Shazam! La Trobe University Automated Research Data Pipeline

Ghulam Murtaza[1,2], Michael Filippidis[2], Connie Darmanin[3]

[1] Intersect Australia Ltd, Sydney, Australia, ghulam.murtaza@intersect.org.au
[2] La Trobe University, Bundoora, Australia, m.fillippidis@latrobe.edu.au
[3] Australian Research Council (ARC) Centre of Excellence in Advanced Molecular Imaging, Department of Chemistry and Physics, La Trobe Institute for Molecular Sciences, La Trobe University, Melbourne, Australia, c.darmanin@latrobe.edu.au

PROBLEM STATEMENT

La Trobe University (LTU) researchers from both the SHE[1] and ASSC[2] colleges across a number of research disciplines are using internal and external instruments to capture large quantities of research data as part of their ongoing research. The movement of the captured data from the instruments to analysis environments was a manual, ad-hoc, error prone and time consuming process for the researchers. Whilst ICT[3] provides research network drives to researchers, these do not scale well for the larger datasets being generated by the instruments in question, resulting in researchers using their own data storage solutions leading to data security and retention issues, and difficulties for the researcher to share the data with their collaborators, particularly with those external to the university. In addition, the researchers have separate and distinct analysis environments, which are difficult for ICT to support.

SOLUTION

La Trobe University got together with Intersect Australia to develop a cloud-based data capture and storage solution that addresses these issues and increases researcher efficiency. Intersect, in consultation with LTU, developed Shazam as a key automated research data pipeline solution. Shazam is an automated research data capture design pattern along with a visualisation and analytical toolkit created by Intersect to implement a robust, fault-tolerant data management, storage and processing solution for research projects generating huge (terabytes) amounts of research data. Shazam client components runs within LTU infrastructure, while server and analytical hub components are based within the Intersect Space/Time cloud. Figure 1 shows the solution architecture of Shazam for La Trobe university.

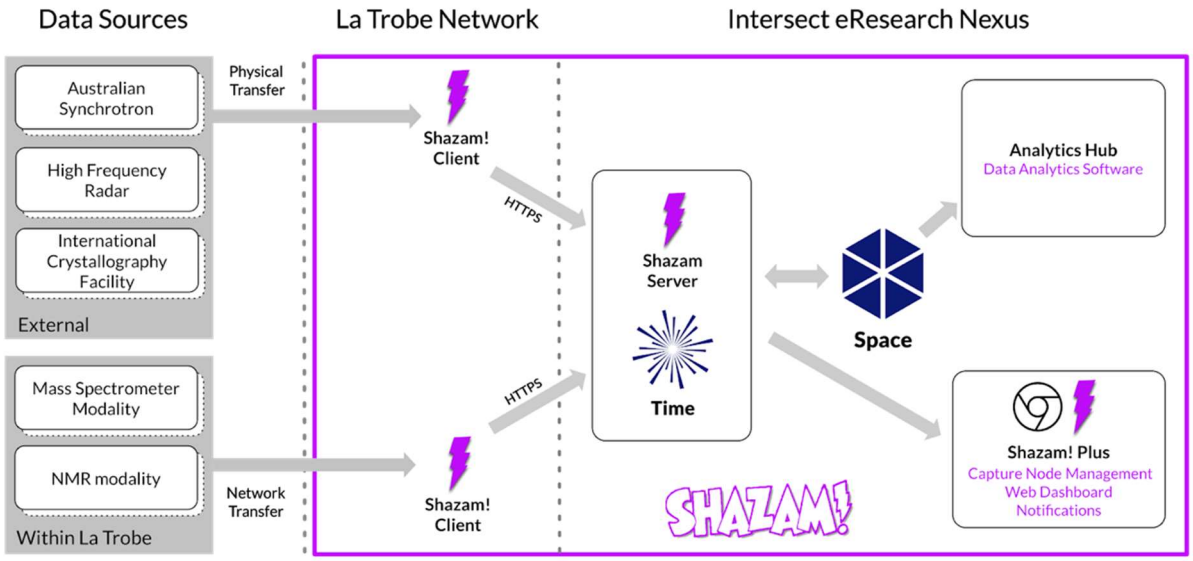

Figure 1 Shazam deployment for La Trobe University

[1] Science, Health and Engineering College
[2] Arts, Social Science and Commerce College
[3] Information and Communications Technology Division

INITIAL DEPLOYMENT

The first Shazam client was installed within the La Trobe Institute for Molecular Science (LIMS). Over Christmas, it began acquiring over 15TB of molecular data in the form of thousands of files from many external USB disks consisting of historical data from the Australian and international Synchrotrons, plus various X-ray Free Electron Laser (XFEL) sources; every last byte of data was captured, safely transmitted via the Internet, validated for assurance, and stored. This data is now held securely in Intersect Space, and available as needed to an analytics system hosted in the Intersect Time platform. Because the data is directly related to observations, batches and experiments – not just filesystems – and because Shazam automates so many tedious manual processes and points where errors could be introduced, we call this “data teleporting”.

CONCLUSION

In this presentation, we would like to present the solution design of the Shazam architecture and demonstrate the approach deployed by La Trobe University. We will also explain how the system is helping researchers at La Trobe university to efficiently teleport their data without a hands-on approach. We will also present a brief overview of the roadmap for Shazam that would outline how LTU envisage Shazam to contribute to the broader Research Systems landscape within the university.