Institution: University of Liverpool

Unit of Assessment: 36 – Communication, Cultural and Media Studies, Library and Information Management

Title of case study: The application of embedded analytics to hyper-scale and distributed data archives

1. Summary of the impact (indicative maximum 100 words)

The research improves digital data archives by embedding computation into the storage controllers that maintain the integrity of the data within the archive. This opens up a number of possibilities:

- Data analysis can be automated and incorporated into the archiving process;
- The approach improves the archiving of all types of digital objects, from television broadcasts to genomes;
- The approach can be applied to distributed data and to datasets that are too big for traditional approaches.

This has impact on three different classes of beneficiary:

- Providers of national data infrastructure in the UK and US, who are incorporating Cheshire 3 into national data repositories;
- Data Users, such as Astra Zeneca, RAI, Sanger Institute, who are using Cheshire 3 to extract valuable information from their data;
- Equipment vendors, such as NetApp, Xerox and Bellerophon Mobile, who are developing commercial systems that will use Cheshire 3.

2. Underpinning research (indicative maximum 500 words)

The chief innovation at Liverpool University is the development of "Cheshire 3", a data analysis and processing system in which the analysis and processing of the data are integral to the data archive. Workflows to analyse data and discover information are implemented in a virtual machine that operates directly on the data archive.

This approach, of embedding the computer that does the data processing in the data archive, has a big advantage. The traditional approach, of moving the data to be processed from an archive to the computer that does the processing, becomes impossible when very large amounts of data, petabytes (millions of gigabytes) or exabytes (billions of gigabytes), are involved. Storing such large quantities of data requires a great deal of hardware. Moving the data from archive to computer consumes a great deal of bandwidth.

The research began in 1997 as a collaboration between Paul Watry (employed continuously at the University of Liverpool) and Ray Larson of the University of California at Berkeley. The collaboration, which continues to the present, was funded jointly by JISC and the National Science Foundation and produced Cheshire 2, a prototype that worked with text, numerical and geospatial data based on an earlier text-discovery system developed by Larson.

The next phase, which was supported by a grant from JISC in 2004 to Watry, developed the prototype of Cheshire 3, which has the added capability of being able to deal with distributed data (see reference 6), and is published as US Patent 20060277170, issued in December 2006.

Subsequent work developed capabilities for "big data" archive and analysis and developed applications to support particular needs and projects by integrating Cheshire 3 with different data management technologies. The work was funded by a variety of collaborative grants including the SHAMAN Framework 7 project, JISC funding under the "Digging into Data" programme and the PERICLES Framework 7 project which is now being prepared for industrial production.

The Cheshire 3 system is used internationally on a production basis and is integrated with policy-

based data management systems such as the integrated Rule-Oriented Data System (iRODS). As discussed in section 4, it now forms a significant part of the national information infrastructure in the UK and the US.

3. References to the research (indicative maximum of six references)

1. Software publication: Cheshire3 Digital Library System.

2. Sanderson, R., Watry, P. "Integrating data and text mining processes for digital library applications". ACM/IEEE Joint Conferences on Digital Libraries, JCDL2007, Vancouver, BC Canada. ISBN: 978-1-59593-644-8. DOI: 10.1145/1255175.1255188.

3. Watry, P, Larson, R., Sanderson, R. "Knowledge generation from digital libraries and persistent archives", Research and Advanced Technology for Digital Libraries, 10th European Conference ECDL 2006. Research and Advanced Technology for Digital Libraries Lectures Notes in Computer Science Volume 4172 (2006). DOI: 10.1007/11863878_54. ISBN 978-3-540-44638-5.

4. Watry, P., Sanderson, R. Larson, R. US Patent: "Digital Library System", Patent number 20060277170, issued 7 December 2006. The published disclosure relates to a digital library system that will operate in both single processor and grid distributed computing requirements.

5. Watry, P. "Digital preservation theory and application: Transcontinental persistent archives testbed activity", International Journal of Digital Curation, vol. 2, No. 2 (2007), pp. 41-68. ISSN: 1746-8256. DOI: 10.2218.ijdc.v212.28.

6. Watry, P., Larson, R. "Cheshire 3 framework white paper: implementing support for digital repositories in a data grid environment", Local to Global Data Interoperability – Challenges and Technologies (2005), 60-64. ISBN 0-7803-9228-0. DOI: 10.1109/LGDI.2005.1612466.

Collaborative research grants:

- The Sustaining Heritage Access through Multivalent ArchiviNg project (SHAMAN), European Commission (Framework Programme 7 project), 2007-2011, £1M
- Promoting and Enhancing the Reuse of Information throughout the Content Lifecycle exploiting Evolving Semantics (PERICLES), European Commission (Framework Programme 7 project), 2013-2017, £978k
- Integrating data mining and data management technologies for scholarly enquiry, JISC funding under the "Digging into Data" programme, 2012-2013, £81k

4. Details of the impact (indicative maximum 750 words)

Overview

Cheshire 3 has had high impact because its development has been guided by rich 2-way interactions with different networks of potential beneficiaries. The beneficiaries include the bodies that provide national data infrastructures, data users, and equipment vendors. Our interactions with them increase the reach and the significance of the impact in 3 ways:

- The interactions have ensured that Cheshire 3 is useful. Its development has been guided by the needs of potential users. It supports the latest policies, methods and practices for the analysis and sharing of data and its technologies are closely integrated with the latest data management technologies.
- The interactions ensure that public and private bodies that provide data infrastructures are aware of and understand the advantages of Cheshire 3.
- The interactions ensure that data users know that Cheshire 3 provides the kinds of enhanced capabilities that they need in order to archive their data reliably and to extract value from it.

Interactions leading to Impact

Our interactions with **providers of national data infrastructure** include long-term collaboration with teams internationally responding to needs assessment and requirements analysis; integration with other projects already involved with "big data" management and analytics areas; and long-term involvement with standards working groups and implementation of the software at STFC for the Virtual Engineering Centre at Daresbury. In 2009 the software was demonstrated to 13 Federal Agencies as a primary technology, which resulted in the formation of the DataNet Federation Consortium (DFC). This demonstration also formed the basis of the SHAMAN Integrated Project prototype (2008-2012), which served to integrate the Cheshire system with the iRODS adaptive middleware, thus supporting long-term curation and analysis of archived data.

Our interactions with **data users** include participation in groups working to develop social consensus for the analysis and sharing of data, policies, methods, and practice; and the interoperability mechanisms to support technology integration and data analysis. This has involved sustained interaction with emerging networks of expertise across digital curation projects.

Our interactions with **equipment vendors** include discussions with NetApp about the possible adoption of the software for storage devices, working with corporate developers in London; employment of software developers across SME (Archive Analytics, SpaceApps) to provide services and support throughout the EU; use of the software by value added resellers.

1) Impact on National Data Infrastructures

The software was developed at Liverpool and partly funded by the NARA US National Archives as a means of supporting long-term curation and analysis of digital data.

a. The software has supported services across multiple agencies, in the US, including NSF, NARA, NASA, NIH, DOD, NHPRC, IMLS, and Europe, including DOE, JISC, EU, EPSRC.
b. The software forms a foundational component of the US National Science Foundation data infrastructure based at DataNet Federation Consortium, for national (US) e-Science research applications.
c. The software contributed to all of these services the ability to index material within a data grid (cloud) and the ability to provide data analytics workflows across active collections, across all these organizations. Cheshire3 has provided this functionality to the DFC and other NSF research projects that support the national data cyberinfrastructure for NSF research projects in the United States and which rely heavily on the extensive set of metadata that allows for the discovery, analysis, and preservation of studies. The expanded ability of these initiatives to federate using iRODS and Cheshire discovery services means that scientific research datasets can be analysed and managed with policy based rules that protect the authenticity, privacy, provenance, context, and integrity of datasets.
d. The software is widely used on a service-oriented basis internationally, and forms the infrastructure of multiple national digital library services in the UK, including the Archives Hub (JISC) (1999-present), and the Incunabula Short Title Catalogue (British Library) (2005-present).

2) Impact on Data Users

a. The integrated software forms the basis of commercial prototypes for AstraZeneca, Drexel, and Virtual Engineering Centre, for managing data driven collections; and in 2012 the Virtual Engineering Centre commissioned a prototype for use in the automotive and aerospace sectors at the Science and Technology Facilities Council.
b. The integrated system is currently being rolled out for research organizations, such as Sanger Institute, Science and Technology Facilities Council (STFC), University of Edinburgh (EPCC), University of Liverpool (N8 grid).
c. The system has been prototyped for use in managing research data for the pharmaceutical industry (AstraZeneca) (2008-2010).
d. The research advances in distributed data analytics applied to petascale (millions of gigabytes) collections have resulted in measurable impact improving the analysis and management of scientific research data, including research collaborations across different communities of practice.

3) Impact on Equipment Developers

a. NetApp, DataDirectNetworks, and Xerox are promoting use of the system and software for cloud appliance and managing “big data technologies” for research or predictive use cases. The Business Development Manager for Cloud Services at NetApp UK Ltd will confirm that the Cheshire technology developed by Professor Watry's group will be embedded in upcoming NetApp enterprise products on a commercial basis from 2014.
b. The software is driving investment within the SME communities, both in the US and EU. The Founder and Senior Science Advisor of Bellerophon Mobile in the United States, for example, states that “the use of Cheshire3 workflows to aid discovery of distributed data could contribute greatly to effectiveness of building mobile knowledge-based systems”.

5. Sources to corroborate the impact (indicative maximum of 10 references)

1. The Business Development Manager for Cloud Services at NetApp UK Ltd can be contacted to confirm the impact of Cheshire3 on Equipment Developers (3a, 3b). He will confirm that NetApp are releasing commercial products incorporating the technology.

2. Head of Search, Taxonomy, and Enteprise Content Management at Astra Zeneca Pharmaceuticals can be contacted to confirm the impact of Cheshire3 on data users and its adoption by Astra Zeneca for managing pharmaceutical data (2a, 2c).

3. The Director of the Sustainable Archive Institute at the School of Information and Library Science, University of North Carolina, can be contacted to confirm the impact of Cheshire3 on the National data infrastructure in the US and the UK and its impact on data users (1a-d; 2a, 2b, 2d).

4. The Co-founder and Science Advisor of Bellerophon Mobile can be contacted to confirm the impact of Cheshire3 on data users (2a, 2b, 2d) and SME investment (3b).

5. The Enterprise Architecture Area Manager of the Services Innovation Laboratory, Xerox Research Centre Europe, has provided a statement to confirm that Cheshire3 produces measurable impact on the analysis and management of scientific data and that it is driving commercial investment (2d, 3b).