Big Data Infrastructure 101

Timothy M. Chester
Vice President for Information Technology
University of Georgia
University of Georgia

- First state-chartered institution of higher education, oldest public institution of higher education in the United States

- Organized into 17 schools and colleges, 140 degree programs offered across multiple campuses

- 89th among U.S. institutions in research expenditures
Conflicting Objectives

• Traditional approach to innovation (find smart people, resource them, and get out of their way) leads to redundancies, inefficiencies, and unnecessary duplication

• Symptoms of Infrastructure Problem
  • Server/Storage Proliferation
  • Non-professional IT Staff
  • Poor IT Practices (Data Loss)
  • Power / Facilities / Other Hidden Costs not visible until crisis

• Signs of Data Organization Problems
  • Many versions of the truth
  • Lack of standards for data definition
  • Lack of integration between production, distribution, and analysis of data
Infrastructure + standards + shared governance + relationship building all linked together
Current Trends

• Academic, administrative, and research-oriented computing are increasingly linked together

• Recognition that data is most valuable resource across all three areas

• Decision-making further decentralized as power of social networks is brought to bear on key problems

• Current systems (people and IT) do not scale well to support emerging needs
Current Trends

• Big Data Problems
  • Admitting Students
  • Student Placement
  • Student Performance
  • Student Retention
  • Degree Completion
  • Research Innovation
  • Increasing External Support

• IT Organization Linkages
  • Infrastructure
  • Administrative Applications
  • Academic Computing
  • Research Computing
  • Information Security
Current Trends

- Campus Networks
  - Size of data continues to grow
  - Data must move to where compute power resides
  - Sharing of data across domains / centers inhibited by information security barriers
UGA Internet Use (GB)

- 2/2011 Eliminated 1 GB Bottleneck
- 9/2011 1.36 GB
- 9/2013 - 2.78 GB
- 9/2014 - 3.9 GB
- 9/2015 - 5.1 GB
Registered Wireless Devices

- 9/2011 - 6,000
- 9/2012 - 14,000
- 9/2013 - 22,000
- 9/2014 - 29,500
- 9/2015 - 53,296
Things to Think About

- Never a better time to move to fully managed networks, requires
  - $$$ Incentives
  - Strong, Reliable Service
  - Patience
- Partitioning of Networks into Science / Research DMZ’s has promise
  - Reduce Information Security related bottlenecks
- Direct upgrade Capex $$$ to data intensive areas, if speed is acceptable don’t upgrade it
Current Trends

• Compute / Servers
  • Data and Compute go hand-in-hand; moving data to compute and not the other way around
  • Traditional backup approaches do not scale well, are too $$$ for research data
  • Cloud is not a one-size fits all solution
    • Continuous research workloads are not cost effective
    • Cloud storage often requires local caches, negating scale advantage of cloud storage
Things to Think About

• Cloud-hosting for administrative applications is mature, no reason not to take advantage of hosted services for these applications

• Cloud compute resources for research only work for episodic, burst-able loads; otherwise local compute + storage resources required

• Leverage policy and regulatory frameworks to reduce scope of what is expected and required - particularly for backups and archives

• Leverage service partners higher up the food chain: labs -> colleges; colleges -> central IT; central IT -> external systems and resources
Centralized vs Decentralized IT

Figure 1. Scholarly Infrastructure

Wheeler, Brad. 2008. “E-Research is a Fad: Scholarship 2.0, Cyberinfrastructure, and IT Governance”
Standards & Accountability

• Overarching frameworks for controls and standards often inhibit innovation at best, often the cause of difficult relationships in large enterprises at worst

• Array of tools and controls, apply them in situations where they work best
  • Sensitive Information
  • Authoritative Data
  • Network Management
  • Compute & Storage
  • Tools & Applications

• Executive Support -> Must have to keep decision-making, responsibility, and accountability aligned, otherwise CIO / IT gets squeezed
Shared Governance
Shared Governance

• Must “be a fit” for the culture and history of the institution, and line-up with philosophy of current executive team

• Any successful IT Leadership strategy will incorporate
  • Information Sharing Forums
  • Advisory Bodies
  • Formal, Project Specific Governance Bodies
  • Linkages and reporting lines to Senior Leadership of the Institution

• Shared governance is not a solution for central IT services that are not reliable, responsive, or cost-effective
Relationship Building

• The “Secret Sauce” that makes things work across large, decentralized enterprises

• Investing in competencies oriented around people, relationship building, and negotiation skills is more important than technical skills

• Special emphasis on:
  • Culture of accountability to end user community; manifested in consistent, responsive, and reliable IT services
  • Regular communication to key executives on a monthly basis about key initiatives and activities
  • Regular face-to-face meetings with key executives and their local IT support; particularly with distributed locations
Wrapping Up

- Infrastructure + standards + shared governance + relationship building are all linked together

- 2/3 of challenges with providing infrastructure for “Big Data” initiatives are people related, not technology related

- No one-size fits all solution, approaches must fit well with institutional context, resource levels, strategic objectives, and build on current baselines

- Progress comes in small incremental steps, not in big, sweeping solutions
Acknowledgements

• Diana Oblinger
• Pierce Cantrell
• Brad Wheeler
• Greg Jackson
• Richard Katz
• Jim Davis
• Phil Goldstein
• Guy Almes
• Curtis W. Hillegas
• IT Staff @ University of Georgia