Big Data in Libraries

Robert Olendorf and Yan Wang

Abstract The term Big Data is somewhat loose. Roughly defined, it refers to any data that exceeds the users ability to analyze it in one of three dimensions (the three Vs): Volume, Velocity and Variety. Laney [1, 2] Each of these has different challenges. Huge volumes of data require the ability to store and retrieve the data efficiently. High velocity data requires the ability to ingest the data as it is created, essentially very fast internet connections. Highly variable data can be difficult to organize and process due to its unpredictability and unstructured nature. Bieraugel [3] Also, multiple data streams can be combined to answer a variety of question. All forms of big data can require high performance computing and specialized software to analyze. Given the fuzziness of defining big data,

1 Introduction

The term Big Data is somewhat loose. Roughly defined, it refers to any data that exceeds the users ability to analyze it in one of three dimensions (the three Vs): Volume, Velocity and Variety. Laney [1, 2] Each of these has different challenges. Huge volumes of data require the ability to store and retrieve the data efficiently. High velocity data requires the ability to ingest the data as it is created, essentially very fast internet connections. Highly variable data can be difficult to organize and process due to its unpredictability and unstructured nature. Bieraugel [3] Also, multiple data streams can be combined to answer a variety of question. All forms of big data can require high performance computing and specialized software to analyze. Given the fuzziness of defining big data,

Libraries can have to primary interests in big data. First is using big data to help with their operations. Libraries can use data to optimize their collections, better utilize space, asses their instruction and to provide information to their users. As more data is recorded in libraries, and more data is available externally via both

R. Olendorf
PennState University, Pennsylvania, PA, USA

Y. Wang (✉)
University of Alabama at Birmingham, Birmingham, AL, USA
e-mail: yanwang3@uab.edu

S.C. Suh, T. Anthony (eds.), *Big Data and Visual Analytics*,
https://doi.org/10.1007/978-3-319-63917-8_11

open and commercial sources, the opportunity for using these data streams is both a promising avenue for libraries, but also potentially risky in terms of allocating resources.

The second area in which libraries can work with big data is in the field of research data services. More and more funding agencies require data from funded research to be made public. Researchers often find themselves without sufficient skills and resources to adequately manage the data during their projects, and even more commonly do not have the skills, time or resources to prepare the data for archiving and often also are unable to find appropriate repositories for their data. Libraries have always played the role of housing societies academic and research output. While in previous generations, this has primarily focused on books and articles, expanding the scope to research data is a natural extension to the libraries role.

Here we explore these to realms of big data use in libraries. We explore the benefits, costs and risks associated with using big data. We also provide use cases, guidance to getting started and briefly outline tools and resources for working with big data. McCoy et al. [4] Given the speed with which the technology can change, we focus more on general guidance. Where specific technologies are referenced, it should be understood that technological progress can render some of the discussion obsolete, however, the general concepts should still be applicable.

2 Using Big Data In Libraries

Libraries are increasingly interested in using data to better manage their collections, utilize space and assess their services. Chen et al. [5] Traditional library services alone can generate huge amounts of data, especially for larger libraries. Some of the data are obvious, such as circulation data, patron counts and knowledge database usage. Other data streams might not be so obvious, such as harvesting social media feeds. It can be difficult for libraries to fully utilize the data create, and including additional data can present further difficulties.

However, using data to improve and optimize library function is a real benefit to be considered. Before diving into using data, it is important for the library to understand the full extent of the issue. Libraries should understand what data streams they wish to include. They should be aware of the costs in terms of both staffing and infrastructure will be required. They should also have a good idea of the benefits to be realized.

Big data projects can be of two types, a one-off project, or a standing service. The development of a either will generally take the following steps the biggest difference is a standing service will require a more durable production data store while a one off project can use less resource intensive data stores. One-off projects are ideal for answering questions that only need to be answered once. However, if the need is to repeat analyses, or have data ready for a variety of analyses or real time analyses, then a standing service should be preferred.

Either type of project will typically follow these steps should follow the order outlined. For instance, if benefits aren't identified early on, there is a risk of spending a great deal of time on resources that has no real need.

1. Identify benefits
2. Identify current resources
3. Identify solutions
4. Identify unmet needs
5. Implementation

2.1 Needs Analysis

Using Big Data can sound sexy and current. The temptation is often to assume that there are benefits because other people are doing it, it's in the news or because its the new technology. It is important to first understand the potential benefits and current needs. This will not only prevent unnecessary allocation of resources, but it will also result in a better product.

Ideally, a comprehensive list of current or potential services should be drawn up, with associated problems, difficulties these services face due to lack of sufficient information or analysis. In addition, ideas on how services can be improved with additional data should be considered as well. While not all of the listed items will directly involve big data, it is likely many will. Also, the list shouldn't, at this point, consider costs or even feasibility. The idea is to get a good list to work with.

Anytime a service can be improved or created through the application of data and analysis is an opportunity to use Big Data.

2.2 Current Resources

A comprehensive list of resources should be developed as well. Again, be inclusive here. Include any resource that could potentially be relevant. This will be useful both in estimating the cost of a solution and identifying solutions that fit well with resources you already have.

Computing resources should also be cataloged. Servers, storage and software should all be noted. Include cloud based resources that you currently use. List any database software (MySQL, PostGres), operating systems, statistical and analysis software you own or are licensed to use.

Staffing resources are critical. Big Data and data analysis are new to many libraries. However, there may be untapped skills within your library. Also, you should ensure that you list all of the Information Technology staff you currently employ. Stanton [6]

The first type of resource to consider is the data itself. A list of current data streams the library currently creates or has access to. This might include current holdings, circulation statistics, gate counts, patron counts, reference interactions and online chat. Also include external data sources which you may currently subscribe to as well as social media feeds.

A final critical resource to consider are external resources. Perhaps the most important would be the High Performance Computing (HPC) center and Cyber-Infrastructure (CI) center at your institution. Additional resources might be Statistical consultants and Risk Management.

2.3 Identify Solutions

With your needs in hand, you can start planning solutions. The solution will typically involve several components.

- Data store
- Administration
- Policies and access
- Analysis and visualization
- Query access (optional)

These may seem pretty standard and they are. However, when working with big data care must be taken to choose the right solutions. As noted previously the solutions you choose will depend on the nature of the project, one-off or service based. Ideally you would tackle your data store first.

2.4 Data Stores

For one-off solutions, the type of data store may be less critical. CSV or excel files may be sufficient for some projects, but they may not scale well with true big data projects, even one-offs, so it may be necessary to use SQL databases or document based databases such as Mongo DB or Hadoop. The cost of making a mistake isn't so great however, mainly in extra time spent fixing the mistake or time spent working with an inefficiency.

Service level data stores must be reliable and robust. Unfortunately, they also typically require maintenance and development. Also, because the the data store is intended for long term use, significant time in planning and development will be required. The benefit, of course, is the ability to make use of the data for a variety of purposes continually over time. Service level big data stores, often utilize technologies such a data warehouse architecture or document based or NoSQL databases Hecht and Jablonski [7].

Data Warehouses take data from a variety of sources, transform it and then load it into a central data store Inmon [8]. The focus with a Data Warehouse is to transform or pre-process the data so that it is more efficient for the types of queries you will perform. For instance, data is often timestamped. Libraries might be interested in how a variety of resources are used on an hourly basis over the course of a week. With just a time stamp this can be an expensive query, especially if you are merging different data streams, where you must first determine the day of the week and hour of the day for each record before joining the records together. Using Data Warehouse type architecture, the data is preprocessed, for instance assigning each record a day of the week and an hour based on its timestamp.

Document based repositories (NoSQL) forgo the structure of data base schema and are often said to scale better than SQL. Additionally, products such as Kibana for Elasticsearch are designed for big data analysis and visualization.

The choice will depend on a variety of factors. Before deciding it is important to determine your needs and also the technical skills and support your institution can provide for a given solution.

2.5 Administration

For service level data stores you will need to provide an administrative layer. Some solutions come with built in administrative front ends. Ideally, when researching your data store, you should also determine what administrative tools are available. Not having to build your own can save a great deal of time and resources.

2.6 Policies and Access

One important non-technical issue to resolve are policies of use and access. At the start, a document that states the scope of the project, and policies governing use should be drafted. The scope specifies the data that are expected to be included in the data store and what they will be used for. Defining the scope helps to focus development and limit" scope creep" which happens as new ideas come to the light. Stating the scope also helps better set stakeholder expectations. Policies concerning use will help define the users, and also provide avenues for interested parties to gain access to the data.

One very important aspect to consider is how to handle sensitive data. This includes data with personally identifying information, trade secrets or other data that could be potentially damaging to other agents.

When mashing different data sets together, additional care must be taken that the combined data doesn't cause additional risks. For instance, data that identifies a student's school may not be risky. Data about student's ethnicity or economic status

may not be risky. Combining the two however, may suddenly make it much easier to identify individuals with a fair degree of certainty.

2.7 Analysis and Visualization

The types of analyses and visualizations that are needed may drive the solutions chosen. While most solutions provide a wide array of analyses and solutions some are limited in what they can do.

Also, the skills of the potential users should be taken account. Analysis and Visualizations can be divided into two types of interfaces, point and click, and scripted. Point and click interfaces are easier to use, especially for users with little or no background in data analysis. The trade-off is that these are usually more limited in what they will do, and perhaps more importantly they often make it difficult document how the analyses and visualizations were created. Also, the ease of entry can allow new users to use the analyses and visualizations wrongly.

Scripted solutions essentially require the user to write short programs to accomplish the analyses and visualizations. They require considerable more skill to use. However, they typically allow the user to a perform a much wider range of analyses and visualizations. Because the analyses are written as a script, it is easy to know how they are done and additional documentation can usually be added as needed. Because the user will typically already be fairly skilled in analysis and visualization, the risk of inappropriate analyses is reduced. Using scripted solutions should be considered a best practice, but may not be practical for all institutions.

Analysis solutions should also be evaluated on how the results can be exported. Some solutions only allow tables to be exported in CSV or even proprietary formats. Graphs and charts may only be exportable in PNG or other image formats. If one of the goals is to make data available on the web and to make it interactive, solutions should be chose that provide the ability to export content as interactive graphs usually in SVG.

2.8 Query Access

Query access is providing refers to providing a low level interface to the data, usually in the form of an HTTP based API. This feature is not necessary, but many data warehouses provide. Implementing this requires some thought, including access control.

Query access can also be used to allow the extraction of data to be used in online visualizations. For instance, a good API can generate json outputs that can then be input into a product such as Highcharts that generates very nice and interactive charts and graphs.

2.9 Implementation

One-off research projects are often practical, and several examples of research projects conducted from within libraries exist [citations needed]. Typically, these projects do not require extensive technical infrastructure, and if the required expertise does not exist within the library, collaborations with other departments or libraries will typically suffice.

Service level projects, will exceed the capabilities of many libraries to support them with the necessary infrastructure and expertise. However, collaborations with other institutional departments such as high performance computing, advanced cyber-infrastructure, or similar units may be advantageous.

3 Library Big Data Services

Libraries are in a unique position to not only use big data for their own purposes, but also to provide services around big data. Many libraries are already offering research data services such as data management planning, data collection, data curation and data archiving. Sugimoto et al. [9] Some libraries also offer some level of analytic and visualization support, much of this driven by demand created by various mandates to provide data management plans and make data more accessible. At the same time, researchers are constantly pushing the limits of their data sets, increasing both their size and the complexity. While in some realms, this has been going on for some time, other domains are just starting to explore the possibilities of big data. Libraries can play an important role in facilitating their institutions research, although as before, this will often require cooperating with other units.

We will explore some of the possible services libraries can provide using the data life cycle developed by DataONE 1. While libraries can play an important role, libraries will not have great a role in each stage of the life cycle. Some stages may present too great of challenges or require too many researchers, so libraries will need to choose carefully what services to provide and how to resource them.

3.1 Planning

Assisting with Data Management Plans (DMPs) is one of the most common services offered by academic libraries. Working with a researcher on a big data project only differs from typical DMPs in several ways. Depending on the funder, a typical DMP covers most of the topics Most of the details will be covered below. Librarians who assist in writing DMPs involving big data projects will need to be aware of all the services the researcher might require to adequately address the issues in a DMP. Also, they should be aware that much of the boiler plate text that many libraries

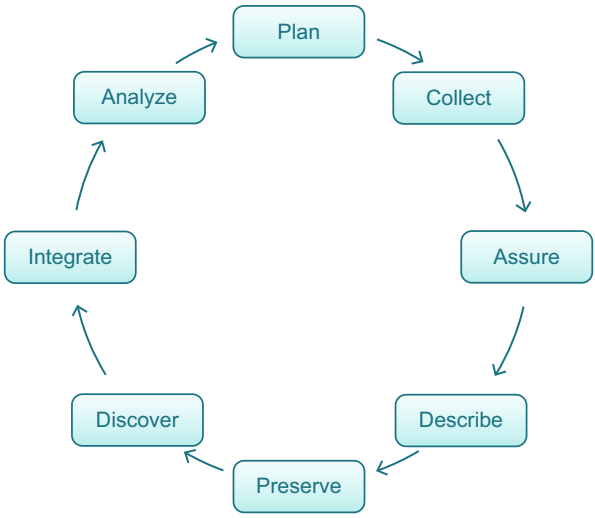

Fig. 1 The data life cycle developed by DataONE

provide may not be applicable. For instance, many libraries provide text describing how their institutional repository can accept data, however, big data projects may be too large or complex for these repositories (Fig. 1).

3.2 Collection

Collection of big data may seem to be outside the range of typical library services. However, there are some opportunities to be found. Some libraries may already have in house skills common big data sources, such as social media feeds. Libraries some commercially available big data sources might be acquired and treated as part of the library collection.

Acquiring and working with big data can be computer intensive. Libraries can collaborate with other units such as high performance computing to provide computing assistance and also the technical infrastructure to work with big data. Libraries primary role here might be to ensure that the data collected is properly organized and documented to ensure that it can be easily archived later.

3.3 Assure, Analyze and Integrate

Quality insurance (QA), analysis and integration of the data is typically the responsibility of the researcher. However, if librarians are supplying data as part of their collection, there may be some responsibility to perform quality assurance on this data and provide adequate documentation for researchers to conduct their

own QA methods. Also, in general, libraries should provide a supporting role by providing tools and resources to aid researchers in conducting and documenting their QA protocols. This might include version control, tutorials, and workshops.

3.4 Describe

Providing data with adequate metadata and documentation fall squarely within the strengths of libraries and this is a task researchers struggle with providing proper documentation and metadata as well. Libraries should provide services to help researchers document their big data and supply it with metadata. Here we differentiate between documentation and metadata. Documentation is human readable information that provides details about various aspects of the data. Metadata is typically machine readable information, often machine generated and in addition to providing some information that aids in using the data, also functions to identify and describe the data.

Ideally, librarians would work proactively with researchers so that they begin documenting their work as it is created. Otherwise, especially with very large complex projects, much information is lost or forgotten over the course of the project. Also, if researchers wait until the end of the project to document their work, it can become overwhelming and often does not get done. Two fairly simple steps to take at the start of the project are to create a README text file, and to create an initial file structure and naming convention. Big data projects can be large and complex and often involve multiple collaborators. By starting the project in an organized fashion with solid documentation, the project will typically proceed far more smoothly than otherwise.

README files are simple text files frequently used in software development to describe the project and document how to use the software. The contents can vary quite a bit, but the overall goal of the README is to provide enough information that other researchers can replicate, validate and reuse the data with confidence. The README should be updated throughout the project and acts as both a means to communicate with the current collaborators and also as a method for providing needed information to others who may use the data in the future. Some common components include

- A list of creators with roles contact information.
- A list of contributors.
- A description of the project, similar to an article abstract.
- A description of the data that includes.
- A description of the contents.
- A description of the naming conventions.
- How to view the data, what software is needed.
- How to run the analysis to replicate the data.
- License information
- An example citation

The project's file structure and naming conventions should also be set at the beginning of the project. A simple, high level directory structure should be created with naming conventions that help to describe the contents and purpose of the directory, such as "data", "analyses", "output". Likewise, naming conventions for other directories and their files should be consistent and provide information about what they are.

3.5 Preserve

Along with data description, archiving and preservation is perhaps the other primary service libraries can provide. Many funders, including most STEM agencies such as NSF and NIH require that researchers make their data available upon completion of the project. NIH [10] For big data projects this can be especially problematic for the researchers due to the quantity and complexity of the data. It can also present unique problems to the libraries.

Many libraries have institutional repositories and some also have dedicated data repositories. However, in most cases they are limited by their technology to smaller and simpler datasets. This presents a problem to researchers who are using big data and whose data sets may not easily fit into the institutional repository solutions. An additional difficulty occurs when data is derived from commercial sources such as financial data or social media feeds. In these cases, researchers are limited by what they can make public, while still often required to release enough information to validate their results.

To better handle big data sets in repositories there are several possible paths. First, it may often be possible to reduce the size of the data by reducing repetitive content, or removing content that can be generated again. For instance, Twitter data can be reduced to a list of just the Tweet IDs vastly reducing the amount of storage needed for the data. This is also another opportunity to collaborate with other institutional units such as who have more resources in the form of storage and computing power. Often these units lack the skills in curation to make the data findable and reusable which libraries often have. Finally, there may be other external repositories that are more appropriate for the data. This is common in the genomics realm for instance.

3.6 Discover

Aside from some of the more common big data sources like Twitter, discovering useful data can be difficult. Librarians involved in research data support can serve a useful role in working with researchers to find relevant and useful datasets. These can range from support of genomics research to locating corpora for digital humanities projects. In addition to locating the datasets, librarians as informaticists should also be familiar with techniques for efficiently acquiring the data, storing and processing the data.

Another issue with data discovery is negotiating intellectual property issues. While much data in the STEM disciplines may be open, there is considerable data that comes with a variety of licensing arrangements. Librarians can be helpful to researchers in understanding these, but importantly, also to determine how to best make the resulting data as open as possible so that the research can be disseminated in as open a manner as possible.

4 Conclusion

Although it has not always been recognized as such, data has always been fundamental to academic research. While we typically think of data as numbers and measurements, it also includes text, images and anything else that is used to inform our understanding of the natural world, and the artifacts we create. As the digital age moves on, we are increasingly seeing data exposed as part of the process and the understanding that access to the data increases our ability to understand what has been done and also to advance faster by reusing data that has already been created.

In the last 15 years or so, we have seen an increase in the velocity, volume and variety of data created to the point where our ability to comprehend, let alone work with the data stretches our limits. Librarians have always been custodians of our shared store of knowledge. As the digital and data age move forward, librarians must adapt to not only handling information in the form of written text but also working with information in all its variety and volume. This will require partnerships with other experts such as high performance computing, cyber-infrastructure and also the researchers themselves.

References

1. Laney, D.: META delta. Application delivery strategies, 949:4. (2001)
2. McCoy, C., Marcinkowski, M., Sawyer, S., Sanfilippo, M.R., Meyer, E.T., Rosenbaum, H.: Social informatics of data norms. Proc. Assoc. Inform. Sci. Technol. **53**(1), 1–4 (2016)
3. Bieraugel, M.: Keeping Up with Big Data. (2013)
4. Rosenbaum, H.: Social informatics of data norms. Proc. Assoc. Inform. Sci. Technol. **53**(1), 1–4 (2016)
5. Chen, H.-I., Doty, P., Mollman, C., Niu, X., Yu, J.-C., Zhang, T.: Library assessment and data analytics in the big data era: practice and policies. Proc. Assoc. Inform. Sci. Technol. **52**(1), 1–4 (2015)
6. Stanton, J.M.: Data science: what's in It for the new librarian? https://ischool.syr.edu/infospace/2012/07/16/data-science-whats-in-it-for-the-new-librarian/ (2012). Accessed 12 Jan 2017
7. Hecht, R., Jablonski, S.: NoSQL evaluation: a use case oriented survey. In: Proceedings—2011 International Conference on Cloud and Service Computing, CSC 2011, pp. 336–341 (2011)
8. Inmon, W.: Building the Data Warehouse. Wiley, Hoboken, NJ (2005)

9. Sugimoto, C.R., Ding, Y., Thelwall, M.: Library and Information Science in the Big Data Era: Funding, Projects, and Future [A Panel Proposal]. (2012)
10. National Institutes of Health: Nih Data Sharing Policy. https://grants.nih.gov/grants/policy/data_sharing/ (2006). Accessed 14 Jan 2017