Big Data – How to Manage Large Volumes of Data:

Focus on Issues in Higher Education Research

Sharon O'Boyle

George Mason University

IT 214, Section DL1

December 1, 2013

Abstract

Managing Big Data is an issue facing all organizations. Many technologies and strategies are being developed to deal with Big Data. This paper will give an overview of Big Data, and then discuss Big Data issues as they relate to higher education. Examples of current technologies and strategies will be described. In addition, advantages and disadvantages of these technologies and strategies will be given.

Introduction

There are many definitions of "Big Data." But, in general, Big Data means very large data sets that cannot be stored or analyzed in traditional ways. Professionals in all industries are tasked with the challenge of finding innovative ways to store and utilize these massive amounts of data. This paper will describe how institutions of higher learning are developing methods to store and analyze "Big Data" required for their research.

History and Background Information

Although data in some form have been in existence forever, today's data is being generated, stored and analyzed like never before. A recent report by the Bureau of Labor Statistics claims that "According to some experts, 90 percent of the data that exists in the world today was created in the last 2 years" (Royster, 2013). Furthermore, this data is not always the traditional numbers and text data of the past (i.e. structured data). Today's data is much more complex – social networking feeds, photos, videos, etc. (i.e. unstructured data). Traditional methods of storing data are not sufficient to handle the new data needs. For example, the volume of data can require hundreds of servers. In addition to storage issues, the data must be easily accessible at all times.

Applications in Today's Society

Research requires large amounts of data. Colleges and universities confront Big Data issues in the many research studies that they undertake. They are using various technologies and strategies to store and process their massive amounts of data.

One example that illustrates the data storage strategy of a research university is the human genome study at the University of California, Berkeley College of Engineering. This research requires hundreds of terabytes, and perhaps petabytes of data. The AMPLab at UC

Berkeley has sufficient disk storage available. But access to the disk storage can cause problems because it can be slow. They have implemented two technologies to improve their processing. One technology introduced at the AMPLab is solid-state drives (SSDs). These drives have high-speed flash memory instead of the traditional spinning disks. This allows for greater throughput. The other technology being used is in-memory storage. They are using a "data cluster" organization comprised of 30 Intel servers. Ten of these servers include almost 1TB of SSD storage each, and all 30 servers include 256GB of memory (Grimes, 2013).

The University of North Carolina at Chapel Hill Renaissance Computing Institute (RENCI) also performs genomics research and is developing methods for handling Big Data. RENCI is using a Quantum StorNext G302 File System to move some of their storage from spinning disks to a Quantum Scaler i1600 tape library. This will improve efficiency of data access (Grimes, 2013).

In Holyoke, Massachusetts a new life sciences computing facility was recently built. Dedicated 10-gigabit-per-second fiber-optic lines link the Massachusetts Institute of Technology, the University of Massachusetts, Harvard University, Northeastern University and Boston University to this new computing facility. This allows for local, efficient storage and also for collaboration among the universities (Moreira, 2013).

Our own George Mason University has implemented Hadoop and installed R software to help handle the Big Data that the faculty and administrators are storing and processing (Hughes, 2013).

Universities are also using "the cloud" for storage of Big Data. According to a 2013 study by CDW Government, data storage is the top application in higher education, and 31 percent of campuses are using the cloud to store data (Nagel, 2013).

Advantages

The various Big Data storage solutions that are being implemented on university campuses each have their advantages and disadvantages. The advantages of the different solutions will be described in this section.

The advantage of solid-state drives (SSDs) such as those at UC Berkeley is that they allow for greater throughput than traditional drives. An advantage of the Quantum StorNext G302 File System used by University of North Carolina at Chapel Hill is that it greatly improves efficiency of data access. The advantage of the new computing center in Massachusetts is that, for the five associated universities, the data can be stored locally. For them, it is a less-expensive option. It is also ideal for collaborative research among the universities.

In some cases using the cloud for storage of Big Data can reduce costs. The CDW Government survey states that organizations implementing cloud services are saving an average of 13 percent annually. The same survey also reported increased efficiency (Nagel, 2013). Cloud storage can increase flexibility as space can be added or reduced "on-demand". For example, Amazon's E2C product allows customers to order virtual servers as needed and pay by the hour. Expectations are higher accessibility, availability, efficiency and improved recovery services (Nicholson, 2009).

Disadvantages

Just as the various Big Data storage solutions have advantages, they also have some disadvantages. The disadvantages of the different solutions will be described in this section.

The disadvantage of solid-state drives (SSDs) is that are more expensive than traditional disk drives. A disadvantage or the large computing center, such as the one at Holyoke, is that,

for maximum efficiency, the users of the data must be located relatively close to the center. So this is an option only in certain localities.

Cloud storage has its own set of potential disadvantages. Cloud storage can be costly and slow. The size of the Big Data files can overwhelm the networks. Other potential disadvantages of cloud storage are privacy and security concerns, performance, and ambiguity in service level agreements. Storing sensitive data in remote storage centers, potentially anywhere in the world, presents legal, compliance and political issues. In the cloud, institutions give up some degree of access control and also transparency (seeing who has accessed the data). And of course, any failure along the cloud network can cause significant access problems.

A disadvantage of implementing any Big Data solution is that in universities it is often difficult to gain the support of Chief Information Officers (CIOs) for investment in new technologies to manage Big Data. Time and effort is required to gain support and funding for Big Data projects. But according to Joy Hughes, CIO at George Mason University for 16 years, Big Data is "a transformative technology environment that is needed in higher education as in the corporate world." (Hughes, 2013).

Conclusion and Future Outlook

Big Data will continue to grow bigger, at even faster rates than we are seeing now. Existing technologies will be improved and new technologies are always being developed. Universities will need to form cross-functional teams to analyze the complex individual needs of their institutions. A "one size fits all" approach will never be appropriate. The many factors involved (cost, efficiency, speed, reliability, privacy, security, compliance, etc.) must be carefully evaluated so that a successful Big Data storage strategy can be developed and implemented.

References

Grimes, B. (2013, May 9). A Big Data storage tip: One size does not fit all. Retrieved November 13, 2013, from http://www.edtechmagazine.com/higher/article/2013/05/big-data-storage-tip-one-size-does-not-fit-all

This is an article that describes Big Data storage strategies at the University of California, Berkeley College of Engineering and the University of North Carolina at Chapel Hill Renaissance Computing Institute.

Hughes, J. (2013, June 26). Education CIOs: Get involved in Big Data. Retrieved November 13, 2013, from http://www.informationweek.com/government/leadership/education-cios-get-involved-in-big-data/d/d-id/1110516?

This is an article that describes the difficulties of persuading Higher Education CIOs to develop an effective strategy for managing Big Data.

Moreira, N. (2013, June 17). Big-data crunching hits the fast lane in Holyoke. Retrieved November 13, 2013, from http://www.bostonglobe.com/business/2013/06/16/datacenter/l4wkiDu1bZPSWuUjr6ceIN/story.html

This is an article that describes a new storage center for Big Data in Holyoke, Massachusetts.

Nagel, D. (2013, February 19). Storage, conferencing drive campuses to the cloud. Retrieved November 12, 2013, from http://campustechnology.com/articles/2013/02/19/storage-conferencing-drive-campuses-to-the-cloud.aspx

This is an article that discusses the use of cloud computing in higher education.

Nicholson, J. (2009, June). Cloud computing's top issues for higher education. Retrieved November 12, 2013, from http://www.universitybusiness.com/article/cloud-computings-top-issues-higher-education

This is an article that discusses the issues related to the use of cloud computing in higher education.

Royster, S (2013). *Working with Big Data.* [White paper]. Washington, DC: Bureau of Labor Statistics.

This is a white paper published by the Bureau of Labor Statistics that gives a general description of Big Data. Royster is an economist in the Office of Occupational Statistics and Employment Projections.