Chapter 18
A Dual-Database Trusted Broker System for Resolving, Protecting, and Utilizing Multi-Sourced Data

Neal Gibson
Arkansas Research Center, USA

Greg Holland
Arkansas Research Center, USA

ABSTRACT

A longitudinal database structure, which allows for the joining of data between disparate systems and government agencies, is outlined. While this approach is specific to government agencies, many of the ideas implemented are from the commercial world and have relevance to problems associated with data integration in all domains. The goal of the system is to allow for the sharing of data between agencies while upholding the strictest interpretations of rules and regulations protecting individual privacy and confidentiality. The ability to link records over time is central to such a system, so a knowledge-based approach to entity resolution is outlined along with how this system that integrates longitudinal data from multiple sources can still protect individual privacy and confidentiality. Central to this protection is that personally identifiable information should not be proliferated on multiple systems. The system, TrustEd, is a hybrid model that provides the simplicity of a centralized model with the privacy protection of a federated model.

DOI: 10.4018/978-1-4666-4892-0.ch018

INTRODUCTION

The majority of state governments are for the first time creating interagency data systems that link data from disparate agencies together. The purpose of these systems is to measure the outcomes of publically funded programs, for example, how college graduates are faring in the workforce. The creation of such systems raises a number of concerns about privacy and confidentiality, and there are numerous regulations regarding how each agency's data must be handled. Such integrated systems also represent a number of technical hurdles as well, especially in terms of matching records to individuals over a long time span.

The architecture of one such system, TrustEd, is described in detail. Its design addresses both privacy and technical issues. While this solution is specific to a single state, the issues involved and solutions developed have significance for any organization that must integrate data from multiple disparate systems. A great deal of the research for this system revolved around identity management. The two main issues are how to get the best possible match rates and how to manage changes once a bad match is discovered. At the center of the design is the protection of individual privacy and confidentiality, achieved through a dual-database design which keeps personally identifiable information (PII) and the data for research in separate systems.

BACKGROUND

There has been a push in recent years for government agencies to share data, especially at the federal level. The impetus for this was 9/11. Subsequent hearings showed that a number of agencies were aware of and even actively monitoring some of the hijackers involved, but that the full scope of the danger remained somewhat hidden by the fact that data were not being shared between agencies (National Commission on Terrorist Attacks Upon the United States [9/11 Commission], 2004, p. 79). The response was to create an "Information Sharing Environment," to facilitate the exchange and integration of data on matters dealing with national security (http://www.ise.gov/). The federal government has also encouraged state agencies to share data, by awarding funds through competitive grants for states to create longitudinal data systems through which such programs as Early Learning, K12 Education, Higher Education, and Workforce can be integrated into what are known as state P20W systems. The U.S. Department of Education and the U.S. Department of Workforce have provided significant grant monies to encourage states to integrate data from early childhood, elementary and secondary education, higher education, and workforce.

Two general models are being used by states to create these longitudinal systems that can track individuals from early learning programs, on to school and college, and their outcomes in the workforce (Garcia & L'Orange, 2012, p. 13). A centralized model is the simplest, where data is brought together in a central data warehouse. For the purposes of privacy, PII can be stripped from the data and a unique identifier used to link data together. A federated model allows agencies to retain the data within their institutions, and a system is put in place to allow for the querying of data across disparate systems.

Issues and Problems: Privacy

Initiatives which link data from multiple agencies give rise to questions regarding privacy. While the term "privacy" is often used as a general term to cover a host of topics, it is important to differentiate between privacy and confidentiality. The difference between the two, as Kenneth Prewitt, former Director of the U.S. Census Bureau, states is between "*don't ask*" and "*don't tell*" (Prewitt, 2011, p. 42). Data breaches, such as the theft of a laptop from the Bureau of Veterans affairs which contained the names, dates of birth, and Social