Explorations in Data Mining

CM Perspectives into the Power of Business Intelligence

Laura Stagner *GSA*
Dennis Ryan *GSA*
Sergio Aranda *E-Builder*
Marty Turner *CH2M*

A CMAA Emerging Technologies Committee White Paper

The authors have worked to ensure that all information in this white paper is accurate at the time of publication and consistent with standards of good practice in the construction management industry. As research and practices advance, standards may change. For this reason, it is recommended that readers evaluate the applicability of recommendations in light of particular situations and changing standards.

CMAA: Construction Management Association of America
7926 Jones Branch Drive, Suite 800, McLean, VA 22102
Phone: 703.356.2622
Email: info@cmaanet.org

...promoting the profession of Construction Management and the use of qualified Construction Managers on capital projects and programs.

Printed in the United States of America

CMAA Emerging Technologies Committee

Explorations in Data Mining

CM Perspectives into the Power of Business Intelligence

Introduction

Data is essential to life. Data shapes our understanding and helps us make decisions. Data helps us convey our ideas. We have relied on data throughout human history. Now, this dependence is deeper than ever. The relationship is now better quantified, too. According to Vouchercloud's 2016 market research[1], more than 2.5 quintillion bytes of data are now created each and every day. That equates to approximately 4 million Google searches, 200 million emails, 277,000 tweets, 48,000 downloaded apps, nearly 3,500 Pinterest pins, and 400,000 Tinder swipes—every 60 seconds.[2, 3]

To put that number into perspective, consider the following chart, which highlights the size of an exabyte (quintillion) compared to the more common megabyte or gigabyte.

Unit of Measure	Symbol	Number of Bytes		
kilobyte	kB	10^3	Thousand	1000
megabyte	MB	10^6	Million	1000000
gigabyte	GB	10^9	Billion	1000000000
terabyte	TB	10^{12}	Trillion	1000000000000
petabyte	PB	10^{15}	Quadrillion	1000000000000000
exabyte	**EB**	**10^{18}**	**Quintillion**	**1000000000000000000**
zettabyte	ZB	10^{21}	Sextillion	1000000000000000000000
yottabyte	YB	10^{24}	Septellion	1000000000000000000000000

Quora, 2017

Included in this stream of daily data generation are the facts and figures created by the construction management industry related to projects, contracts, performance metrics, safety statistics, and more. Is there a way to sift through the overload—to find the truly useful information that can help us answer pressing questions, guide our decision making, and improve results? The solution to this quandary lies within the powerful process known as **data mining**.

This white paper, prepared by the CMAA Emerging Technologies Committee (ETC), delves into data mining—namely, the role that business intelligence tools and applications (BI) play in gathering and transforming quantitative data into actionable information that advances the industry. ETC asked several CMAA member firms to share the challenges posed by those working on the leading edge of technology. The effort yielded incredible insights into the benefits, challenges, and solutions associated with data mining implementation, system integration, data governance, and data sorting. This white paper also features a case study of one of CMAA's premier member owner agencies, the U.S. General Services Administration (GSA). GSA's story provides practical information and lessons learned, which are invaluable for any industry professional or organization managing disparate databases.

Definition and Functionality

The massive compounding effects of data volume and velocity in the era of cloud computing are commonly referred to as Big Data. **Data mining** represents a systematic approach to managing Big Data contained within databases or larger data warehouses, and is defined as the process of transforming raw data into actionable data. McCombs Business School professor Prabhudev Konana, Ph.D, defines data mining further as "***the extraction of new knowledge and insights from data or simply the identification or discovery of meaningful correlations, patterns, and trends using simple queries, statistical methods, and mathematic techniques***."[7] Ideally, data mining not only transforms raw data from databases and warehouses into knowledge, but also produces knowledge on which better decision making is based.[4]

The typical functionality applied by data mining tools include key steps to: normalize and correlate data; identify data outliers and missing data; process the data to find patterns using advanced algorithms and standardized methods such as data clustering, decision trees, regression analysis; and other techniques.

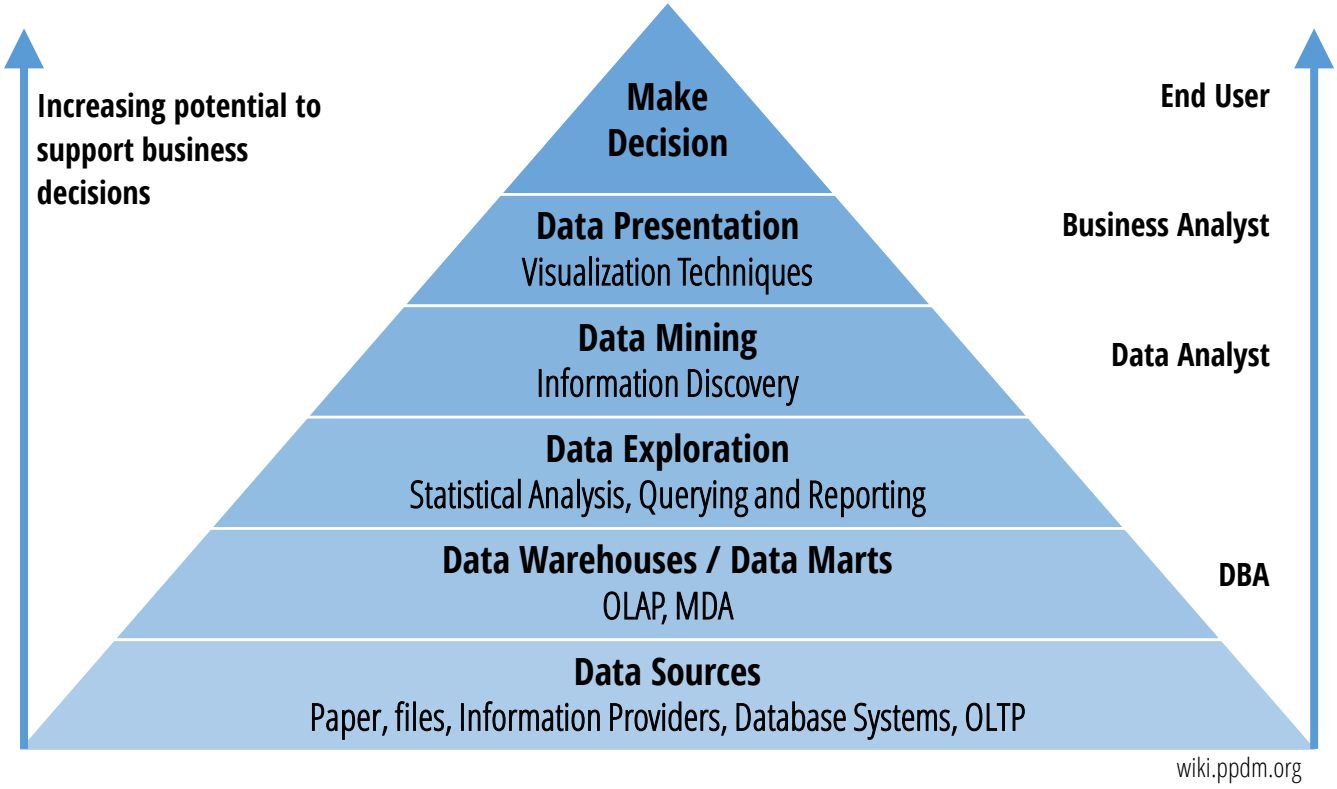
Early renditions of data mining tools were expensive, overly complicated, and not always user-friendly. They were effective at processing structured, homogeneous data sets to identify underlying patterns and trends; they could not consistently produce dependable results when source data was not commonly aligned. First-generation tools proved the benefits of data mining, but "business intelligence" did not always translate into "intelligent business."

In contrast, next-generation data mining tools (produced within the last two years) provide tremendously more sophisticated algorithms that are capable of analyzing both structured and unstructured data. This is important, because the vast majority of data remains largely unstructured and text-based, as it is primarily derived from email and web-based sources such as transactions, news reports, research articles, shared postings, and blogs. And thanks to large investments in medical research as well as the efforts of the academic and scientific communities, contemporary analytical algorithms have become increasingly refined. As a result, the robustness of applications is continually on the rise.

Mechanics

New sophistication in data mining should not be confused with automation. Construction management BI is much more than tapping a smartphone app. Successful BI requires negotiating disparate data systems, for one. Equally critical is human expertise. Without real-life comprehension of construction management, a database may be subjected to irrelevant queries, or the database's answers may be misunderstood or deployed to insignificant ends.

Before an organization can grasp the basics, it must understand the foundational definition of data mining. It is the transformation of raw data into usable knowledge:



Raw data exists in myriad forms. In industries that are slow to embrace change, much data is accessible only as hard copy. The first requirement of data mining, then, is to digitize that information. In construction management cases, recurring stakeholder questions concerning acquisition, financial management, and project management become data elements in an information system; systems can range in type from Excel spreadsheets to a more complex transactional platform. For the purposes of this white paper, the term "source data" refers solely to digitized elements. It should be noted that human participation is presumed from the start of data mining—information managers are collecting and digitizing information.

Source data is inherently messy. Even if an organization were to digitize all hard-copy data using a single information system, the data elements within that information system are governed and populated by their respective business lines. As such, source data often resides in silos, making it subject to different units of measurement, workplace cultures, and other conditions.

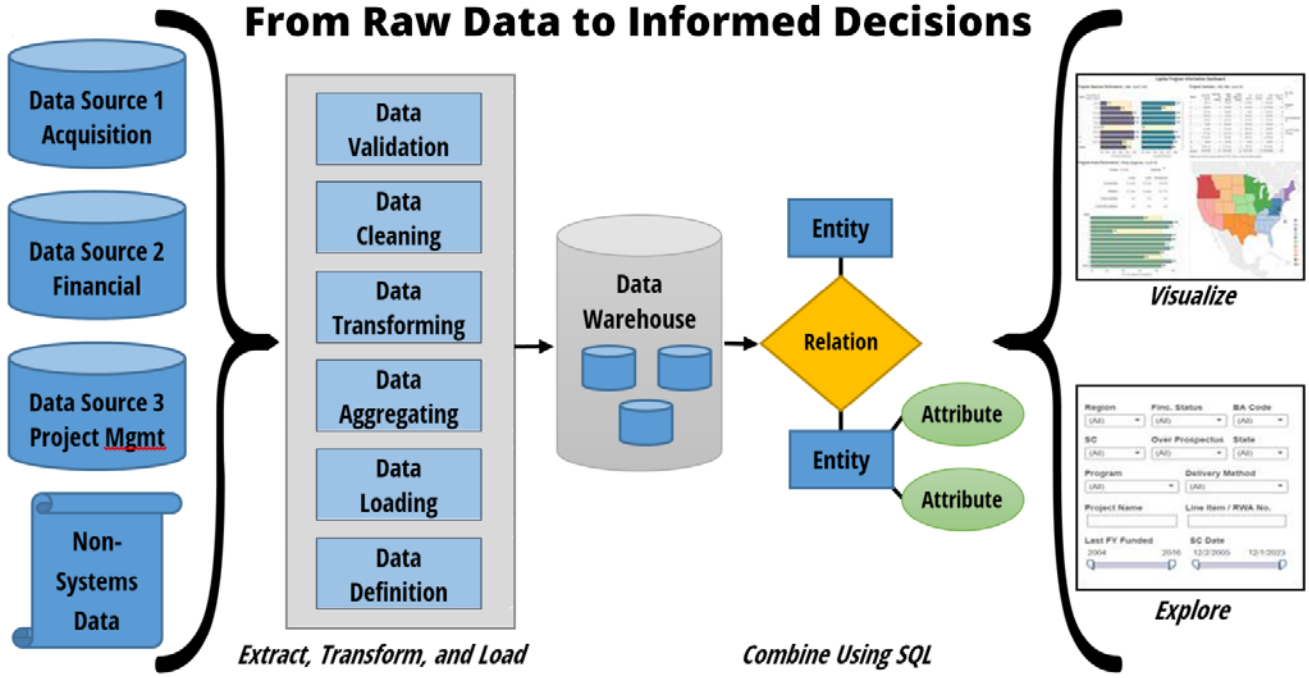
Because data typically does not originate from the same source, or is sourced inconsistently, those bits and bytes must be consolidated to a single platform. A data warehouse convenes information in one place where it can be organized and defined. Here, information managers are again called to action, specifically to scrub data so that a subsequent database query can identify meaningful patterns within them. Scrubbing data can involve many tasks, such as detecting outlier data or missing data, mostly performed via algorithm. This effort is also known as normalizing the source data. Scrubbing data may include a more sophisticated activity, like grouping data according to logical relationships or organizational preferences; preexisting knowledge of construction management distinguishes this task as more sophisticated than others.

To conduct subsequent database queries, an organization's data warehouse must connect to a **relational database management system such as SQL server** (Microsoft), Oracle Database, or MySQL.[6] To ensure that the servers correctly interact, the organization must also have skilled SQL programmers. These programmers efficiently extract and transform scrubbed data to reveal patterns, such as correlations between data elements or long-term trends. What could those correlations mean? What future behaviors can be predicted? Interpretation demands the review and dialogue of CMs, who often must be presented with the patterns in visual form to best analyze them. Visualization tools allow readers to manipulate and export data, to maximize their actionable knowledge.

In a similar vein, a data-mining investment includes development of a site where data patterns, analysis of those patterns, and actionable knowledge are all documented. This communicates decision making and, perhaps more important, the rationale underlying those choices, to an entire

organization. A publication site also establishes a system of data governance: **Data governance (DG)** *refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise. A sound data governance program includes a governing body [or group of leaders], a defined set of procedures, and a plan to execute those procedures*.[5]

In sum, data mining both organizes data beautifully and yields real knowledge. Executed properly, an organization not only understands data correlations and insights, but also knows what to do with them. Success therefore relies on tools, technical expertise, strong organization, and commitment to data governance.



With these basic elements and system requirements in mind, the next obvious question is, What can a properly deployed data mining system do for the modern CM, owner, or agency?[8]

The Benefits and Results

Based on feedback obtained from more than 2,200 confirmed users and 389 consultants on www.BI-Survey.com, the top benefits achieved from a robust data mining and BI system include[9]:

1. Faster and more accurate reporting, analysis, and planning as reported by 88%–91% of survey respondents.
2. Better business decisions (data-based decision making) based on 80%–84% of responses.
3. Improved data quality, observed by 75%–78% of respondents.

4. Improved operational efficiency, observed by 70%–74% of respondents.
5. Improved customer satisfaction was noted by 60%–66% of all respondents.

Many of these same results are being realized within our own industry, according to interviews conducted by the ETC with several leading companies, agencies, and universities within CMAA membership. Their feedback is below, edited for length.[10, 11]

Operational efficiency was cited as a key benefit at the Pittsburgh Water and Sewer Authority (PWSA), where a team from **Mott MacDonald** helped the agency establish a set of policies and procedures that are enforced using the data mining features within an established Program Management Information System (PMIS). This system allows the program manager to run reports against specific steps within the workflow, to evaluate and determine process cycle times. In turn, data mining has allowed Mott MacDonald to identify and resolve bottlenecks within the process and improve the agency's overall system performance. The time savings from this process improvement has allowed the client to avoid late fee penalties on payments to their vendors and subcontractors which in turn has led to substantial cost savings.

Faster, more accurate reporting was identified as a major outcome for **Washington Suburban Sanitary Commission (WSSC)**, where data mining has helped track key metrics such as how many miles of replacement pipe have been installed from year to year. WSSC is under a consent decree, which requires completion of several milestones addressing environmental risks. Tracking and accurately measuring progress according to the specific terms of the consent decree is essential to meeting the agency's legal obligations and community commitments.

WSSC has also benefited from the tool's **ability to work with multiple and disparate data** systems that include construction data, files stored in its enterprise file management tool, data from work orders, and data from the assessment group for sewers, all of which can be combined and correlated to display on a GIS system. The resulting GIS output helps field personnel with sewer assessments, including visual representations of that data, which helps identify and address issues with more precision. Keith Tyson, WSSC Technical Services Group Leader explains, "we recently had a discolored water issue and created GIS maps of the work order for water complaints. By tracking this information in the Maintenance Management System as a separate work order, we can pull that data into a map and show the relative strength of the issues for a particular region based on number of calls. The data is summarized in a 'heat map' to illustrate where the most calls are located, so we can see where the problem areas lie. One insight this provides is immediately identifying the reason for the calls, being that these homes are the closest to the treatment plant where we're having an issue. We can also correlate that data to compare the age of pipes versus discolored water complaints, and figure where we need to target our flushing—open fire hydrants to flush out the manganese particles in the water or try to scour the pipes to get rid of the film that's causing this."

In another example, **Northeastern University** explained how its project management office manages approximately $150 million in annual spending across 130 campus projects. Gary Younger, Project Controls Manager, has developed a series of executive dashboards that combine metrics from processes, as well as budget and schedule data, to provide a complete view of project status to executives. The facilities team can now track and compare data related to the cost per square foot for different facility types, projects trending over budget, process aging data (number of days to complete a process), and more.

Carlo Orsenigo, **CH2M**'s Senior Vice President of Global Enterprise Services, shared his perspective that "each company has a different starting point. For CH2M, the transition to BI and data mining was born out of necessity. Our company had tremendous customer support and great delivery, but in many cases we were suffering for scarce project management discipline and poor change management. Prior to making the change to BI, we had a variety of tools designed to track our internal cost performance, but they tended to be historically based and had us driving with the rearview mirror, affecting our forecasting accuracy. Once we made the switch, we realized how powerful it was for us to be able to summarize and synthesize our key performance drivers. BI allowed all levels of the organization to access the data, from the project manager up to the CEO. Previously our data had been predominantly tabular, so it was very powerful to see the data visualized, particularly in terms of the traditional S-Curve of actual versus plan. The visual impact was immediate. Today we no longer have discussions without the data being right there in the room. High-level decisions have advanced from perception to a real-time, data-driven process."

Data Mining Case Study: The GSA Story

Contributing editor, David Sokol

This case study chronicles how the U.S. General Services Administration (GSA) went from identifying a need for data mining to full deployment of an effective BI system. Included is a detailed account of the decisions, challenges, and rewards surrounding system deployment and implementation.

Introduction

GSA is responsible for managing 370 million square feet of owned and leased real estate, in which 1.1 million federal civilian employees provide services to the American people. The cost-efficient, timely delivery of federal courthouses, office buildings, land ports of entry, and other facilities impacts the productivity of civil servants, and influences taxpayers' attitudes toward the federal government.

As an owner and operator of federal civilian facilities, GSA undertakes renovations, restorations, modernizations, and new construction on behalf of its federal tenant agencies through the Public Buildings Service. The Office of Project Delivery (OPD) is responsible for this work, which is partly paid by the Federal Buildings Fund; alterations and service upgrades are reimbursed by GSA's federal-agency tenants. In any given year, OPD delivers approximately $9 billion in capital projects and tenant improvements. This figure normally includes an annual infusion of $1 billion in new Federal Buildings Fund moneys via Congressional appropriation.

Because proximity is essential to excellent project delivery, individual projects are delivered through 11 regional offices, which report to a national office. Formulating the capital program takes place first. This effort requires synthesizing regional priorities into one national investment strategy; identifying delivery strategies for each project; and ensuring that the proposed project will fulfill tenant mission, operate efficiently, and provide value to the taxpayer. Individual projects are then executed by regional teams, while the national office oversees and optimizes regional performance

In this structure, OPD's national office fields queries from Congress, federal tenants, taxpayers, and the media in addition to conducting internal oversight. OPD must manage significant data to accomplish all these tasks. It tracks scope, schedule, and budget to synchronize regional and national understanding of project delivery within GSA. National and regional OPD colleagues also rely on data to provide the best customer service to federal tenants, some of which are especially proactive about accessing data.

Data has assumed different forms since OPD's inception. Printed fact sheets have given way to multiple electronic platforms for deploying and documenting contract awards and modifications, design and construction progress, funding outlays, and more (hereafter referred to as enterprise systems). Yet whether analog or digital, historically GSA has struggled with data. Over time, for example, data systems have grown to respond to the needs of specific business lines, and inconsistent data structures

have developed. As a result, information was inconsistently coded at its source, or the information collected from a weekly data call would become obsolete by the time it reached a tenant or other key stakeholder several days later.

An early solution to this dilemma was the Project Information Portal (PIP), an Oracle-based application. The portal extracted some data from GSA's enterprise systems; remaining data elements required hard-coding. Despite its shortcomings, the PIP provided an unprecedented enterprise-wide view of the capital program in whole.

Nascent Dashboards

In 2009, GSA was appropriated more than $5.5 billion to convert federal buildings into high-performance buildings, and to build new, energy-efficient federal buildings, courthouses, and land ports of entry—in other words, more than five times its average capital work. New information requirements accompanied that funding, as did numerous stakeholder requests for accurate data. In fact, OPD received at least three inquiries concerning timing, budgets, and job creation and other economic metrics every day—an unprecedented rate.

This period of intense activity in GSA's capital program underscored the need for a Project Information Portal. It also revealed the existing tool's shortcomings.

Security issues concerning firewalls and database platforms rendered the PIP obsolete during this period, and it had to be terminated. Because the PIP's shutdown coincided with persistent and rising data demands, OPD's national office contracted for surge support from consultants to meet the ongoing need for an enterprise-wide view of the capital program. In turn, the consultants tracked and reported scope, schedule, and budget of hundreds of capital projects, again through a combination of data harvesting and hard-coding. That data was transferred to, and manipulated in, spreadsheet files, and the native functionality of the spreadsheet program provided a nascent dashboard perspective of nationwide delivery.

The consultants' activity revealed that available delivery data was trustworthy, but that hard coding data from enterprise systems was both redundant and labor-intensive. Response times to queries showed little improvement. A Congressional, media, or Freedom of Information Act request could demand as much as five hours of research. Because the process relied on outside services, it could not be responsibly sustained over the long term, from a taxpayer perspective.

OPD's national leaders and their regional confidants envisioned a more organic alternative—a dashboard that eliminated hard-coding. The tool would harvest obligations, outlays, and other new delivery data as project managers directly entered it into enterprise systems. By leveraging existing enterprise activity, OPD sought to eliminate data calls, eliminate duplicate data entry, and improve data quality overall.

The Road to CPID

Moving toward a sophisticated dashboard required buy-in from various stakeholders within the agency, including the Office of the CIO. Past experience indicated that, when GSA had invested in tools like the PIP, significant funds were required for data upkeep and maintenance of custom elements. OPD ensured the CIO that it instead preferred a BI platform that gathered existing enterprise information from several systems, and which converted that information into an interactive dashboard controlled by GSA credentials.

As OPD earned the CIO's trust about BI, it also sought ways to trim GSA's operating budget. Due diligence reaffirmed a tool that automates data aggregation and display across enterprise systems. OPD also assured its stakeholders that undertaking BI would leverage existing resources and sweat equity, and require only spending new funds for software licensure. The Capital Project Information Dashboard (CPID) Initiative was ultimately authorized to proceed, because it promised to leverage existing data, reduce maintenance costs for custom features, and automate data aggregation.

Because one cannot expect accurate reporting from information that is merely dumped into a data warehouse, the CPID Initiative focused initially on governance. By establishing a project information intake process for enterprise systems, the CPID team standardized key common data elements. It also identified errors and misinterpretations that could corrupt data: one enterprise system may consider a project with all its phases as one project, whereas the next system might separate out each phase as its own 'project,' for example: OPD tracks projects at many levels, be it work items within a task order, task orders within a phase, or phases within a project. Therefore, once a program of projects is authorized and funded, the project information intake process ensures that key data elements that serve as the glue among the enterprise systems follow the correct standard or data validation. Adhesion of enterprise systems makes the automation of tracking projects at any level possible.

Upon resolving data leaks, the team produced a multiple-element BI, and coded domains for the platform using SQL. The framework can combine information from any business line, as long as their enterprise systems have one data element in common. The CPID team was responsible for conceiving the most relevant and enlightening combinations of data, and for coding them efficiently so that information could be refreshed or queried as quickly as possible. It also deployed Tableau to create graphic data representations from CPID selections, as even a world-class dashboard can benefit from a comprehensible, aesthetically appealing presentation.

GSA Findings and Results

CPID has transformed the query process. Today, the average self-serve page view costs GSA about $1 and 20 minutes of the individual user's time. It is no longer necessary for a designated professional to perform a five-hour query. Confidence in data accuracy also is higher than ever, which has impacted

credibility external to GSA. Time- and source-stamped information has reduced contention between stakeholders, in particular.

OPD's regional project managers have embraced CPID simultaneously. They understand that everything a federal employee does on a government computer is the official business of government, and they have enthused to data harvesting as a method for minimizing data calls. The time savings has created goodwill between OPD's national and regional offices. Thanks to CPID's data standardization, regional project managers also have greater confidence that national-level colleagues and federal leadership are interpreting their job performance objectively.

Since CPID rolled out in 2015, OPD's national office has begun shifting its reliance on contracted consultants to in-house talent. These colleagues are rewriting SQL code to improve existing data combinations. Rewrites are also dedicated to increasing the efficiency of dashboard updates, so that future updates can accommodate increased project volume or data load without sacrificing speed. GSA is writing all-new SQL code internally, in addition, to reflect the reporting requirements that evolve over time: OPD now pulls data from the previously untouched Occupancy Agreement enterprise system, so that tenant agencies can understand the relationship between a construction project and its rent, for example. Staffers are working in Tableau to continually improve the presentation of data, as well.

The CPID Initiative has had wide ramifications for a variety of data management issues. The effort is shedding new light on data quality in enterprise systems. It underscores the need to comply with data dictionary rules, and has shown how small improvements like data-validating the contract number field can make a big difference in data aggregation generally. The CPID Initiative also is illuminating new goals. Digitizing still-analog source documents and dovetailing that information into enterprise systems promise to impact the dashboard and its usefulness profoundly.

Conclusions

The CM Industry is slowly coming to terms with the reality that technology is changing everything around us, and that data management is fast becoming a key business differentiator for many leading companies. Events like CMAA's National Conference and other industry educational and information-sharing initiatives are raising awareness of these new rules of engagement. Meanwhile, industry leaders, including the members of the ETC, should continue raising the bar for applying new technology to construction and program management, to advance the profession overall.

References

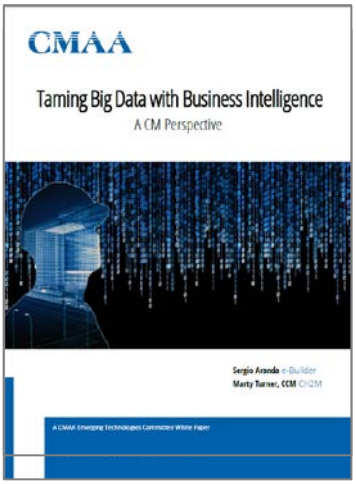
These references are noted throughout the paper and describe the source of the quotation and/or material included in this paper.

1. Vcloud News, "*Everyday Big Data Statistics*," April 5, 2015
2. http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily test
3. Kimmorley, S. "*INFOGRAPHIC: Here's how much data is created on the web every minute*," Business Insider Australia April 19, 2015 https://www.businessinsider.com.au/infographic-heres-how-much-data-is-created-on-the-web-every-minute-2015-8
4. Woodie, A. "*Why Self-Service Prep Is a Killer App for Big Data*," May 31, 2016
5. Data Governance definition adapted from www.whatIs.com search.
6. SQL definition adapted from www.sqlcourse.com/intro.html
7. Konana, P. "*Basics of Data Mining*," YouTube Video Published on Apr 6, 2016. https://www.youtube.com/watch?v=u2oSiVOQRmg
8. Berger, C. " *Predictive Analytics is Data Mining's Future*," BioIT World, Oct 25, 2017. http://www.bio-itworld.com/newsitems/2005/06-05/06-23-05-news-oracle/
9. www.BI-Survey.com
10. Gartners Magic Quadrant – 2017 Version. https://www.gartner.com/doc/reprints?id=1-3TDP4W1&ct=170213&st=sb
11. Dresner, Howard. "*2017 Wisdom of the Crowds Business Intelligence Market Study*," Dresner Advisory Service, LLC, May 31, 2017.

Just for fun, check out https://www.usafacts.org/

Incredible data mining site that compiles US Government Data

Also by the CMAA Emerging Technologies Committee:

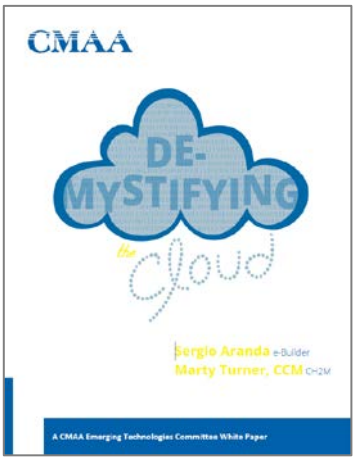


Taming Big Data with Buisiness Intelligence: A CM Perspective

Sergio Aranda, e-Builder | Marty Tuner, CCM, CH2M

Television programming abounds with commercials promoting the virtues of Business Intelligence (BI). You have all seen the ads telling of "a new way to work" or "how you can make smarter decisions" or "how you can act faster and perform better." Sounds great, but were you left wondering if Business Intelligence was really needed in the world of construction management or if the promises of mining Big Data had any useful applications in helping you run your project, your region or your business operations?

The CMAA Emerging Technologies Committee pondered these very same questions and set about to find the answers. This Whitepaper seeks to explain what Business Intelligence is, what it does and why every CM professional should be applying this technology on their projects and in their organizations.

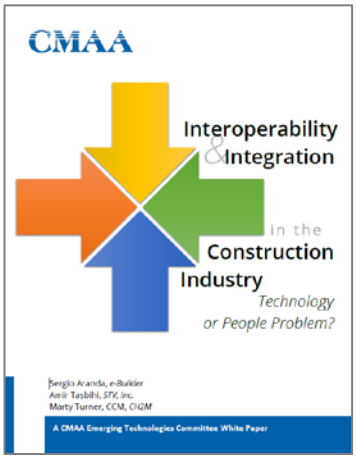


De-Mystifying the Cloud

Sergio Aranda, e-Builder | Marty Tuner, CCM, CH2M

You've probably heard a lot about the cloud and how it is revolutionizing the way "everyone" does business these days. But what is the cloud and why is it important to you and your business as a CM practitioner? The CMAA Emerging Technologies Committee is here to help take away the mystery and provide you with a starting point to understanding what cloud computing is, and how you can make it work for you.

So what is the cloud?



Interoperability & Integration in the Construction Industry: Technology or People Problem?

Sergio Aranda, e-Builder | Amir Tasbihi, STV, Inc. | Marty Tuner, CCM, CH2M

interoperability is achievable, and there are tremendous benefits waiting to be reaped by those willing to invest in it. This white paper by the CMAA Emerging Technologies Committee will provide an in-depth look at the characteristics, benefits, challenges and effective techniques used to select the right programs for a project and then subsequently integrate them all into a one, highly efficient system in order to meet your project needs. It argues that the people using the technology present a bigger obstacle to achieving their goals than the technology itself. It argues that the people using the technology present a bigger obstacle to achieving their goals than the technology itself.

Download these and other articles & white papers at www.cmaanet.org/articles-white-papers

CMAA Advancing Professional Construction and Program Management Worldwide

CMAA is North America's only organization dedicated exclusively to the interests of professional construction and program management and is uniquely qualified to serve both owner and service provider practitioners. CMAA builds business opportunities for its members by enhancing the identification of CM/PM as a valuable professional service. Members gain unsurpassed professional development and networking opportunities, and organizations can strengthen both their capabilities and their marketing opportunities. CMAA responds to the needs of the CM industry – owners and service providers alike.

Our CM Standards of Practice and the CMAA Body of Knowledge are the cornerstone of certification by the Construction Manager Certification Institute. Today almost 3,000 professionals have earned the distinction of becoming Certified Construction Manager.

If your organization is serious about providing CM/PM services, then membership in CMAA is a necessity.

www.cmaanet.org

THE MISSION of CMAA is to promote the profession of Construction Management and the use of qualified Construction Managers on capital projects and programs.

THE VISION of CMAA is that all owners will realize capital project and program success by using professionally qualified Construction Managers.

CMAA

7926 Jones Branch Drive

McLean, VA 22102 USA

703.356.2622

www.cmaanet.org