A Systematic Approach to Determine the Amount of Data Required for Asset Management Decisions

Brendan Maestas, Sean Stuntz, Joey Applebee, William Bentley, Jared Breuker, Andrew Davenport and Andrew Hoisington

September 18, 2019

A Systematic Approach to Determine the Amount of Data Required for Asset Management Decisions

Brendan Maestas[1], **Sean Stuntz, Joey Applebee, William Bentley, Jared Breuker, Andrew Davenport, Andrew Hoisington**

Facility, infrastructure, and asset related data is being generated at an unprecedented rate, usually without specific purposes or goals. Data is collected in large amounts for exploratory science, achieving significant statistical power, due to the relatively cheap cost of storing data in the cloud. In many cases however, organizations do not consider the negative issues with indiscriminate data collection to include diminishing returns to reduce uncertainty in asset management decisions and the cumulative costs of the data. This paper proposes a novel 4-step framework for determining the correct amount of data required for asset management decisions. The framework is built upon the following steps: 1) identify the problem, 2) establish context, 3) verify/collect data, and 4) analyze/decide (IEVA). The IEVA framework can be used as a baseline that orients asset managers to collect decision-focused data and make data-informed decisions.

1 Introduction

Advances in computing and communication technology have propelled the world into an unprecedented information age. For example, the computer system that supports the Large Hadron Collider (LHC), the world's largest particle physics laboratory, can process approximately one petabyte of data every day (European Organization for Nuclear Research (CERN) 2017), equivalent to 210,000 DVDs. The server grid that supports the LHC is only able to actually store 45 petabytes of data, and must rely on networked computers around the world for an additional 15 petabytes (European Organization for Nuclear Research (CERN) 2017). Meaning scientists at the LHC are unable to store 84% of the data that they process, which says nothing about the overwhelming amount of data that is being

[1] Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

generated and not processed. Although the magnitude of data processed and ignored by the LHC is certainly an extreme, the trend of collecting more data than feasible to use is common. In the construction and asset management industries, large amounts of data are generated, collected, and stored that is never actually analyzed (Hammad et al. 2014). Examples like these show how data managers around the world, are “drowning in data while thirsting for information” (Herrmann 2001).

Datakleptomania is an unconscious desire to collect increasing amounts of data under the premise that more data is better (Hellawell 1991). However, for many businesses, success is the direct result of collecting meaningful data and extracting useful information to support strategic decisions (Woldesenbet et al. 2016). There are some significant statistical benefits of collecting more data. Through increasing the number of observations, researchers are able to achieve greater confidence that their sample is representative of the population. When a sample can be said to be representative of the population, then conclusions about the sample are more likely to be valid for the population (Cohen 1992).

The practice of datakleptomania is enabled by the relatively cheap cost of storing data. As of this writing, Google’s cloud service offers the ability to store 100 gigabytes of data for a subscription cost of $12/year (Google 2017). With the low cost and high accessibility of cloud storage, additional companies are storing data and doing business in the cloud. Because of this low cost of data storage, and the technologies that make collecting data even easier, the world is trending toward unquestioned data collection with less scrutiny regarding its necessity (Hanley 2012).

Since data is a resource that can provide value to an organization, data should be purposefully managed using asset management principles. In doing so, organizations will need to view their data within a lifecycle analysis context that considers costs, conditions, and performance of their data from creation to the deletion. However, many data collection activities aren’t designed to support decision making processes (Flintsch, G, W; Bryant 2009). Data should be collected in such a way that the intent for its use is clearly defined, ensuring that the methods for collection, analysis, and use can be appropriately tailored from the start (Hanley 2012).

2 Problems with Big Data

Beyond the diminishing returns of additional data, the quantity of data that is being collected can be problematic. Datakleptomania can be useful in exploratory sciences where little is understood and questions will be developed later. However, the field of asset management may not be the best example of exploratory science. Instead of collecting every piece of information available, infrastructure asset managers should focus on collecting information that helps answer known questions. Indeed, collection of large quantities of data can lead to 'analysis paralysis' where decision makers have so much data they don't how to move forward with decisions.

Another problem is the time that it takes to process vast quantities of data can still be a challenge. Reducing the amount of data reduces the amount of time it takes to process and analyze that data. This in turn allows decisions to be made faster. Pat Helland explains another problem with analyzing too much data. He states that when the time it takes to process large amounts of data exceeds the window of time in which the decision must be made, the data being processed becomes obsolete by the time the decision is made (Helland 2011). Helland recommends that in such situations, approximating a good answer can be more valuable than taking time to develop the perfect answer.

Labovitz et al. (1993) developed a notional rule to describe the cost of process correction based on the quality of management to that process. The rule was meant to be applied to a wide range of applications and is known as the 1-10-100 rule (Labovitz et al. 1993). Doyle (2014) tailored this rule to apply it to the costs of data collection and management, namely, that as the quality of data management efforts decreases, the costs of using the data later increase significantly. Data collected early and deliberately with the intent to be applied to later decision-making efforts yield relatively low cost to the organization. The organization has failed to implement effective data management strategies and will more than likely result in poorly informed decisions which must be corrected down the road at a high cost to the organization. Data collected without proper forethought will more than likely result in time intensive analysis and require a large cost in human capital.

Once data is collected, an organization may desire a storage capability to use the date at a later date. Although the storage cost of data has decreased steadily as technology and management tools improve, it is still a portion of the costs (LaChapelle 2016). This data storage capability requires either data infrastructure investment or regular storage fees for use on the cloud. Several factors influence data storage requirements. The organization must first determine the useful life of the data stored and also decide on a meaningful format for analysis and communication. When optimizing data management, an organization must also consider the lifecycle cost of storage. The cost of keeping a single gigabyte of data indefinitely can cost $100 (Omaar 2017), a cost far greater than the monthly cost of using the Google cloud or similar service. This cost is further compounded for organizations whose security policies prohibit the use of the cloud for storage purposes. Redundancy and security also play a heavy role in this process. The organization must build into their storage plan their requirement of backing up the data and ensuring its security from outside threats. They must decide how much risk they are willing to take and in which areas they are willing to accept this risk.

A final problem with big data occurs during data transformation, which is the process of transitioning from a set of stored raw data to usable data ready for analysis. Although it is recommended that the data is already in a useable format when collected, it may not be possible until new equipment or systems are installed. In this transformation effort, there are generally aspects of the raw data which are lost due to translation or conversion issues inherent with diverse formats and equipment systems. The goal is to perform these transformation efforts with minimal costs to the organization by way of lost or mistranslated data. Establishing a consistent transformation strategy to minimize these losses is critical to an organization ensuring their efforts are effective and useful. This transformation cost can also be incurred if the organization decides to upgrade operating systems or interfaces. In these situations, a well-developed plan and strategy are critical to successful transformation of data with minimal loss of information. This effort can be costly and time consuming in and of itself, but will be worth it in the future.

3 Notional Framework for Right Size Data

Any organization which seeks to optimize their decision making and data collection strategies benefit from a simple and effective framework to guide them through that process. A decision-making framework allows organization to simplify their data collection and analysis process. This in turn will enable trend tracking and decision making in the future by eliminating shortfalls in their data, eliminating unnecessary data, reducing decision making time and cost, and do this all to the organization's self-set standards. The 4-step IEVA framework presents a strategic level approach for an asset manager to identify the problem, establish the context of the problem, verify and collect appropriate data, and finally analyze and make a decision. Each step of this strategic asset management decision making framework can be seen in Fig. 1 and is explained in further detail throughout this section.

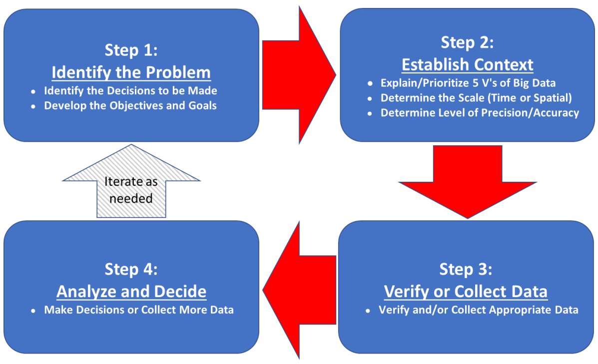

Fig. 1. IEVA Framework Steps

3.1 Step 1: Identify the Problem

The initial step of this strategic framework is determining the overall decision to be made. Asset managers and their respective teams must brainstorm during this step to narrow down and identify the overall decision to be made. Once the decision is understood, the next step is to develop objectives and goals, so that the decision can be fully developed. Finally, this decision should be emphasized throughout the 4-step framework.

Now that a decision has been identified, set goals and objectives specifically for this decision. These goals will be specific to each asset manager and their respective objectives. For example, if the decision is to save energy costs across an installation, the goal could be to reduce electrical usage by twenty percent in the next five years. To increase effectiveness, goals and objectives should be specific, measurable, attainable, relevant, and time bound (SMART) (Mind Tools Content Team 2014). A specific goal must be clear and well defined, and this goal should guide decision making throughout the rest of the framework. A measurable goal is one that sets a clearly defined objective. In addition to measurable, the goal should be attainable and realistically accomplished. A relevant goal will help keep the decision in mind and keep the goals focused. Finally, a time-bound goal will create a sense of urgency and allow for an achievement if and when that goal is met.

3.2 Subheading: Step 2: Establish Context

Once the problem is understood, a user should prioritize the five V's of data for the specific objective and goal (Marr 2015). The five V's are volume, velocity, variety, veracity, and value. Volume is the vast amount of stored data which is gathered every day from multiple sources. The value of data is defined simply as the ability of a person to convert raw data to performance metrics. When applying a value determination to this framework's steps, it is important to ensure that the data which will be leveraged is being collected in the first place. Velocity refers to the speed at which data is gathered, which can be dependent on the equipment or data storage capacity. Veracity is the trustworthiness of the data. Lastly, variety refers to the multiple different types of data which can be collected (Marr 2015).

Overall, examining the five V's of data provides insight on how data is collected and measured (Marr 2015). It is important for users of this framework to ask these data questions before going ahead in their decision making. The order at which the five V's are considered are dependent upon the questions being asked and they should be weighted accordingly. The overall goal of this step is for the user to look at the question they are asking and mold the data collection for this purpose. In many cases where the five V's are not considered, the available data forces decision makers to alter the question they originally wanted to ask (Marr 2015).

Understanding the scale of the data will aid in this process to determine the right amount of data needed to make a decision. This step is to ensure there are boundary conditions for the problem being solved. As an organization steps through the process, it can be easy to add scope to the original question. Data scales can either be spatial or time related. A spatial scale would evaluate or collect data based on regional zones, facility types, utility infrastructure, or rooms in a facility. A time scale refers to where the beginning and end of the useful data occurs. Overall, any scale could be used by the decision-making team or individual, and picking one and justifying it will help make the overall decision and help identify the amount of data needed for it.

In addition to scale, precision and accuracy are important factors to consider to create a useful framework. Precision can be modelled by looking at the variance of the data set. As the variance increases in the data set, the data is less precise. On the other hand, accuracy is the trueness of the data set. The data's velocity can change the accuracy and precision of the data due to the number of samples that are usable. Once the five V's have been adjudicated, the next step is to collect data.

3.3 Step 3: Verify or Collect the Data

Deciding how to collect and aggregate data contributes to data management life-cycle costs and data quality. There are many software options which offer a degree of automation and analysis. It is therefore pertinent for each asset manager to research these options and balance the potential for errors during collection with software acquisition costs. In addition to software, some data collection may be dependent upon manpower, also contributing to increased data acquisition costs. Overall, there are many different ways to collect data, and each asset manager must balance which way is most efficient and practical for the decision they are making using this model.

3.4 Step 4: Analyze and Decide

After the appropriate data has been collected and consolidated, the next step of the framework is data analysis. This analysis will be dependent upon the objectives and questions which were developed earlier in the framework. These objectives will aid in distinguishing which statistical measurements and tests are required for a proper conclusion to be made. Each asset manager's analysis will be completely dependent to the overall decision being made. There are multiple methods of statistical analysis that can be simple or complex requiring different levels of mathematical foundation. Understanding the desired method and using any available aids, such as excel and JMP software, will decrease the time required and increase the useful output of the analysis effort.

The final step of this framework is to make the decision. At this point, the user of the framework has collected the data and analyzed it accordingly. The data has been specifically gathered to ensure the overall question will be answered in an effective and efficient manner. However, if the situation presents itself where data is insufficient, the team should return to Step 2 to re-evaluate what needs to be measured to accomplish the goal. This is designed to ensure that the data being collected truly relates to the original objective set forth by the team. If the data is sufficient and executable, the team shall move to the decision-making.

These steps were created to ensure that data collection is closely monitored and gathered with a specific purpose. The best-case scenario of this framework is that the user finishes the steps and comes to a data-validated answer to their specific question. They have the ability to make a decision and influence positive changes in their organization. In a worst-case scenario, the user will complete the framework with a better understanding of how to effectively return to the prior steps to ask a question better suited to the desired results.

4 Conclusion

Modeling a data management framework is an important tool for asset managers to sift through the large data quantities being collected. The 4-step IEVA framework provides a method to give realistic goals and objectives for asset managers to utilize while limiting the effects of uncertainty, quantity of data, and cost factors. The 4-step IEVA framework is the first attempt to address the issues of big data and utilizing a system to tailor the data for effective analysis. For future research, this framework can be applied to other asset management decisions regarding condition assessments for facilities and utility systems. Regardless of future applications, the 4-step IEVA process provides a framework for determining the amount of data required for asset management decisions.

Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government.

References

Cohen, J. (1992). “Statistical Power Analysis.” Current Directions in Psychological Science, 1(3), 98–101.

Department of Defense. (2016). “DoD Building Code (General Building Requirements).” Washington DC.

Doyle, M. (2014). “Why Data Should Be a Business Asset – The 1-10-100 Rule.”

European Organization for Nuclear Research (CERN). (2017). “Computing | CERN.” <https://home.cern/about/computing> (Dec. 8, 2017).

Flintsch, G, W; Bryant, J. . (2009). “Asset Management Data Collection for Supporting Decision Processes Asset Management Data Collection for Supporting Decision Processes.” 2–97.

Google. (2017). “Google Drive Pricing Guide.” <https://www.google.com/drive/pricing/> (Dec. 8, 2017).

Hanley, J. R. (2012). “Marine environmental monitoring programs: tips on design, measurement and interpretation.” The APPEA Journal, 52(1), 317.

Helland, P. (2011). “If you have too much data, then ‘good enough’ is good enough.” Communications of the ACM, 54(6), 40.

Herrmann, K. R. (2001). Visualizing Your Business: Let Graphics Tell the Story. Wiley, Hoboken, New Jersey.

Hubbard, D. (2014). How to measure anything: finding the value of intangibles in business. John Wiley & Sonds Inc., Hoboken, New Jersey.

Labovitz, G. H., Chang, Y. S., and Rosansky, V. (1993). Making quality work: a leadership guide for the results-driven manager. Harper Business, New York, NY.

LaChapelle, C. (2016). “The Cost of Data Storage and Management: Where Is It Headed in 2016? - The Data Center Journal.” Data Center Journal, <http://www.datacenterjournal.com/cost-data-storage-management-headed-2016/> (Dec. 5, 2017).

Marr, B. (2015). “Why only one of the 5 Vs of big data really matters.” IBM Big Data & Analytics Hub, <http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters> (Dec. 8, 2017).

Mind Tools Content Team. (2014). “Five Golden Rules for Successful Goal Setting.” <https://www.mindtools.com/pages/article/newHTE_90.htm> (Dec. 8, 2017).

Omaar, J. (2017). “Forever Isn’t Free: The Cost of Storage on a Blockchain Database.” <https://medium.com/ipdb-blog/forever-isnt-free-the-cost-of-storage-on-a-blockchain-database-59003f63e01> (Dec. 5, 2017).