# Big Data

## Laying the Groundwork

ECAR

ECAR Working Group Paper | September 17, 2014

This paper is a publication of the ECAR Campus Cyberinfrastructure (ECAR-CCI) Working Group. ECAR working groups bring together higher education IT leaders to address core technology challenges. Individuals at EDUCAUSE member institutions are invited to collaborate on projects that advance emerging technologies important to colleges and universities. ECAR-CCI helps educational institutions develop institutional strategies, plan resource deployment, and help users harness and optimize the power and capabilities of these new integrated IT tools and systems for educational and research applications in higher education.

# Introduction

The topic of big data continues to receive a great deal of publicity because of its promise for opening new avenues of scholarly discovery and commercial opportunity. The ability to sift rapidly through massive amounts of data, for example, is resulting in new kinds of scientific discoveries and is making information about Internet-browsing habits more accessible and usable to the commercial sector. Understanding the issues associated with this topic is particularly important on our campuses, where the Internet plays a vital role in managing and providing access to big data sets for research and in helping generate big enterprise data sets from the day-to-day business of the university. Questions arise such as "How do we take precautions to preserve the security and privacy of systems like admissions, e-mail, and course management systems without detracting from our ability to do research?"

This paper introduces and sets the groundwork for a series of briefs on the topic of big data as it relates to the academic campus. The series has the following aims:

- Establish the role that big data plays in campus cyberinfrastructure (CI).
- Point out issues of concern around big data and campus CI, support, and development.
- Present findings and recommendations to CIOs, vice presidents for infrastructure, and IT computing and networking staff; we expect that other parts of the campus leadership, including vice presidents for research, library leadership, and perhaps risk management will also find the materials of use.

# Defining Big Data

There are many definitions for "big data." Depending on the context (e.g., business, research, campus enterprise, or sports), the meanings of "large" and "complex" vary. Other definitions go beyond size (sometimes called volume) to include rate of change of the data (sometimes called velocity) and the complexity of the data and metadata (sometimes called variety). Often there's an assumption that big



EDUCAUSE

data represents a collection of related data sets from multiple sources that need to be integrated, correlated, or otherwise analyzed together. Perhaps the one consistency across these definitions is simply that big data is data that is "big" and doesn't readily lend itself to conventional IT practices or casual treatment.

Our growing ability to create, move, and analyze big data sets is responsible, among other things, for

- Improved modeling and more accurate forecasting in the fields of storm analysis and weather prediction.
- Data mining on a massive scale involved in the discovery of the Higgs boson in the field of particle physics.
- Verification of the cosmic inflation hypothesis.[1]

> Big Data: A loosely defined term used to describe data sets so large and complex that they become difficult to process and store using data management tools.
>
> EDUCAUSE ACTI Data Management Working Group, Data Management Glossary

# Big Data and Campus CI

When considering big data for research and the impact on campus CI of storing, moving, and curating massive amounts of data, features that distinguish big data sets emerge. As important as issues of storage, movement, and curation are, we must keep them in perspective as we go about the day-to-day business of the campus. Some questions that arise when we do that include:

- How do we enable new architectural models for big data computation and transport without undermining existing enterprise architectures?
- How do we protect enterprise data without impeding traffic for research data?
- How do we manage big research data efficiently without adversely impacting the rest of the campus CI?

We plan to address these questions in context, focusing on the specific issues that arise when working with big data to understand what their implications may be on the campus infrastructure as a whole.

The first three topics that will be addressed in the series are infrastructure support, security, and curation.

## Basic Infrastructure Support

In the next paper in the series, we intend to address basic campus infrastructure support, including network design, server hosting strategies, backup and disaster recovery implications, and campus bridging. Research computing infrastructure support services have often been specialized and at times isolated. The size and scale of research big data by default will impact campus IT. A well-coordinated strategy is essential.

> Based on years of experience moving large data sets between research universities and national labs, the ScienceDMZ separates the wide-area end-to-end performance needs of moving big data from the more general needs of the campus LAN.

Similarly, network architecture must be reconsidered because conventional campus network architectures that seek to minimize security threats and manage costs are not adequate to meet the performance needs for efficiently transporting research big data. In a follow-up paper, we will explore emerging approaches to addressing these issues.

## Security and Privacy

Whether data contain personally identifiable information, are confidential, or are restricted for other reasons, big data presents challenges that are not present with smaller data sets. For example, many data sets have restricted data use requirements, and conventional ways of dealing with these security and privacy risks often fail. Another complication arises when big data sets are distributed globally or when an institution has big data sets that it wants to share with collaborators who are distributed globally. It is unrealistic to replicate these (often dynamic) data sets, so access management needs to include authentication and authorization methods, as well as data-transfer mechanisms that are secure and mutually trusted.

> Big Data creates new security and privacy challenges that de-identification can't meet.
>
> —David Geer, CSO Online

## Curation

Data curation is the process of ensuring that data can be understood and reused by interested parties across disciplines, organizations, and the passage of time; it also implies making choices about where to invest limited resources and understanding likely needs for the data, ranging from experimental reproducibility to genuine repurposing. It subsumes preservation but goes beyond it. Stakeholders in the process include scholars, librarians, IT staff, funders, and policy makers. There has been considerable progress in developing standards and best practices within and across disciplines and types of data, but substantial challenges remain. New policies being put in place by major funding agencies have also added a significant compliance dimension to research data management. Data curation challenges are not unique to big data, but the particular characteristics of large data resources imply both a sizable investment of resources and some particular technical complexities and approaches that we will emphasize in our discussion of the topic.

# Conclusion

We wrestle with the complexities and logistics of big research data, not as an end in itself but because mastering these complexities is a means to new scientific discovery. We aim to strike a balance between the demands of enterprise computing and the needs of research computing. Future papers in this series will address that balance.

## Authors and Contributors

*Special thanks go to the following ECAR Campus Cyberinfrastructure Working Group authors of and contributors to this report.*


**Guy T. Almes, ECAR-CCI Co-Chair**
Director, Academy for Advanced Telecommunications and Learning Technologies
Texas A&M University

**Curtis W. Hillegas, ECAR-CCI Co-Chair**
Director of Research Computing
Princeton University

**Timothy Lance**
President and Board Chair
NYSERNet, Inc.

**Clifford A. Lynch**
Executive Director
Coalition for Networked Information

**Gregory E. Monaco**
Director for Research and Cyberinfrastructure Initiatives/Great Plains Network
Kansas State University

**Michael R. Mundrane**
Chief Information Officer
University of Connecticut

**Ralph J. Zottola**
CTO, Research Computing
University of Massachusetts Central Office


## Note

1. Harvard-Smithsonian Center for Astrophysics, "First Direct Evidence of Cosmic Inflation," Release No.: 2014–05, March 17, 2014.