Cloud-Based Self Service Analytics

Andrew G Naish*
Chief Technical Officer, Space-Time Research, Melbourne, Australia
andrew.naish@spacetimeresearch.com

Abstracts

Traditionally, the means by which official statistics are analyzed and disseminated by both commercial and private industries consisted of a large capital outlay for data lifecycle management and infrastructure. Additionally the business processes by which official statistics are extracted and disseminated are inefficient, costly and require specific expertise.

With the introduction of a cloud based self-service model, official statistics providers can now maintain data sovereignty and security while disseminating to a wider audience with much more efficiency, and at a lower cost.

Providing an easy to use, self-service interface can not only alleviate the burden of such a lengthy process but can also provide new opportunities to popularize statistics and promote evidence based decision making. Given standard ontologies, statistical based sub systems now have the capability to join on common dimensions between agencies and produce new insights, offer trend reporting and predictive analysis.

Space-Time Research provides the means for common people to explore, build, visualize and share official statistics online.

Key Words: Self-Service Analytics, Cloud, Official Statistics, Data Sovereignty.

1. Introduction

The online self-service dissemination model can be described as an evolution of the traditional means of dissemination of statistics commonly involving manual aggregation, perturbation and de-identification steps before releasing a 'static', predefined data view (usually in the form of a table, chart, map or aggregated data cube) of specific official statistical data topics.

In 2006 the Australian Bureau of Statistics (ABS) led the world in the online dissemination of statistics based on a self-service model. With the release of ABS TableBuilder (formerly CDATA Online) in 2008, powered by Space-Time Research's SuperWEB2 platform, we saw the very first instance of government dissemination of self-service statistics extracted directly from unit record microdata[1].

The online self-service model introduced by Space-Time Research and the Australian Bureau of Statistics allowed anyone to create ad-hoc queries on unit record micro data without the risk of disclosing individual unit record details. The three main components to such a system were:

a) An enterprise data warehouse with fact tables and dimension tables containing individual unit records.
b) A data aggregation server that will aggregate and apply disclosure control methods to unit records based on predefined classification definitions.
c) A web based portal to allow online access for creating cross tabulations and visualizations based on the predefined classification definitions.

Figure 1.0 depicts the current self-service model used at the Australian Bureau of Statistics:

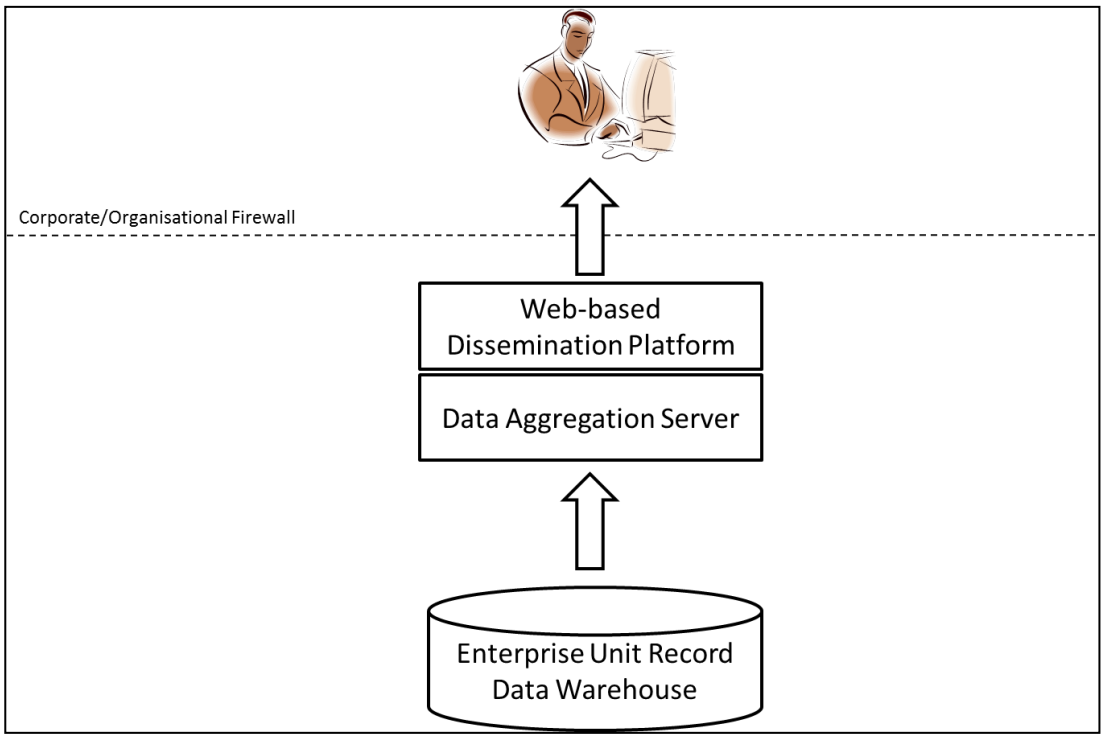

Figure 1.0 – Traditional Self-Service statistical data dissemination model

2. Evolution of a new model for cloud-based self-service analytics

In 2012 Space-Time Research embarked on a research and development project to re-architect the technology involved in self-service analytics to provide for the following capabilities:

a) Allow organisations that produce official statistics to maintain data sovereignty without restricting interagency data merging or third-party application development.

b) Allow end users to upload and merge proprietary data with official statistics without granting organisational firewall access.

c) Provide a cloud based platform to enable community based sharing and commentary on official statistics in order to further promote evidence-based decision making.

In order to support these capabilities a component-based architecture was developed to split the data aggregation server and disclosure control methods with a cloud-based query and dissemination platform. Figure 1.1 depicts the basic architecture for the new model:

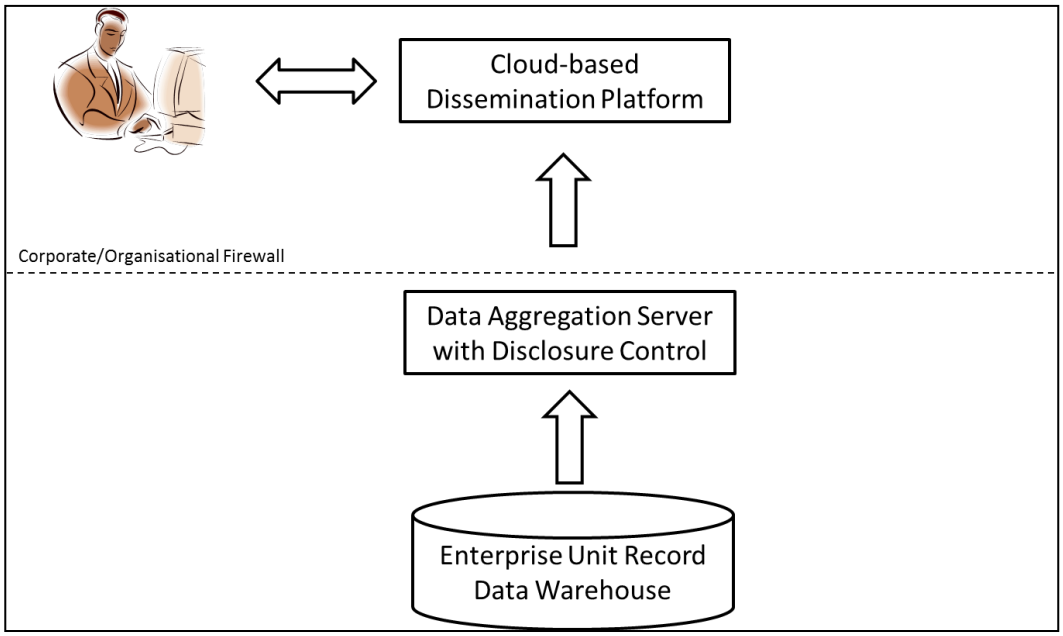

Figure 2.0 – Component based architecture to split between confidentilised aggregation server and web dissemination.

Advantages of the split between a confidentialised aggregation server and the cloud-based based dissemination platform are numerous[2] including, but not limited to,

- Infrastructure cost savings,
- Work flexibility
- Time to market (for reports and datasets) and,
- Social networking capabilities

One subtle advantage is the capability to derive new insights from existing datasets sourced from individual disparate official statistics providers. By joining datasets with common, standard dimensions (the most obvious examples being time and/or geography) a galaxy schema can be derived enabling 'on the fly' interagency dataset querying.

Figure 2.1 depicts the component based architecture involved in allowing inter-agency data querying.

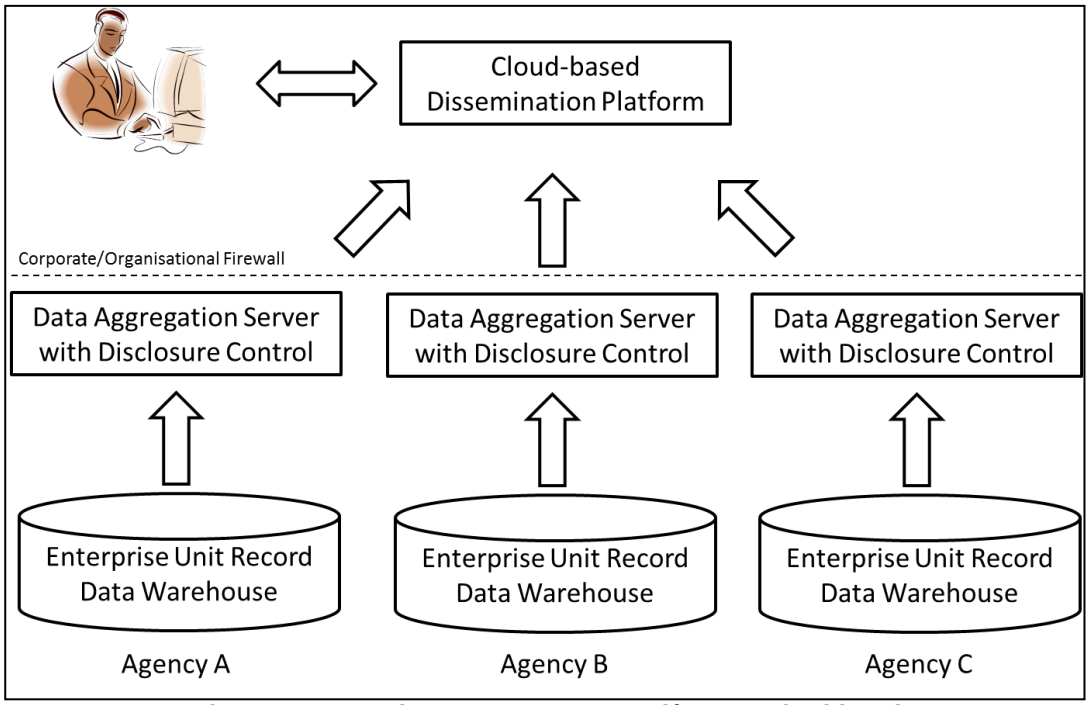

Figure 2.1 – Seamless interagency data querying using a self-service cloud-based component model.

3. Results

Implementation of a cloud based self-service analytics system executing queries on unit record micro data was completed and successfully launched online at http://www.superdatahub.com on May 15th 2013. Anyone can register to the system at http://www.superdatahub.com/register.html

The system was developed using modern, popular languages including a JavaScript web based client, Scala/Java web components, RDF DB stores and various proprietary tabulation components. Protocols between components primarily consist of a RESTful http calls transmitting JSON representations of database queries and results.

The system has been utilised by a number of users from varying backgrounds, including students from Australian universities such as Monash University (VIC), The University of Melbourne (VIC) and Griffith University (QLD) along with major enterprises such as Westpac Bank, ANZ Bank and Optus and also international organisations such as the OECD[3].

5. Conclusions

Fundamentally, cloud based services are increasingly becoming more relevant as the cost savings involved in 'renting' architecture (as opposed to a large capital infrastructure investment) become more apparent.

Overcoming the increasing value and organizational awareness of data sovereignty (especially with government) has been one of the most difficult challenges, however the technology and architecture developed as part of SuperDataHub has been demonstrated to completely solve this problem through the separation of tabulation and disclosure control with cloud-based visualisation dissemination.

Furthermore, with the ever increasing capability of social networks to deliver data views at almost real-time capacity, we can be somewhat optimistic that better, evidence based decisions will be made from more accurate data by more people. Additionally, with higher velocity data, better computing techniques and semantic networks this type of model could lead the way to automated correlation of datasets providing users with information opportunities we have never seen (or thought of) before.

References

1 - "Making Census Data Available with CDATA Online", Department of Innovation, May 25, 2010.
http://showcase.govspace.gov.au/item/making-census-data-available-with-cdata-online/

2 - "Benefits of cloud computing", Queensland Government, April 16, 2013.
http://www.business.qld.gov.au/business/running/technology-for-business/cloud-computing-business/cloud-computing-benefits

3 - "Super Data Hub – Alpha Trials", Space–Time Research, May, 2013. Email: andrew.naish@spacetimeresearch.com