordinary Cost?	NULL	1.025,00 €
Material1	Doka beam H20 top	Premix concrete C40/45
Material1 Quantity	10	9
Material1 Unit	unit	m3
Material1 Cost	1.000,00 €	250,00 €
Material2	Doka formwork sheet 3-SO 250x100	NULL
Material2 Quantity	12	NULL
Material2 Unit	unit	NULL
Material2 Cost	350,00 €	NULL
Material3	Formwork sheet 200x100	NULL
Material3 Quantity	3	NULL
Material3 Unit	unit	NULL
Material3 Cost	30,00 €	NULL

Total Material Cost	1.391,00 €	250,00 €
Labor Cost	187,33 €	NULL
Equipment	Craine	Concrete Pump Truck
Equipment Cost	5,45 €	550,00 €
Subcontractor	NULL	Concrete Facade Co.
Subcontractor Cost	NULL	687,45 €
Planned Direct Cost	1.696,00 €	1.252,50 €
Actual Direct Cost	1.583,78 €	1.487,45 €
Primary Quantity	21,20	8,35
Type	area	volume
Unit	m2	m3
Planned Primary Unit Cost	80,00 €	150,00 €
Actual Primary Unit Cost	74,71 €	178,14 €
Unit Labor Cost	8,84 €	NULL
Unit Equipment Cost	0,26 €	65,87 €
Unit Subcontractor Cost	NULL	82,33 €

Every time the query is run, the table obtained is a snapshot of data from the four dimensions to be added to the Data Warehouse.

4 FRONT END ROUTINE PROPOSAL

This topic presents the routine proposal for the implementation of the Business Intelligence data center for the Construction Industry in the front end. Figure 9 enumerates the steps following the Information Flow.

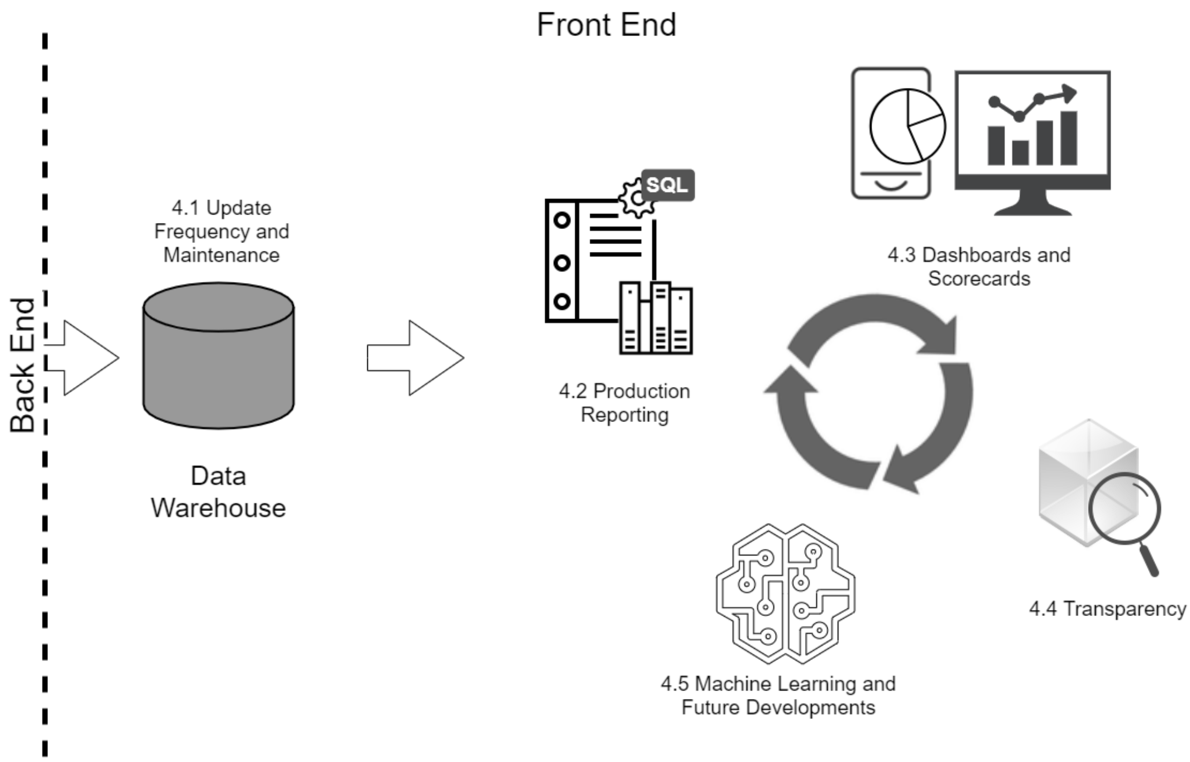


Figure 9 Information Flow Front End

BI's front end tools are defined as the following features [17]. They will be the focus of the next topics: 4.1 **Error! Not a valid bookmark self-reference.**, 4.2 Production Reporting, 4.3 Dashboards and Scorecards, 4.4 Transparency, 4.5 Machine Learning and Future Developments.

4.1 Update Frequency and Performance Maintenance

The operation of adding an RDMS snapshot to the Data Warehouse is called extraction. It can be triggered by an action in the central dimension (Schedule) or happen periodically. The data is never replaced, but there are cleaning and maintenance procedures that can be done to optimize the storage. Both of these matters, extraction, and maintenance, will be explored in this topic.

There are two types of extraction: full and incremental. The former capture all data from the source system while the latter only collect the data that has changed since the previous operation. The choice of type and frequency at which it will be done affects the Data Warehouse file size. It will have 42 columns and will gain as many rows as there are tasks in the schedule for each snapshot. This operation will demand about 300 extra KB for a 600 tasks schedule full extraction.

Due to the file structure of each dimension, it is challenging to keep track of which data has been updated since the last extraction, the solution proposed is to extract all the data available and later to clean the rows without changes. The methodology can differ as to how long one should wait before erasing the redundant rows from the Data Warehouse, and some approaches reduce the granularity from days to weeks after a given time reducing to 15% of its original size.

Extracting it daily would demand an additional 9 MB of storage per building per month, which is not significant compared to file sizes in BIM implementation, where 1 GB size files are standard. However, if the extraction is triggered by a change or update in the schedule, it could happen multiple times per day or, even worse, none at all. For the intended use, it is not a disadvantage to miss track of changes during the day; the daily update is the best solution.

4.2 Production Reporting

As stated earlier, the average engineer or project manager requires very little training before performing queries to access the Data Warehouse or the RDMS directly to extract ordinated data. It is useful to empower the non-technical end-users to generate their own reports quickly, without burdening the IT department for information that was not included in the programmed visualization tools.

The scenario painted to exemplify this feature is the construction project manager who is preparing for a meeting and wants to know how many times a subcontractor delayed his deadline and the ranking of suppliers ordered according to its purchased value.

None of these features were foreseen when developing the dashboard. Therefore, the RDMS must be accessed directly by writing an SQL. First, to obtain the subcontractor's information, the code shall filter tasks in which that subcontractor is listed and count every time the Planned End occurred before the Actual End. Second, to create the rank, the code shall work only within the COS_Contract frame to filter the contracts where ResourceType is equal to Material, then sum the Cost accordingly to the CompanyName and display the CompanyName in the Cost order.

These are two scenarios that could be solved by a four-lines script, which can be easily written by a naïve programmer, it could be done by the manager himself or by an intern with 8h YouTube training in SQL.

4.3 Dashboards and Scorecards

Dashboards were thoroughly explained in section 2.4 Information Visualization. This feature aims to fit a real-time business overview into one computer screen displaying graphs, charts, summaries, and other information reports. The concepts proposed in 2.4 provides the company's significant stakeholders with crucial information at a glance to make smarter, faster, and better decisions.

There are many tools available to translate the Data Warehouse into a "human-readable" display, most of which allow interaction between the user and the wide variety of possible views; this customization is considered essential to improve usability. Also, tailoring the visualization to diverse player profiles and using online data access via mobile devices is a huge advantage in promoting holistic business knowledge at the top of the pyramid and improving synergy at the bottom.

Section 2.4 exemplified two uses of dashboards for meetings and overall project and task's indicators. The present chapter will add by introducing the Scorecard concept and elaborating two examples: one for the top of the pyramid and another for the bottom.

Scorecards differ from dashboards in that it focuses on comparing a forecast or a goal to a given metric; the selected metric can use data from multiple dimensions. Scorecards can be used to synthesize performance against benchmarks providing insights for different levels of stakeholders.

For instance, at the top of the pyramid, it can provide a strategic-level analysis: gross value added per month worked. As previously explained, it can only be realized after the asset is liquidated or with an accurate forecast of revenues. It consists of subtracting the total material and equipment cost from the total income and divide this value by the time invested in months. An online scorecard that displays this indicator across projects using a similar metric from different industries as a target leverages a construction company to the state of the art business intelligence application in the construction industry.

For the bottom of the pyramid, the scorecard can be used to display the subcontractor's productivity while utilizing the benchmark required by the contract to receive the bonus as a target. For both scenarios illustrated, having a single source of truth is mandatory to support the data against skepticism.

4.4 Transparency

The proposed Business Intelligence routine increases transparency as a whole, even though it was not its primary objective, this brings major advantages to the construction industries. There are two aspects in the routine that enhance transparency: the strict system that encompasses all expenses and the effort to make all data available.

First, the fact that every contract must be linked to tasks, every task is quantified, and the ratio of productivity and cost is stored in a single source of truth makes it easier to locate errors. Second, the development of a tool for data visualization and sharing it with the main stakeholders without taking them away from the big table. Both add significant value to the business.

The documentation of the construction reality in real-time, along with the user-friendly data visualization, signals the general contractor as trustworthy and, by doing so, attracts staff and subcontractors who share the same view. This information sharing is a must to implement modern project delivery models, such as Integrated Project Delivery.

4.5 Machine Learning and Future Developments

The proposed routine paves the path to growth, as it is in all BI tools, the constant improvement is one of its keys to success. This topic proposes ways to further develop the proposed framework in both: Back and Front End. Back End growth in the sense that future developments could cover more processes inside each dimension to capture data to better describe the reality through data frames. The Front End growth is adding different features to the current data, without adding demands to the Back End.

To link the current Data Warehouse with external sources would increase usability. For instance, to integrate the Data Warehouse with external data, such as material suppliers' price sheets to have a real-time cost of resources yet to be purchased or the national data frame for unitary cost. Some other external sources could be another company's productivity in case of a unified bidding proposal.

To create tailored tools to speed up procedures, such as the development of integrated tools to automate the inventory transactions data frame with code scanners readable by mobile devices.

Machine learning application, even though this study does not engage it, it cannot go unmentioned. The same way the collected data can be used for statistic inference about the relationship between variables; it can feed a machine learning model to make the most accurate prediction.

To have unnormalized historical data gathered in the Data Warehouse provides all the necessary data for creating the Analytical Base Table (ABT) from which the machine will actually learn. An ABT is defined as a flat table used for building analytical models and scoring (predicting) the future behavior of a subject. The proposed workflow allows the creation of an ABT with around 30 descriptive features, excluding the redundant columns, for to system to target total cost or task duration.

It is also essential to highlight the Data Warehouse is a very accurate reality description, not only for the tasks but for the project itself, a lot of questions can be answered from it with increases its potential as ABT. This accuracy is assured by the methodology used in its development.

5 BEXEL MANAGER SOFTWARE ASSESSMENT

According to BuildingSMART, BEXEL is a high-tech, construction, and engineering consultancy company. The BEXEL Manager is the software they developed for internal use and later on became their main product. BEXEL Manager will be referred only as "BEXEL" for the rest of the present topic. BEXEL is capable of handling data from the four dimensions discussed previously, and therefore it is useful as a benchmark for the proposed Business Intelligence data center for the Construction Industry, which will be referred to as "BICI".

This topic aims to explore how it stores, manages, and analyzes data. Before starting, it is essential to understand that exploring BEXEL was done by reverse engineering it and the original and exported files from the software were used in this task. The author did not have access to the code; therefore, the information presented are assumptions based on elements of the software's front end, to make assumptions for the back end by exploring the front end is far from an exact science. Reverse engineering is a redesign methodology that uses a variety of procedures to dissect and fully understand a product; the goal is to compare the front end features and to identify system components interrelationships.

The "Bexel Sample Project – V1", the demo downloaded with the program, was used as a sample to this analysis. It was developed by BEXEL to get the most from their software; therefore, it will most likely enable all the software features. The files used in the process were the following: BEXEL original file; IFC exported file; Power BI exported file (still on beta phase); Excel Schedule exported file; Excel Cost Breakdown Structure exported file.

The BEXEL assessment was done file by file; however, it will not be presented as such. This topic will be divided into dimensions. The following will be explored for each dimension: what data does it contain, how it is collected, how it differs from the proposed approach. Then, the last three sections will analyze how it manages and structures data to link the dimensions, which analyses it performs, how the data is visualized.

5.1 Cost Dimension

The Excel Cost Breakdown Structure exported file displays the cost data and illustrates how BEXEL links data. The information flow in the data frame is visible by the dependency between cells, its inputs, and outputs. The purpose of this table is to calculate the total cost of each task. In the sample file, BEXEL follows two parallels routines: one for labor and equipment cost and another for material cost.

The labor and equipment come from the combination of three data frames, one which lists the labor and equipment demanded by each task, another which lists the hourly cost of each labor and equipment, and the last data frame, which is the name and duration of each task, the schedule. Using a process similar to the SQL software, it is capable of retrieving the total labor and equipment cost per task.

The information path for the material cost is not clear in the Excel Cost Breakdown Structure exported file, the BEXEL original file was explored to understand how it is obtained. Following BEXEL's online seminars about "5D Cost Estimation", it was clear the unit cost is the input and the total cost, the output. To first retrieve the unit cost of each task, previously provided by the user, is the opposite of the BICI and states that BEXEL is designed for cost estimation.

With the unit material cost at hand, it must then calculate the sum of the primary quantity of each task. BEXEL's routine is to sum of primary quantity starts from the task's WBS, using queries to link BIM elements to tasks and formulas to calculate the primary quantity from the element's properties. It obtains the sum of the primary quantity of all the elements to associate with each task. The primary quantity of each task has a defined type and unit.

Finally, it multiplies the unit cost by the primary quantity of each task and adds it to the total labor and equipment cost per task resulting in the total cost of each task. An example of each one of the discussed columns is in Table 7.

Table 7 - Excel Cost Breakdown Structure exported file, Value Example.

Column Name	Example Value
Outline Level	1.1.1.1.1.1.0
Code	03 31 0570 0001
Name	Structural concrete, placing, pile caps, direct chute, under 3.83m3, includes vibrating, excludes material
Resource Name	Cement Finishers; Common Building Laborers; Common Building Laborers Forman (outside); Gas Engine Vibrator.
Resource Type	Labor; Labor; Labor; Equipment.
Resource Quantity Type	Time; Time; Time; Time.
Resource Quantity Unit	h; h;

	h; d.
Resource Quantity	8,00; 32,00; 8,00; 2,00.
Resource Unit Cost	20,86; 78,31; 81,81; 47,08.
Unit Cost	$49.72
Quantity Type	Volume
Quantity Unit	m^3
Quantity Formula	[Volume]
Element Query	[FAMILY] = '%ST-Pile_Cap_2%'
Unit Cost	49,7230053771254

BEXEL has two ways to insert the cost data, it can be done manually, one by one, or it can import it from an Excel template, using data frames. The process of capturing data from the Cost dimension is very similar to the BICI even though the approach is from unitary to total instead of from total to unitary.

5.2 Schedule Dimension

The data management from the Schedule dimension in BEXEL is the same as in the BICI. This is clear since it features analyses of actual vs. planned schedule but does not have an analog tool for the Cost and Resource dimensions.

BEXEL is effective at collecting data from external sources through data frames, the same as the BICI; however, it only stores the present data and keeps no track of changes. One could store old files, but to analyze one against the other would be ineffective since it is not an embedded functionality of the software.

5.3 BIM Dimension

The BIM dimension data flow starts by importing an IFC file to create a federated model, as it was expected to any non-authoring tool such as BEXEL. It contains all the elements of the project and a lot of properties are attached to each one of them; a lot of unnecessary data is captured when importing the IFC file. The advantage of this method is it does not demand the conversion of the BIM model into data frames, even though it is a simple process; it is a new process.

This dimension is linked to the Schedule by assigning elements to tasks; it can be done by queries or by selecting the elements. Similar to the BICI, it is the most laborious task in implementing BEXEL.

An exciting possibility to the BICI would be to add the zones division to the BIM dimension; there are many advantages in BEXEL due to it. Adding zones would demand very little since it can be done as a property to the elements in a BIM authoring tool. It would then be extracted with the data frames to be added in the RDMS and later to the Data Warehouse.

5.4 Resource Dimension

BEXEL combines data from Cost, Schedule, and BIM; however, it does not collect data directly from the Resources, it replaces this last dimension with formulas relating the primary quantity from the BIM to the material resource quantity. It is a different approach that works well for planning but presents limitations to construction monitoring since there is no data related to actual resources. Still, it would be beneficial to add this feature to the BICI enabling waste analyses.

Since BEXEL does not capture data from the inventory, the actual resources are not obtained; therefore, it cannot perform actual vs. planned for this dimension either. There is an up-side to it; the implementation demands fewer changes in the current workflow and less commitment from a company when choosing the tool. It makes sense to do so for software, but not for a BI Data Center.

5.5 Back End Architecture

The fact BEXEL is not able to export an IFC file with data from the four dimensions is evidence that it does not store all data in an object-orientated way. Its exported IFC file uses IfcTask, IfcRelNests, IfcRelSequence, IfcCalendarDate, IfcDateAndTime, IfcRelAssignsTasks, IfcRelAssignsToProcess, and IfcScheduleTimeControl combined to store Schedule dimension data. However, the Cost and Resource dimension are lost in the IFC format, which indicates a similarity to the BICI as it uses the Schedule dimensions as the central link to BIM, Cost, and Resource.

A similar feature from BEXEL is to be able to export a Data Warehouse with data from the four dimensions; the Excel Schedule exported file is a 48x6384 data frame that contains a snapshot of data from the four dimensions. The data frame can be categorized as 25 columns about Schedule; 10 columns about Cost; 5 columns about BIM; 8 columns about Progress Monitoring, leaving Resource dimension out of it. The granularity is also different; for instance, it stores the total cost instead of the cost of each resource. It would be presumptuous to point is as a mistake; it only confirms that the propose of the Data Warehouse is not the same and that historical analyses are not among BEXEL's many features. This feature could be added to BEXEL by the end-user, it would take very little to code a daily extraction of its Data Warehouse and compile it with the previous one instead of replacing it.

BEXEL data architecture differs from the BICI, mainly for two reasons. First, it does not collect actual data from the Cost and Resources dimension. Missing this data is not good or bad; it is a choice that leads to a more straightforward implementation. Second, it does not keep track of previous snapshots; neither was it designed to analyze trends, missing the historical analyses and the Business Intelligence with it.

5.6 Front End Features

This topic will follow the organization established by BEXEL dividing the analyses into 4D and 5D, 4D is the collection of features related to schedule and monitoring progress, and 5D is the cost related hub. Both use data from multiple dimensions. There is a grey area between then, this topic will start by the most clearly related to 4D and shift to 5D as it develops.

Scheduling and planning can be called a feature from BEXEL even though it is most likely developed in different software and imported later on. However, an exciting component of BEXEL is to custom calendars for each subcontractor; this feature could be surely added to the presented BI routine.

The schedule optimization could not be deeply analyzed due to lack of information, even though the way the problem was modeled could be inferred, the bottleneck is the uncertainty of the solution approach and the algorithms used. The brief explanation provided in BEXEL's web site point to a strategy of using budget, deadline, and task's dependability as constraints while changing the task's orders and resource allocation. It is safe to assume it has a different approach to the proposed one of using probabilistic analyses to provide two outputs: the chance of finishing the project within the time and cost constraints, and given the desired probability, what is the completion time and cost. This analogy also highlights the different focus; BEXEL applies optimization on planning while BICI focuses on assessing risk with actual production rates.

Progress tracking is the contrast between the actual schedule with the planned one, BEXEL adopts the Earned Value Management combining cost and time, similar to the proposed BI routine. However, the unit cost itself does not change; only the time the task it is associated with is executed or its duration, and therefore, the cost is also affected.

BEXEL also creates de project plan, combining the Schedule with the Cost dimensions through the process explained in 5.1 Cost Dimension; resource and cost management are created during the process as well. The differences and similarities of this feature were already covered in 5.1.

Another BEXEL feature that could be added to the BICI is the generation of interim payment certificates. In BEXEL, this document is created with the actual schedule combined with the cost of each task divided into resources, the disadvantage of this method is to be based on unitary cost and hardly ever the actual cost will match the planned cost. Applying the same logic to the BICI would require combining data from the cost of each contract with the list of tasks the contract is linked to and the primary quantities of each task. This routine would generate the exact interim payment certificate to be approved by the construction manager.

BEXEL has an embedded Application Programming Interface console on the beta phase; it is as explained in 4.2 Production Reporting, there is an advantage in enabling and encouraging the project manager to perform queries to extract ordinated data. An example of an API application aligned with the clash detection feature is to organize clashes per cost of the elements involved in the clash.

5.7 Frond End Visualization

There are many ways to visualize data in BEXEL; the present topic will follow the same logic as the previous one, starting from 4D to 5D.

The construction simulation is among BEXEL's features; it is the most visual approach to display the construction sequence. It requires the same data as the previously presented analyses; however, it demands fine-tuning to appear professional; it can be worth the effort since the contractor tends to appreciate it highly even though, in the author's view, it provides shallow information. BEXEL uses color schemes to visualize construction sequences, allowing the user to color code by resource, task, and resource. There is no parallel to it in the proposed BI routine.

The Gantt chart is a shared tool for both routines; however, BEXEL also uses Line of Balance diagram to illustrate the schedule's repetitive work as lines of progression on a graph, it indicates the production rate. This representation is useful whenever dealing with construction projects that are repetitive or linear in nature, such as highways and pipelines.

The Power BI exported file does not present different visualization schemes; it changes the formatting of tools already embedded in the BEXEL original file. It also transforms cost tables into bar charts; the level of information add by this representation is trivial. The main advantage is it enables the interactive use of dashboards exploring multiple granularities; this feature would be a valuable addition to the BICI.

6 CONCLUSION

The high level of commitment required for implementing the Business Intelligence routine and the wide range of processes that would be affected made it impossible to perform on a case study. Without this validation procedure, the thesis focused on specific problems and challenges faced by the Construction Industry, applied established methodologies following consecrated authors step by step to ensure a valid and viable solution. Also, it explored software with similar proposals to leverage the achievements of the Business Intelligence data center for the Construction Industry.

The routine did not propose new types of performance indicators, metrics, or innovative solutions. It uses ordinary data, which leading construction companies already have, and arrange it into an architecture tailored for the business characteristics and demands. The question was: how to empower data-driven culture on construction to optimize cost estimation and schedules? The answer is to collect data across dimensions; to use RDMS to connect the four dimensions (Cost, Schedule, Resources, and BIM); to store the information in a Data Warehouse and to present insights through the most visual and dynamic tool available.

The digitalization of Construction Progress Monitoring was divided into the work of each dimension sector, and the interoperability was reduced to sharing specific and fairly simple data frames, which make it straightforward and not that difficult.

The proposed algorithm to optimize schedules does not engage the variable's values to reduce completion time or cost. It does not come up with changes to the current schedule, resource allocation, or the order of execution. It uses probabilistic analyses to provide two outputs: the chance of finishing the project within the time and cost constraints; and given the desired probability, what is the completion time and cost.

The use of BIM models is essential to build the Data Warehouse, as it will be used to assign the number of resources to each element, through weighted average relating resources to the primary unit and, by the group of elements, assign the number of resources to each task. It is essential to highlight the fact that, even though a design office can deliver a model ready to extract the primary resources, it is not reasonable to demand from the design phase a model with the data for the 4D (BIM_4D_Groups from Figure 8), this last piece must be developed and updated during the construction.

The front end is everything the end-user will have contact with; it is as responsible for strengthening the company's data-driven culture as the back end. It was designed to increase digital asset value by tailoring the visualization to diverse player profiles; the outcome is the adoption of the War Room concept along with the access to online data via mobile devices. The proposed solution could leverage a construction company to the state of the art business intelligence application in the construction industry, promoting holistic business knowledge at the top of its pyramid and improving synergy at the bottom.

The comparison between the data management and functionalities of the proposed BI routine with BEXEL Manager was fruitful to acknowledge useful features from the software to improve the BI routine. It was also enlightening to understand the choices made by the software developers to decrease the client's resistance against implementation, such limitations were also considered while developing the BI routine. However, BEXEL Manager ended up compromising monitoring to improve planning, the proposed routine did the opposite. The exact fact that each tool had different priorities was what made the analogy so enriching, many planning features were overseen and would definitely be added to the BI data center, such as waste control, zones division to the BIM dimension, custom calendars for each subcontractor, and generation of interim payment certificates.

7 REFERENCES

[1] R. Agarwal, M. Sridhar, and S. Chadrasekaran, "Imagining construction's digital future | McKinsey," *Article*, Jun-2016. [Online]. Available: https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/imagining-constructions-digital-future. [Accessed: 24-Mar-2020].

[2] J. Koeleman, M. J. Ribeirinho, E. Sjödin, and G. Strube, "Decoding digital transformation in construction | McKinsey," *Article*, Aug-2019. [Online]. Available: https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/decoding-digital-transformation-in-construction. [Accessed: 24-Mar-2020].

[3] H. Boutetière, A. Montagner, and A. Reich, "The keys to a successful digital transformation | McKinsey," *Survey*, 2018. [Online]. Available: https://www.mckinsey.com/business-functions/organization/our-insights/unlocking-success-in-digital-transformations. [Accessed: 23-Mar-2020].

[4] R. Eadie, M. Browne, H. Odeyinka, C. McKeown, and S. McNiff, "BIM implementation throughout the UK construction project lifecycle: An analysis," *Autom. Constr.*, vol. 36, pp. 145–151, 2013.

[5] K. Bråthen and A. Moum, "Bridging the gap: Bringing BIM to construction workers," *Eng. Constr. Archit. Manag.*, vol. 23, no. 6, pp. 751–764, 2016.

[6] K. Kensek and D. Noble, *BIM in the current and future pratice*. New Jersey: John Wiley & Sons, Inc., Hoboken, 2014.

[7] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, Third. Indianapolis, Indiana: John Wiley & Sons, Inc., 2013.

[8] W. Hughes, I. Yohannes, and J.-B. Hillig, "CIB World Building Congress 2007-253 Incentives in Construction Contracts," 2007.

[9] B. Botelho, "Data warehouse vs. data lake vs. data mart: Beyond the RDBMS," 28-Aug-2019. [Online]. Available: https://searchdatamanagement.techtarget.com/feature/Beyond-the-RDBMS-Data-warehouse-vs-data-lake-vs-data-mart. [Accessed: 01-Apr-2020].

[10] T. Trauner, W. Manginelli, S. Lowe, M. Nagata, and B. Furniss, *Stock Image Construction Delays: Understanding Them Clearly, Analyzing Them Correctly*, Second., vol. 2, no. 5. Elsevier Inc., 2009.

[11] A. Watts, *Modern construction : Lean project delivery and integrated practices*. Taylor &

Francis Group, 2011.

[12] International Organization for Standardization, "Information and Documentation --Thesauri and Interoperability With Other Vocabularies -- Part 2: Interoperability With Other Vocabularies.," *ISO 25964-22013*, 2013.

[13] A. M. Fahmy, "Optimization Algorithms in Project Scheduling," in *Optimization Algorithms - Methods and Applications*, InTech, 2016.

[14] J. Zhou, P. E. D. Love, X. Wang, K. L. Teo, and Z. Irani, "A review of methods and algorithms for optimizing construction scheduling," *J. Oper. Res. Soc.*, vol. 64, no. 8, pp. 1091–1105, 2013.

[15] L. Rischmoller, N. Dong, M. Fischer, and A. Khanzode, "Automation of the building information model breakdown structure," *IGLC 2017 - Proc. 25th Annu. Conf. Int. Gr. Lean Constr.*, no. July, pp. 389–396, 2017.

[16] S. Few, *Information Dashboard Design*, First. O'Reilly, 2006.

[17] C. Howson, *Successful Business Intelligence Secrets to Making BI a Killer App*, First. New York: The McGraw-Hill Companies, 2008.