International Journal of Population Data Science

Journal Website: www.ijpds.org

Public Cloud: The Future of Record Linkage?

Brown, A¹, Randall, S¹, Ferrante, A¹, and Boyd, J¹

¹Curtin University

Introduction

Businesses worldwide are increasingly adopting the storage, compute and analytical services provided by cloud computing. Yet, few operational linkage units are keeping pace with this world of technological change - most use legacy systems approaching their limits with the rapidly increasing size and range of datasets now required for linkage.

Objectives and Approach

To meet the demands of linkage for the near future, it is important that new solutions for linkage consider the services provided by public cloud infrastructure for compute, storage and analytics. We examined Platform as a Service (PaaS) offerings for use in the development of a cost-effective cloud model for scalable, privacy-preserving record linkage (PPRL). PPRL techniques were adapted to maximise the quality of linkage and to automate as much of the process as possible. Finally, a prototype was created to demonstrate the capabilities and potential of the model.

Results

We present our cloud model for PPRL, a platform for record linkage that provides rapid scaling of resources to meet demand, and the results of how our prototype performed on massive datasets.

Conclusion/Implications

The application of record linkage using relatively inexpensive cloud infrastructure represents a significant step towards providing an efficient and scalable record linkage service to researchers and government. Larger datasets can be linked efficiently, including national or cross-jurisdictional datasets, with little investment in private infrastructure, and improved turnaround times for researchers.