Paper 9740-2016

Big Data in Higher Education: From Concept to Results

Stephanie R. Thompson, Datamum

ABSTRACT

Higher education is taking on big data initiatives to help improve student outcomes and operational efficiencies. This paper shows how one institution went form White Paper to action. Challenges such as developing a common definition of big data, access to certain types of data, and bringing together institutional groups who had little previous contact are highlighted. The stepwise, collaborative approach taken is highlighted. How the data was eventually brought together and analyzed to generate actionable results is also covered. Learn how to make big data a part of your analytics toolbox.

INTRODUCTION

Higher education can benefit from the insights gained though big data analytics. However, information technology infrastructure may not be ready to handle petabytes of unstructured data. Making the best use of existing technology and skill sets to demonstrate the benefits of big data can pay off in the long run. What constitutes big in terms of data I have found to be very different in higher education than in many commercial sectors. Other than this difference, colleges may have data sources that they have not considered (or even knew about) that can be used to help students succeed or become more efficient. For example, what could room occupancy sensors or computer lab log-ins tell you about building utilization? Could this information have an impact on climate controls, snow removal schedules and patterns, dining service hours, or even campus safety?

This particular project started from the white paper drafted by one of the campus-wide working groups at a private university. The goal of the working groups was to investigate the potential for particular new initiatives to benefit the campus. After the white papers were reviewed by senior administrators, some were chosen to move forward. Big data was one that they felt was promising. This paper provides a general overview of the process taken by the team put together to make this happen.

A LITTLE BACKGROUND

In order for a big data project to have the best chance of success, different groups on campus who do not normally work together may need to. Institutional Research offices tend to have an ongoing relationship with Information Technology (IT) and student student-centered departments. Dining Services and Facilities Management, for example, may not have ever have had to interact with institutional research before. One thing is certain, if you start a big data project, you will get to work with a wide variety of departments across campus.

Sometimes it is better to start small with big data. This seems counterintuitive, but showing the promise of what can be accomplished without committing to hardware and software expenditures will help drive buy-in and take some of the uncertainty out of the project. Having the ability to grow the size of the data used will also build the skills of those managing and analyzing the data as this skill set may not already exist in house.

Lastly, once new data are stored and analyzed, questions will arise regarding data governance. Maybe you are now retaining data that has been purged regularly in the past. Data may also be personally identifiable and privacy protections are now at the forefront. Sometimes institutional research offices do not initially see these concerns as they typically work with until level data for faculty, staff, and students. When others are brought in to the project, this is seldom the case.

HOW BIG IS BIG?

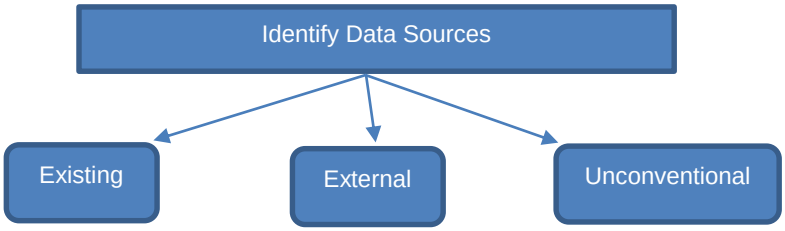
Data in higher education does not tend to have the volume of say retail companies. Colleges may have ten to twenty thousand students per year and even a ten-year longitudinal study does not yield high numbers of observations. However, when other data sources, both internal and external, are looked at data can get big quite fast. Even adding demographic and other student characteristics to a schedule

data set can make it several orders of magnitude larger. In one instance, data collection for a proof of concept had not yet even reached a terabyte and there were concerns as to the size of the data. This type of data volume can be new to IT departments and analysts. Many times the only infrastructure to house your data is a traditional database management system. This is not always the most desirable method and not necessarily geared to unstructured data.

Using existing systems as long as possible will allow you to build up some analytics to show a benefit to the organization. Once benefits are seen, expenses for better systems may be easier to justify. Growing with need as opposed to building for the possible makes sense. It could take time to reach brontobytes of data and having all of that structure in place for just a few terabytes doesn't make sense. Building capacity as you go is easier now than in the past. I recall working in a retail environment where there was such excitement about adding a terabyte of server storage. Now, I have a two terabyte external drive that cost around a hundred dollars compared to the thousands that terabyte cost just over 20 years ago.

POSSIBLE DATA SOURCES YOU DIDN'T KNOW YOU HAD

Once you start investigating data on campus, you may be surprised what is available. Sources like campus debit account purchases, wireless access point connections, and course management system use tracking can provide insight for big data projects. The diagram below highlights the different types of data sources to consider.



The existing data sources on campus tend to be student information, course management, work order, human resources, and dining systems. These tend to be the ones that are most commonly used for reporting and operations. Sometimes data from these sources are combined, but it is typically for a very specific purpose (e.g., determining the number of student employees working in various departments).

External data like weather might prove valuable and are publicly available. Granted, this is not data you already have in hand, but it and other types of public data are readily available. Looking at external and internal data together might provide insights you didn't think of. For example, do purchasing patterns change when it rains or drops below a certain temperature? Does rain impact parking utilization?

Look around for any type of electronic system. If it is powered, there is data there somewhere. This concept leads to the other hot topic if the Internet of Things (IoT). So many devices generate data that can be leveraged. This area is the "Unconventional" in the diagram above. These sources are where some of the interesting analyses can arise. Systems that maintain card access to building and rooms may prove beneficial. Even security camera systems may have features to allow you count people who pass by every hour to look at pedestrian patterns. Are your building controls automated? If so, you could review room light activation or occupancy sensors to see when facilities are used.

BRINGING DEPARTMENTS TOGETHER

The very nature of big data projects tends to bring many different departments together, even if only briefly. Information technology and Institutional Research are two of the most common. However, data stewards and customers can come from a variety of departments. Facilities Management, Dining Services, and Public Safety were some of the departments encountered during this project. These are not typical Institutional Research customers. However, they had needed data sources and opportunities that a big data project could help address.

Sometimes data are stored in atypical locations. Some sensors are controlled by an individual or small department. There may be concern regarding sharing the data or using it inappropriately. As data are institutional resources, they need to be treated as such. Making sure you work with the data owners or stewards will ensure confidence is developed. The analyst needs to understand the data to the fullest extent possible. This avoids misinterpretation and embarrassing errors. Also, once the data owner sees that you are interested in understanding the data they may be more comfortable sharing it.

WORKING WITH THE CUSTOMER

The biggest challenge is determining what can and cannot be answered by the data. Sometimes a customer wants to know something and there is just no data. Other times, while one question is being investigated, answers to other, unasked questions are found. That is all part of big data. Many things can be learned in the data evaluation phase of a project.

Spending the time up-front discussing a department's pain points can lead to a better understanding of their needs and generate some ideas for data sources and possible analyses. The more the analyst understands the "business" of the customer, the more likely the project will yield benefits.

It is also important to explain that a big data project is not like existing reporting endeavors or ad-hoc analyses. While a big data project may yield some type of production-based or live analytics result, it will be different than reporting from a data warehouse or online visualizations.

EXECUTING A PROOF OF CONCEPT

Starting small with big data can be to your advantage. Trying to do too much too soon could derail efforts to build a successful big data program. By selecting a single opportunity that needs insight and a select group of data, you can limit expense and demonstrate the benefits big data can bring to the university. Also, you will not be addressing an entire campus of data silos but just a few.

Make sure you plan for more time than you initially think will be necessary. You can run in to all types of issues that will slow you down: getting approval to access data, understanding data, narrowing down the wants of the customer, and technical issues of data storage and access.

A thorough review and exploration of the various data sources can take some time. For example, one data source in a recent proof of concept had over 700 tables. Trying to learn what to use and what not to use takes time – computing and talking with the subject matter experts.

A completed proof of concept can be the ticket to expanding big data efforts. If benefits from this type of data collection and analysis can be shown on a small scale, there is the possibility of expanding the benefits to other areas. Also, since some new data collection has been started, continuing it may just benefit other areas and projects.

A NOT SO BRIEF STATEMENT ON DATA GOVERNANCE

As data accumulates for a big data project, questions will arise regarding data governance and security. Institutional researchers may see this as somewhat odd as they tend to work regularly with unit record level data that contains personally identifiable information. This is not the general case. These types of projects have the possibility of tracking an individual in much greater detail than other analyses. Knowing an individual across data sources can be a huge advantage in developing models and presenting results. However, these results are more for groups of individuals as opposed to the individuals themselves. For example, knowing that members of a certain athletic team or student group always meet for lunch at a particular location when all of the members have gotten out of class can help to generate a demand model for that location. Since class times shift form term to term, you can reasonably know when to expect that group to show up. If you can do this for a large enough number of groups, you can assist dining services to be better prepared on the first day of classes. To make this happen, you need to identify the individuals but still protect their privacy. It is not necessary to say which groups are going where to the customer, just that they should expect X number of people to arrive at Y location at noon.

Many colleges have formal data governance committees and policies related to data and its use. Trying to fit a big data proof of concept into this mold can stunt the progress of the project. It may be better to work the small scale project and see where the existing rules apply and where policies meet or do not meet the requirements of a big data. This does not mean you should ignore the rules but rather see how things may need to be adapted for this type of project. This adaption may be on the project side or the policy side.

While data governance may not be an issue immediately, it is relevant and will become part of the process. This is particularly true as the big data team starts building a repository of data not typically retained or captured.

CONCLUSION

Big data efforts are becoming more common in commercial industries, but colleges and universities can also benefit from the effort. The data may not be a big as it is elsewhere, but there are many sources of data that can be leveraged. May are not what IT or Institutional Research offices are used to working with. This mutual learning curve is worth the work and can pay dividends with some perseverance.

Please note that additional details regarding big data project work will be part of the discussion during the presentation. However, due to issues of confidentially it is not part of the paper or presentation materials.

REFERENCES

Although the following books have not been directly referenced in this paper, they are nonetheless good sources for information on big data projects.

Lane, Jason E. (Editor). 2014. Building a Smarter University: Big Data, Innovation, and Analytics (SUNY series, Critical Issues in Higher Education). Albany, NY: State University of New York Press

Rijmenam, Mark van. 2014. Think Bigger: Developing a Successful Big Data Strategy for Your Business. New York, NY: AMACOM

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Stephanie R. Thompson
Datamum
901-326-0030
Stephanie@datamum.com
http://www.datamum.com