MANAGING DIGITAL LIBRARIES: THE VIEW FROM 30,000 FEET

Organizational considerations initiating a big data and analytics implementation

Received 9 May 2016
Accepted 9 May 2016

H. Frank Cervone
School of Public Health, University of Illinois at Chicago, Chicago, Illinois, USA

Abstract

Purpose – Organizations are beginning to realize the potential benefits of big data and harnessing all of the data they are creating. However, a major impediment for many organizations is understanding where to start in big data and analytics implementation. In many respects, starting a successful implementation is not much different from any other project managed within the organization. The major stumbling block is knowing what questions to ask to get things going. This paper aims to help libraries and information organizations that are considering big data and analytics implementation to begin their journey by following a checklist of eight aspects to be considered in the development of a big data and analytics strategy.

Design/methodology/approach – The eight aspects to consider in big data and analytics implementation were developed using a combination of existing project management common knowledge, consultant recommendations and real-life experiences.

Findings – Organizations considering big data and analytics implementation need to explore aspects related to the data they have, what organizational problems they are trying to solve, how data governance will work in the new environment, as well as how they will define success in terms of their implementation. These are in addition to the technical issues one would normally expect in a systems implementation.

Originality/value – While there have been many articles written about the implementation of big data and analytics in organizations, most of these focus on technical issues rather than managerial and organizational concerns. In addition, none of these other articles have been from the perspective of library and information science. In this article, the focus is specifically on how information professionals may approach this problem.

Keywords Big data, Analytics, Analytics project implementation,
Big data business concerns, Big data project implementation,
Organizational issues related to big data and analytics in the information professions

Paper type Viewpoint

As the amount of data being generated continues to grow exponentially (CSC, 2012), more and more organizations are beginning to realize the potential benefits of harnessing all of the data they are creating. Traditionally, most of the data created within organizations has been the result of intentional, transaction-based systems. However, today, a far greater amount of data is generated as a byproduct of various activities that are not transactional in nature. This includes such things as text streams

Digital Library Perspectives
Vol. 32 No. 3, 2016
pp. 137-141

2059-5816
DOI 10.1108/DLP-05-2016-0013

from social media postings, computer system logs, sensor information from environmental detection equipment and other devices and sources of data which in the past might not have been considered useful or meaningful inputs for either operational concerns or strategic planning.

As our understanding of how this data could be used to our advantage, one of the major impediments for many organizations is getting a handle on how to start using big data and analytics. While it may seem like a daunting task, in many respects, it is similar to any other project managed within the organization. You just need to know what questions to ask.

To get started on your big data and analytics journey, here are eight factors to consider in the development of your organization's plan.

1. Identify what you have to work with and start collecting it

It is likely you are already generating a significant amount of data within your organization. What you may not be doing is collecting it in an organized or methodical manner. Consequently, the first step you should take is to identify what data is being produced through various systems and input streams and start collecting the data if you are not already doing so.

Some typical activities that might be necessary to gather data not currently being collected include turning on logging within systems or changing the types of data that is being logged to include more of the events that occurring. For example, enabling detailed application logs to supplement the various operating system logs already being captured.

Where the data is already being collected, the major task will be to determine how to ensure the data you need in your "data lake" is captured and stored, so that it is easily accessible for processing later on. For instance, system logs often have a "trim" date which is a predetermined amount of time the data is kept available, such as 60 or 90 days. Once the trim date is met, if the data is not otherwise copied or backed up, the old data is purged. As part of your evaluation process, you will want to ensure your existing retention periods are adequate, if they need to be increased, and if your back up and retention plans and systems are functioning as required.

2. Identify business problems

Some organization may start with this point because the most significant question related to big data or analytics initiative is "What is the problem we are trying to solve?" Many approaches to big data seem to imply that simply mining data will result in useful outcomes. This is not true. Unplanned data mining is unlikely to result in positive outcomes and can create problems for the organization when mined information is used indiscriminately (Waxer, 2013).

Therefore, understanding what particular organizational issues big data can address is critical to ensuring that your efforts are directed in a way that will facilitate success. Based on their work related to big data, the consulting firm Deloitte (2014) has developed a preliminary list of questions a big data project manager should ask to develop a solid footing for big data initiative. Adapted for a library or information organization environment, these questions are:

Q1. Do the problems under consideration for big data initiative directly relate to the strategy of the organization?

Q2. If those problems were to be solved, would those problems have the potential to increase/improve services or reduce costs?

Q3. Who within the organization has expressed interest in big data and why? What influence do they have in either helping or hindering big data effort?

Q4. Which area or functions in the organization may benefit the most from applying big data to operational decisions?

Q5. Are the issues identified so far long-standing problems the organization has not previously been able to address because of the limitations of available technology and data?

Q6. How can we measure the financial impact of applying big data to these problems?

If the answer to *Q1* or *Q2* is "no"; a big data solution is probably not appropriate for the organization. If you cannot provide definitive answers for questions three through six, success may be elusive.

3. Consider data from a functional perspective

While it may be easy to start collecting data, eventually you will need to perform a value analysis to understand the importance of what is being collected. In particular, you will need to ask which data is most relevant and provides the greatest advantage for your effort and focus on that.

For example, consider the case where you need to collect transaction log data to ensure Health Insurance Portability and Accountability Act compliance. In this case, the focus should be on collecting data from sources where protected health information may reside.

Consider also what data might give your organization a competitive advantage or help you better serve your customers. Social media analysis is often a major factor in this area because analyses of social media, such as Twitter sentiment analysis, can provide a good early warning indicator of problems within your organization that may not be self-evident from traditional indicators you may be using.

4. You would not be able to abandon current data management and warehousing solutions

Big data initiative is unlikely to replace your existing systems. In most cases, big data and analytics supplement existing systems by providing additional and complementary information, functionality and services that cannot be implemented in your current environment or with current tools.

An aspect that can cause confusion is how big data relates to traditional data warehousing and business intelligence (BI) platforms. This question is not easily answered mainly because the environment is evolving so rapidly. While most of the attention in big data and analytics has been focused on solutions, such as Hadoop and Cloudera, the reality is that things are not so clear cut. As traditional data warehouse platforms are adapting to the world of big data, the differentiation between the two environments becomes fuzzy.

Perhaps the most important question to be addressed before getting too deep into an initiative is understanding what your current data warehouse or BI vendor (if you have

one) is planning related to big data and analytics functionality. You will need to consider whether those vendor-provided solutions are feasible and cost effective in comparison to open source solutions, such as the Hadoop stack.

5. Consider your ability to support an infrastructure internally

For many organizations, this aspect may not even be an issue because you have already outsourced the hosting of your infrastructure to another organization. However, for environments that do have an internal computing infrastructure, a factor to consider is whether it makes sense to develop an internal big data environment or use a cloud-based solution instead. Especially at the early stages of big data initiative, a cloud-based solution may make a great deal of sense. With capacity on demand solutions, such as those provided through Amazon or Google, your organization does not need to worry about constantly monitoring or upgrading capacity. The downside, of course, is that one must be mindful of the financial implications of unrestricted data bacause costs could quickly mount depending on the nature and volume of what is being collected.

6. Self-service

Just as it has been a priority with on-going BI and data warehousing efforts, the ability to provide self-service access to data for your user community is critical to the success of your big data initiative. With proprietary solutions built off an existing data warehouse infrastructure, implementing dashboards and self-service data may be easier because you can reuse familiar tools.

However, with an open source big data implementation, the path is less clear. There is no consensus over which open source tool is best for implementing a dash board solution and no product has a clear leadership position. Some open source products to consider are Lumify (lumify.io), Bouquet (www.openbouquet.io), H20 (www.h2o.ai) and Shiny (for those who program in R, http://shiny.rstudio.com/).

7. Data governance

Data governance in any environment is crucial but with big data it is critical. In addition to the issues related to security, integrity and provenance of data, the sheer volume of information in a data lake demands that data governance be considered from the beginning to ensure an orderly data enterprise. Therefore, it is not something that can be done as an afterthought or as a late stage in your implementation.

In organizations where data governance is already in place, incorporating big data governance should be relatively straightforward. A complicating factor may include the assimilation of people or organizational units that are new to the concept of data governance but need to be brought to the table given they are providing (or responsible for) the new sources of data. Another potential stumbling block is that data derived from external sources, such as social media data, are unlikely to have an existing owner or steward, so new policies, procedures and guidelines will undoubtedly need to be created to address this.

8. Success criteria, value and impact

Last, but certainly not the least, you will need to consider how success will be defined. To track the on-going progress, you will need to develop and capture metrics that allow you to monitor and evaluate the current state in addition to capturing indicators that help project future requirements.

Circling back to an earlier point, at the beginning of your initiative, one of the critical questions will be what it means to the organization if the problem were solved with big data. This is crucial because it is really the lynchpin for determining whether a big data strategy is right for the organization.

Thinking about this question also helps provide a context for developing a budget and staffing plan. Without being able to demonstrate how your initiative will increase efficiency, drive better service or increase organizational capacity, it will be difficult to justify expenses related to getting big data initiative up and running.

9. Summary

While still considered a buzz word by many, the impact of big data and analytics is increasingly being felt in libraries and other information organizations. Starting big data initiative is a daunting task. In addition to the technical issues to be explored, there are a myriad of "softer" issues that have to be considered. By addressing the eight different factors outlined above and working through the issues methodically, an organization can successfully begin big data initiative and start reaping the benefits big data environment can provide.

References

CSC (2012), "Big data universe beginning to explode", available at: www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode

Deloitte (2014), *A Pragmatic Approach to Getting Started with Big Data*, Deloitte Insights (Oct 2, 2014), available at: http://deloitte.wsj.com/cio/2014/10/02/a-pragmatic-approach-to-getting-started-with-big-data/

Waxer, C. (2013), "Big data blues: the dangers of data mining", *ComputerWorld*, available at: www.computerworld.com/article/2485493/enterprise-applications-big-data-blues-the-dangers-of-data-mining.html

Corresponding author

H. Frank Cervone can be contacted at: fcervone@uic.edu