Schematic Structure of National Data Harmonization System for Identity Management

Osuolale, A. Festus
Adewale, O. Sunday
Department of Computer Science
The Federal University of Technology, Akure, Nigeria
Abimbola, O. Jeremiah
Department of Computer Science, Babcock University,
Ilishan-Remo, Ogun State

doi: 10.19044/esj.2016.v13n3p318 URL:http://dx.doi.org/10.19044/esj.2016.v13n3p318

Abstract

Imagine a web-application in which a user can query varieties of information about an individual (like name, age, state of origin, driver's license number, national identification number, etc.). This information would be stored on different databases, each having its own schema. Often times, these resources are replicated in different locations and on different platforms. Hence, the need for data harmonization or integration. Data harmonization addresses this issue by considering these external resources as materialized views over a virtual mediated schema resulting in "virtual data integration". When mediated schema is queried, the solution transforms the result into appropriate queries over the various and existing data sources. This paper focuses on data harmonization that cuts across different governmental database and incorporating them all into a centralized view and this is possible as a result of development and implementation of a web-based databank application.

Keywords: Database, Databank, schema, harmonization/integration, queries

Introduction

From time immemorial, organizations and government agencies have had to store information of employees, criminals, and operational statistics etc. which are usually in databases located at different sites. Most often, it is such that each department in the organization has their own separate database with its schema containing only information peculiar to the department. By implication, this means a single organization or agency can have more than one database populated and scattered within or without their

operational territory and this can cause redundancy. This phenomenon also applies to databases of different governmental organizations. Data-integration (or data harmonisation) will address this issue by considering these external resources as materialized views over a virtual mediated schema resulting in "virtual data harmonisation" or virtual data integration". As the term implies, data harmonisation actually involves concatenating data from several and different sources, which are stored using various technologies and provide a unified view for the data (Data Intgration Info, 2015). Generally, materialized views are significant in data effective database query because they enable much more efficient access, at the cost of data being potentially out-of-date (Flexview, 2011). In light of this, there will be a virtual and mediated schema that best models the kinds of desired response. Other aspects are the adapters for each data source which will be designed such as census database and crime database. These adapters simply transform the local query results into a processed form for the data-integration solution. When a user queries the mediated schema, the solution transforms this query appropriately over the respective data sources but there is a need for an effective identification management system.

The solution provides a uniform access to a set of autonomous but often heterogeneous data sources in a particular domain. This is typically what is needed when for instance, querying the *deep web* that is composed of plethora of databases accessible through web forms (Abiteboul, et al 2011). We would always want to find relevant data no matter which database provides it even with a single query.

Basically, two approaches to achieve this exist: the mediator and the warehousing approach. For the purpose of this project, we take the *mediator* approach because of its huge advantage of accessing "fresh" information. In this approach, data remain exclusively in data sources and are obtained on request. Advanced Data virtualization is also built on the concept of object-oriented modelling in order to construct virtual mediated schema or virtual metadata repository, using hub and spoke architecture. This is in contrast to a *warehousing* approach where data is extracted from the data source ahead of query time, transformed, and loaded in the warehouse. Unlike the traditional extract, transform, load ("ETL") process and data federation, data remains in place, and real-time access is given to the source system for the record, thus reducing the risk of data errors and reducing the workload of moving data around that may never be used. (Morgan, 2013) A useful way out is to reorganized disparate databases to harmonize the databases without considering the use of ETL. The recast databases support and provide designed data access paths with data value commonality across databases

The biometric recognition of identities using biological and behavioral means has been presented as a natural identity management tool

that offers greater security and convenience than traditional methods of personal recognition. Indeed, many existing government identity management system employ biometrics to ensure that each person has a unique identity in the system. Each update (deletion or insertion) of data in the various databases will also include the biometrics. Identity Management (IM) is establishing identity of a single person using one or more of the biometric or non-biometric features. Biometric trait is a biological and behavioral characteristics of an individual, such as fingerprint, face, gait (i.e. the way the person walks) and signature. Non-biometric feature is anything other than biometric such as pin number, password and name. When an application user queries the mediated schema, the solution transforms this query into appropriate queries over the respective data sources.

The specific objectives of this project are to: develop a national data harmonization system that captures twelve major different organizations in Nigeria as follows: National Population Commission (NPC), Federal Road Safety Commission (FRSC), Nigeria Police Force (NPF), National Youth Service Corp (NYSC), Independent National Electoral Commission (INEC), Federal Inland Revenue Service (FIRS), Federal Service Commission (ServiCom), Nigeria Medical Association (NMA), Federal Ministry of Health, National Identity Management Commission (NIMC), Central Bank Of Nigeria (CBN), Nigerian Communications Commission (NCC) and to design a web application for its implementation.

Related works

The Protein Data bank (PDB) is the single worldwide archive of structural data of biological macromolecules. It is a repository for the three dimensional structural data of large biological molecules such as proteins and nucleic acids. The data stored are received through x-rays crystallography or NMR spectroscopy and submitted by biologist and biochemists from around the world. The submitted data are made available on the internet via websites of member organizations. The PDB is driven by the worldwide Protein Databank (wwPDB). Data may be submitted via email or via the AutoDep Input Tool (ADIT) developed by the RCSB. All the data collected form depositors by the PDB are considered primary data. Thereafter, the captured data are assessed for quality i.e. how well these models fit the experimental data. The PDB validates structures using community standards as parts of ADIT's integrated data processing system. (Wikipedia, 2014). Zhonghua Yu et al, 2008 developed a poison databank in order to establish a comprehensive, easily approached, operated, method to search the internet on professional poison data and knowledge of effective treatment for those consented such as medical staff, and emergency response team in the shortest time. A computer poison databank was established, by

adopting B/S structure, using SQL Server databank, and explores technology, in which all information may easily be explored and obtained by users (Yu, YI, & Chi, 2008). A bioinformatics databank was created in 2014, information from research areas including genomics, protemics, metabolomics, microarray gene expression and phylogenetics. Information contained in the biological databank includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures. The project was done because it was difficult to ensure the consistency of information. This was solved by creating a biological databases cross-reference to other databases with accession numbers to link their related knowledge together. (Biological Databases, 2014) Williams J.S Elliott, 2011 published a report on the development of a DNA databank in canada for the purpose of forensic DNA analysis which aided the job of the security agents in the country during investigation but the kind of information it captures (DNA) posses security and privacy issues (Elliott, 2011). C.G Berbatis in 2003 developed a pharmacy databank funded under the Third Community Pharmacy Agreement Research and Development Grants Program. It was deveopled because of the increasing rate of criminology in the pharmacy department of the state. The public was empowered with the right to ascertain the legitimacy of the license issued to the pharmacist who attended to them. (CG Berbatis, 2003)**.** Pascal and Nyamulinda reported on a program held in Rwanda on National Identification. At the end of the project, different platforms were inter-connected like Ministry of Labor and Public Service, National Police, Immigration Office, Central Bank, Rwanda Revenue Authority, Rwanda databank, Land Center, National Electoral Commission, and Telecommunications like MTN, TIGO, and Airtel (PASCAL & NYAMULINDA, 2014). England embarked on their first DNA databank in 1995 and was reviewed in 2006. This was developed for forensic purpose. Following a series of legislative changes, DNA samples can be taken by the police from anyone arrested and detained in police custody in connection with a recordable offence. These are offenses that that have to be recorded on the Police National Computer to form part of a person's criminal record. (Technology, 2006)

Research methodology

The integration of these repositories into a single databank will be implemented using the concept of resources as "*materialized view*" over a "*virtual mediated schema*", resulting in "*virtual data integration*". This means we would construct a virtual schema. The virtual schema will consist of the unique way of identification, the authentication system all incorporated as a "mediated schema" to best model the kinds of answers the

users want. Next, we design "*wrappers*" or adapters for each data source such as National Identity Management Commission (NIMC) database and Independent National Electoral Commission (INEC). These adapters simply transform the local query from the individual databases into an easily processed form for the data integration solution. When an application user queries the mediated schema, the data-integration solution transforms this query into appropriate queries over the respective data sources. This implies that when information is needed from any of the respective data sources, the mediated schema could be queried as well. Eventually, virtual database harmonizes these results into the answer the user may have requested just as depicted in fig 1.

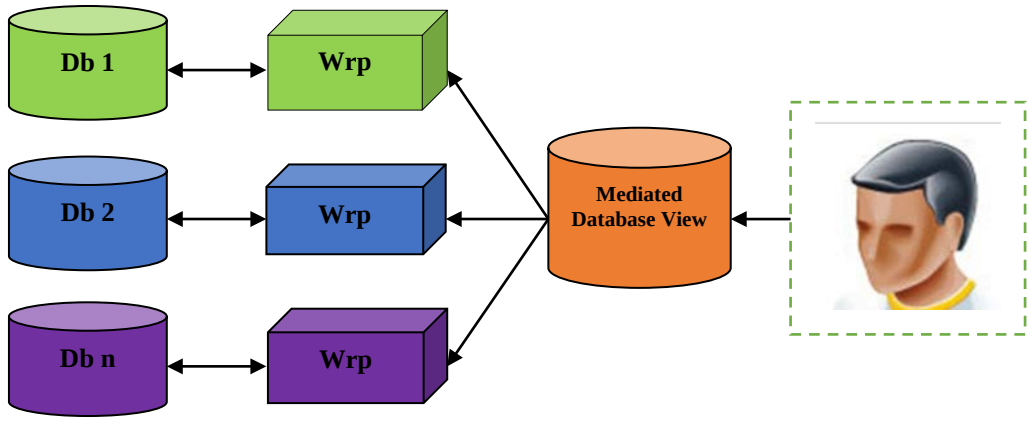

FIGURE 1: Mediated Schema; Db-Database and Wrp-Wrapper

Schematic and mathematical relationship among wrappers, data source and database

From a technical viewpoint, the difficulty comes from the lack of interoperability between the data sources that may use a variety of schemas, specific query processing capabilities and different protocols. However, the real bottleneck for data integration is logical. It comes from the so-called *semantic heterogeneity* between the data sources. They typically organize data using different schemas even in the same application domain. For instance, each organization may choose to model identification of individuals in its own way. The Nigerian Population Commission (NPC) may use the social security number to identify persons while Nigerian Identification Management Commission (NIMC) may use biometrics (fingerprint) to model theirs. The main issue here is to specify the relationships (i.e. semantic mappings) between the schemas of the data sources and the global schema. Based on these mappings, one can answer queries over the global schema using queries over the data sources.

Typically, query answering in the mediator approach is performed as follows. First, independently of the data in the sources, the user's query posed over the global schema is transformed into *local queries* that refer to the schemas of the data sources. A *global query* combines the data provided by sources. Queries are optimized and transformed into query plans. The local query plans are executed and their results combined by the global query plan. We consider that the global schema and the schemas of the data sources to integrate are all relational. In practice, each non-relational data source (e.g., XML or HTML) is abstracted as a relational database with the help of a *wrapper*.

From an article published by L. Libkin, Let us consider in more detail, the specification of semantic mappings between the data sources and the global schema.

If S_1, $S_{2,\ldots,}$ S_n represent the local schemas of n pre-existing data sources, where S_i is the local variable then, to simplify the presentation, we assume that each local schema S_i is made of a single relation that is denoted also by S_n. The relations S_1, S_2,.., S_n are called the *local* relations.

Suppose the global schema G consists of the *global* relations G_1, G_2,...,G_m where G_i is the local variable, the goal is to specify semantic relations between the local relations S_n and the global relations G_m. The G_m are logically and intentionally defined by the *Sn* such that:

$G_1 = S_1$……..…..…1.0

One can find more complicated relationships,

$G_2 = S_1 \cup S_2$ ……..…1.1

$G_3 = S_1 \bullet S_3$…….....1.2

Generally, expressing the semantic mappings between $\{S_1,...,S_n\}$ and $\{G_1,...,G_m\}$, inclusion statements can be used; i.e., logical constraints, of the form $v(S_1,...,S_n) \subseteq v'(G_1,...,G_m)$, with v and v' being query expressions called *views*. Now, given an instance I of $\{S_1,..., S_n\}$ (i.e., an instance of the data sources), we don't know the instance J of the global schema, but we know that:

$V(I(S_1),\ldots ,I(S_n)) \subseteq v'(J(G_1),\ldots ,J(G_m))$…….……..…..…1.3

We consider two kinds of views to model the mappings

i. Global-As-View
ii. Local-As-View

Global-As-View

The semantic mappings are of the form

$V_i(S_1,\ldots, S_n) \subseteq G_i$ ……..1.4

also equivalently denoted

$G_i \supseteq V_i(S_1,\ldots, S_n)$……..…….....…1.5

where each *Vi* is a view over the local schemas, i.e., a query built on local relations. For example, consider the following five different organizations as sources and some of their properties;

S_1.INEC (voting status, fingerprint, Nationality)
S_2.FRSC (Names, Literacy, LicenseID)
S_3.FIRS (Nationality, work status)
S_4.NYSC (Studentname, university)
S_5.NPC (Names, SSN, Nationality, Literacy, Fingerprint, work status)

Now, suppose we define NPC to act as the global source of our schema; this implies that any other organization that needs to store an individual's information in its own database doesn't need to key it in all over but instead pull from the Global using an attribute.

PesonalDetails = (Names, SSN, Nationality, Literacy, Fingerprint, work status…)

These relations are defined in terms of the local relations by the following GAV mappings

S_1.INEC (Fingerprint, Voting status) $\supseteq$ PersonalDetails, VotersDetails
S_2.FRSC (LicenseID) $\supseteq$ PersonalDetails, DrivingLicenceDetails
S_3.FIRS (Work status) $\supseteq$ OccupationDetails, CompanyDetails, PersonalDetails
S_4.NYSC (StudentName) $\supseteq$ StudentDetails, PersonalDetails

Local-As-View (LAV)

For LAV, the semantic mappings are of the form

$V_i(G_1,\ldots,G_m)$……1.6

Where each V_i is a view over the global schema, i.e., a query built on global relations.

LAV mappings enable quite fine-grained descriptions of the contents of data sources. LAV mappings express loose coupling between local and global relations, which is important for flexibility and robustness when the participating data sources change frequently. If we are interested in a student record, we do not need to know in advance (unlike the GAV approach) how to join two sources from different organization to get that. We just define them as a global query.

In summary, when there is need to update a particular record of an individual, a single attribute such as SSN can be used to query the database (Global) and then other information can be added.

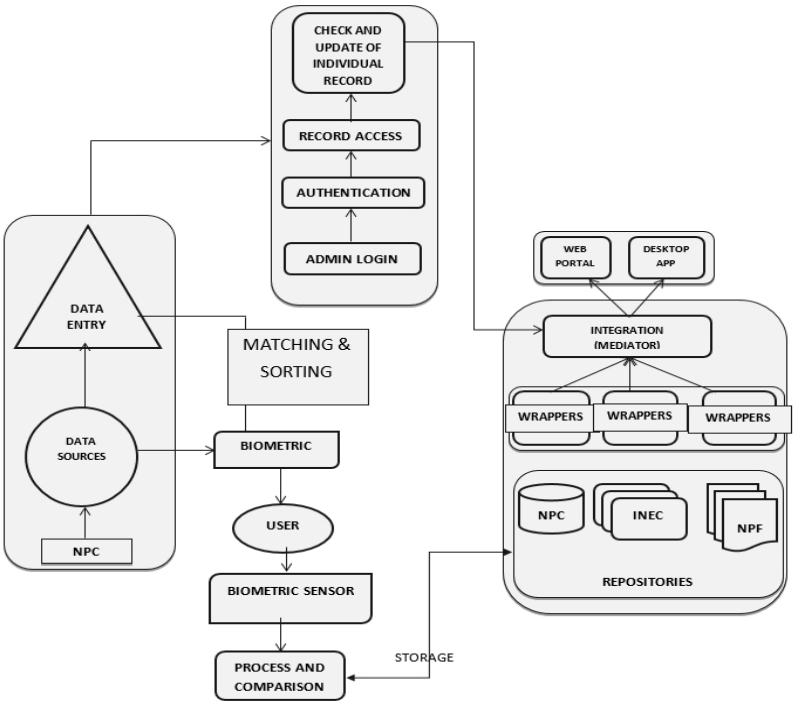

Figure 2 schematic flow of data in the databank

System implementation and requirements

This chapter explains the implementation details of the development of the National Data Harmonization System. This project was developed as a web-based application using ASP.NET C#. During this implementation phase, the software design of this project represented in form of relational diagram in chapter three is transformed into a functional and workable application that users can interact with. Other tools include MSSQL server, Internet Information service (IIS), Visual studio 2013, Google Chrome, Internet Explorer and Biometric Scanner. Implementing this project requires some basic system configurations like 2GB of RAM, 320GB of HDD, a reasonably high processor speed, internet access and stable power source.

Conclusion and future work

In this project, we've been able to present a detailed and robust databank system with an inclusion of biometrics which is an improvement, for cross-reference purposes without any data replication and this in turn has reduced redundancy. Its architecture has also been explicitly explained alongside its implementation with the use of the finger print scanner. There is also room for expansibility in that, new organizations can be included in the databank. However, for further research a stronger biometric machine such as retina scanner, face detector to mention but a few can be used for

implementation. Also, Cloud storage is also recommended as an efficient storage facility for the databank system since the dataset is very large.

References:

1. Abiteboul, S., Manolescu, l., ligaux, P., Rousset, M.-C., & Senellart, P. (2011). Data Integration. In *Web Data Management* (p. 196). Cambridge: Cambridge University Press.
2. Begg, & Carolyn, T. C. (2005). Database Systems, A practical approach to design,implementation and management. In T. C. Begg, *Database Systems, A practical approach to design,implementation and management* (p. 1427). U.S.A.
3. Data Intgration Info. (2015, August 11). *Data Intgration Info.* Retrieved September 17, 2015, from http://dataintegration.info/data-integration
4. Dr., L. (2010). A Proposed Cryptography-Based Identity Management Scheme For Enhancing Enterprise Information Systems Security. *CISDI, i*(2), 6.
5. Elliott, W. J. (2011). *The National DNA Databank of Canada.* Canada.
6. Flexview. (2011, March 03). *Flexview*. Retrieved June 2015, from Flexviews: code.google.com/p/flexviews/
7. Morgan, G. (2013). Data Virtualisation on rise as ETL alternative for data integration. In G. Morgan, *Computer Weekly*. Wikipedia.
8. Nguyen, & Thien-Loc. (2003). *National Identification Systems.* Macssachussetts: Macssachussetts Institute of Technology.
9. PASCAL, & NYAMULINDA. (2014). IMPLEMENTATION OF A NATIONAL IDENTIFCATIOJN PROGRAM. In *WORLD BANK FORUM* (pp. 31-35). Rwanda.
10. Siboniso C. Makhaye, P. T., & Bayaga, A. (2014). Designing dynamic federated identity management framework for reduction of management overhead in cloud computing., (pp. 7-8). Zululand.
11. Vitanen, & Samu. (2015). *Integrating an Open Source Identity Management System into Access Management Software.*
12. Yu, Z., YI, F., & Chi, X. Z. (2008, March). *National Institute of Health*. Retrieved June 2015, from http://www.ncbi.nlm.nih.gov/pubmed/18788587