Enabling Seamless Sharing of Data among Organizations Using the DaaS Model in a Cloud

Addis Mulugeta
Ethiopian Sugar Corporation, Addis Ababa, Ethiopia
addismul@gmail.com

Abrehet Mohammed Omer
Department of Computer Science and IT, Addis Ababa Science and Technology University, Ethiopia
abrehet@gmail.com

Abstract

Organizations are struggling to allow seamless data sharing and synchronization with the intention of data verification and avoid data redundancy. Currently, many organizations in Ethiopia share data manually using a formal letter. This is not economically as well as technically feasible. It is tiresome, inefficient, error prone and impractical as the systems and problems become larger and more complicated. This paper explores data sharing methods anticipated functionalities through inspecting real cases and different existing data sharing mechanisms by proposing an extended framework, including: (1) identifying the core issues that have to be addressed during data sharing, i.e., data owners autonomy, heterogeneity, and need of naming standard; (2) developing a customized conceptual framework, which allows providing data as a service and specifies different components of data as a service provider and service consumer, i.e., Acquaintance manager, Query manager and ECA rule manager; (3) developing a detailed implementation framework; (4) demonstrating the framework by implementing a prototype based on the implementation framework. The proposed approach takes advantage of the newly emerging computing paradigm, cloud computing, that enables providing Data as a Service.

Keywords: Cloud Computing; Data as a Service; Data Sharing

1. Introduction

The need for sharing data between different organizations has been increasing with the technology and economic growth of the country. Sharing data among organizations will help organizations to increase efficiency and performance by reducing manual work, in order to get accurate data, and to improve delivery of the service for the customer. In addition, with the current high mobility and advancement of technologies, few illegal persons take advantage of sharing data between the organizations for illegal purpose that increase the demand of data verification.

Moreover, many organizations build their own database to be utilized only within the organization even if the same data is stored in another organization. As a result, the same data might be stored in multiple locations resulting in data redundancy. This redundancy can be avoided by sharing existing data.

Advances in computing paradigms and technologies triggered a new multitude of data sharing mechanism, which is providing data as a service. This new computing paradigm is called cloud computing which promises the provision of providing anything as a service [2]. Cloud includes the different participants involved in the cloud along with the attributes and technology that are coupled to address their needs and different types of service “XaaS” [1] X is a software, hardware, platform infrastructure, data business, etc. Everything is a service like Software as a Service (SaaS), Hardware as a Service (HaaS), Data as a Service (DaaS), Infrastructure as a Service (IaaS), etc.

DaaS is based on the concept that the product, data in this case, can be provided on demand to the user regardless of geographic or organizational separation of provider and consumer. The concept of DaaS advocates the view that with the emergence of service-oriented architecture (SOA), which includes standardized processes for accessing data "where it lives”, the actual platform on which the data resides doesn't matter. With data-as-a-service, any business process can access data wherever it resides [2].

Currently, many organizations in Ethiopia share data manually using a formal letter for the purpose of data verification. This is not economically as well as technically feasible. It is tiresome, inefficient, error prone, and impractical as the database, systems and problems become larger and more complicated. In addition, from the real case study analysis it is observed that: (1) when any organization builds a database, it only considers internal requirements which imply data is stored in heterogeneous manner; and (2) all organizations decide to share or not to share data autonomously for security purpose. Thus, any data sharing mechanism is expected to be secure, flexible, and personalized.

This paper discusses ideas concerning the data sharing mechanism that includes the expected functionality mentioned above. Our approach relies on the concept of providing data as a service based on demand. We define our approach by developing a conceptual and detailed implementation framework that includes all components to handle heterogeneity, keep autonomy, and allow data synchronization.

The remaining part of the paper is organized as follows: Section 2 provides one of the case studies used to identify the problem at hand. Section 3 gives an overview of the proposed framework, conceptual as well as implementation framework. Section 4 presents the prototype implemented to demonstrate the proposed framework. Section 5 presents related work and compares our approach with some related work and finally Section 6 presents conclusions.

2. Case Study

The main purpose of this study is to know how an organization handles data and the sharing mechanism, the policy of sharing data and a detailed study on the organization, which verifies data of many persons with different organizations.

The case study is done on the Federal Ethics and Anti Corruption Commission (FEACC). One of the tasks of this commission is to register and verify the wealth of an appointee of some government employees. The commission has a directorate that registers and verifies thousands of appointee employees and their family data with all banks to check their account, municipality data to check house and land they owned, investment office to check their investment activities, and with other organizations in which each organization may have hundreds of branches all over Ethiopia.

Even though all Banks, Ethiopian Revenue and Custom Authority, Ministry of Trade, sub cities and other organizations own databases, there is no automated way to share data. When data verification is needed, the organization verifies data manually using a letter. In addition, there are many organizations with similar situation in the country, which need data for sharing and verification purpose; for example, banks to verify the legality of collateral, courts to check the legality of different documents from different organizations.

According to the study in the organization

1. The organizations are Autonomous
2. The databases found in the organizations are heterogeneous
3. There is no attribute naming standard.

The core issues to be addresses are the heterogeneity of the databases and the autonomy of database on each organization.

3. The Proposed Solution

From the different data sharing and the case study, we designed a conceptual framework and a detailed implementation framework. We adopted and modified the approach from the Hyperion project to fit the requirement from the case study.

3.1 Conceptual Framework

The proposed conceptual framework follows a peer-to-peer data sharing mechanism based on the Hyperion project [3]. It has two kinds of peers: the service provider peer and the service receiver peer as shown in Figure 1.

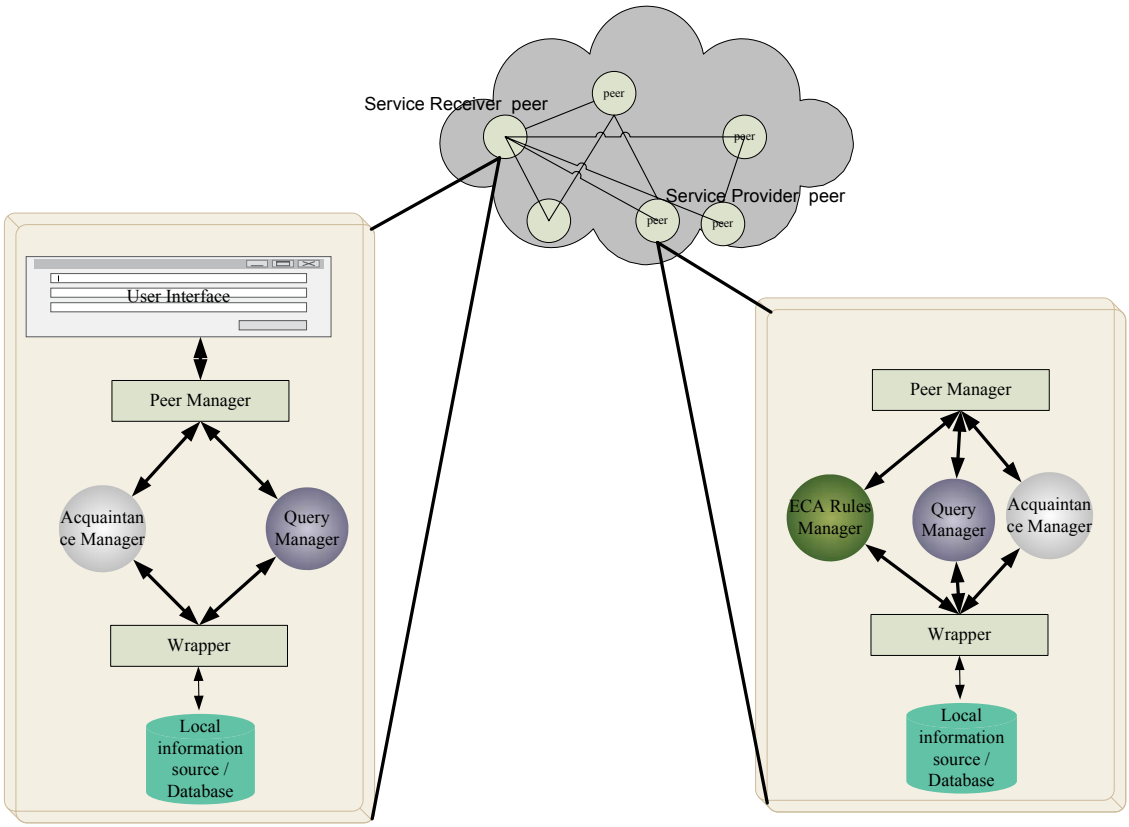

Figure 1: Proposed Conceptual Framework

3.1.1 Services Provide Peer

The major components are:

1. *Acquaintance manager:* manages the exchange of public schemas, mapping tables, and coordination rules between new acquaintances and the inference of new mappings.
2. *Query manager:* facilitates the ability to execute local queries to the format in which the query that the data model holds, and the result from the data model to the users can understand by rewriting the queries and mapping.
3. *ECA rule manager:* manages and executes distributed event condition rules in order to enforce consistency of the shared or provided data.
4. *Wrapper:* handles data conversion and table mapping from the organization's format to the data model format.

3.1.2 The Service Receiver Peers

Similarly, the service receiver includes the following components:

1. Acquaintance manager
2. Query manager
3. Wrapper: the wrapper in the service receiver changes from the data model format to the format which the user needs

3.2 Implementation Framework

For the actual implementation of the conceptual framework, a detailed framework is developed as shown in Figure 2. This framework gives details on how a data provider peer can provide services and how the service consumer peer can receive the service. The different components of service provider peers that enable providing data as a service are described below.

a. Database

The databases in different organizations are in different format. The files vary from simple flat files like Excel or text to commercial databases like Oracle, MS SQL, MySQL, etc.

b. Wrappers

The main purpose of the wrapper is to convert heterogeneous databases into a single format and to resolve conflicts caused by heterogeneous databases and put the data on the data model. XML is selected for the database. XML technology makes the

standardization description of all kinds of irregular information and rule information possible, and gradually becomes the standard to describe the data on the Internet, and sets up an enterprise information integration platform in the XML technology, and is an inevitable trend of information technology development [4]. In the wrapper, there is a schema and a table mapping. On the schema level, all the components of the database are changed to the XML database format. Each attribute on the database is mapped to the attribute of the standard data model. A standard data model is a table which is designed to accommodate all of an organization's databases with different data types and name to the standard name and data type which the user understands and queries easily.

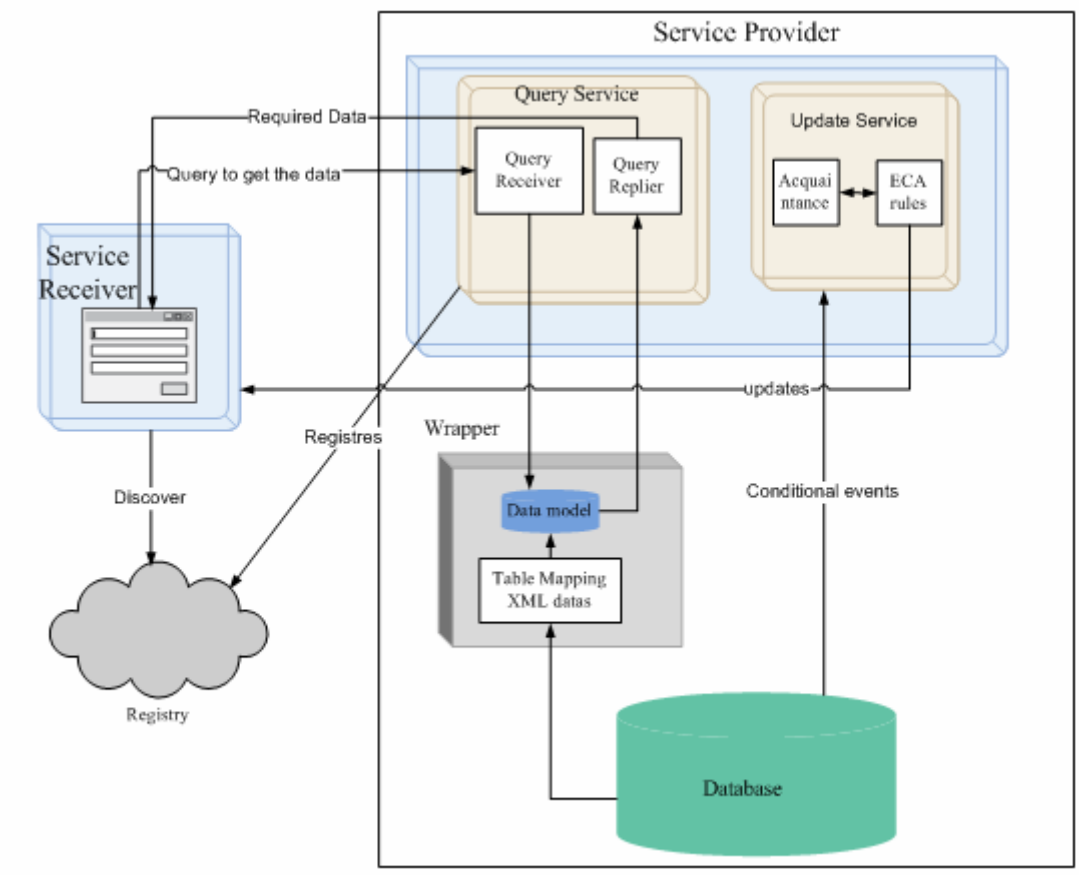

Figure 2: Proposed Implementation Framework

c. Update Service

This service has two components: acquaintance and ECA rules. Acquaintance is used to store and manage the peers. ECA is a coordination rule that describes when, how, and where a message should propagate to the concerned peers based on the agreements of peers (organizations). These components allow data synchronization among peers.

d. Query Service

The main purpose of the query service is to provide a unified interface for the client to query the required data from the heterogeneous databases. The query service accepts the standard query from the interface and changes to the query that is used to fetch the data from the data model that is found in XML format.

In addition to the above listed components, the framework includes the service receiver and registry components which are actually included to show the complete system of DaaS in a Cloud.

The service receiver can be any organization which needs the data from different organizations for verification and sharing. The registry holds constantly evolving information about the available services in the service provider which allows the service receiver efficiently discover and

communicate with service providers. When the service receiver needs data, it gets the available services from the registry and then communicates with the service providers and using a browser the receiver gets the necessary information.

4. Prototype

A prototype is developed to demonstrate and validate the proposed framework. The prototype shows all the implementation except the update service.

The prototype is developed with the Talend open software and Microsoft Visual Studio 2010. Talend open source software is used to implement the wrapper task.

The wrapper

- Copies the data from the organization,
- Changes the database from any format to the XML format or any other standard format,
- Maps the table of the organization into the data model table,
- Puts the data in the data model, and
- The data model is ready for the service.

The Service Receiver

- Searches the available service from the registry and gets the address of the registry,
- Within the address enters into service receiver page,
- In the page the user fills the form accordingly and posts the data,
- The service provider accepts the data query, process it and returns back the result,
- The data which is obtained by the query in XML format is changed to the standard format, and
- The data will be displayed in the page.

The prototype is designed to convert data in the organization that intends to share its data to the XML format and put in a file. Using a web interface from the XML file, the user queries data and the result is shown in Figure 3.

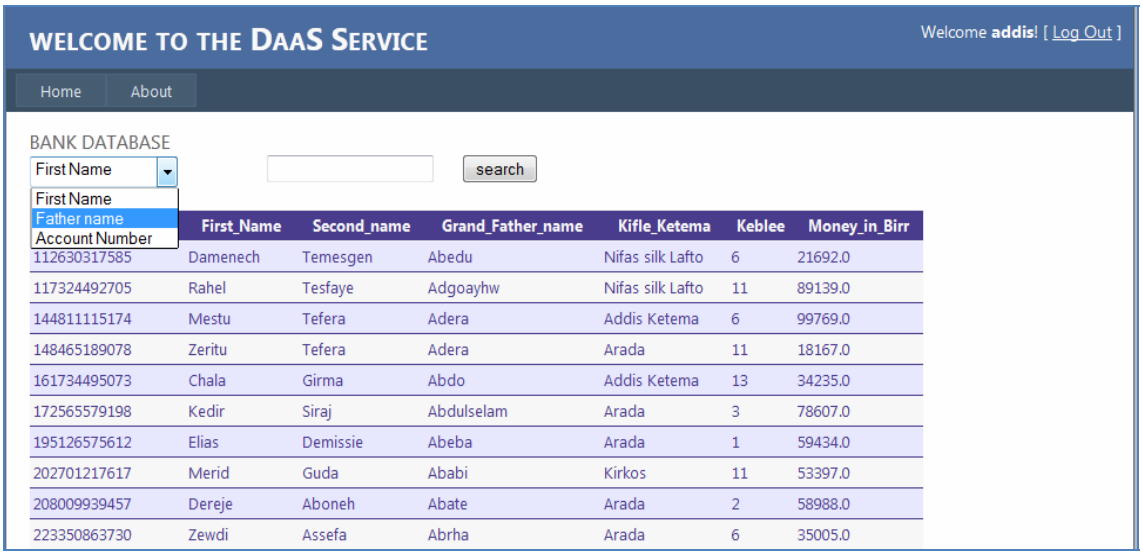

Figure 3: Interface for the user

5. Related Work

5.1 NETDB2

NetDB2 is a data sharing mechanism proposed in [5]. The basic NetDB2 system is implemented as three-tier architecture, namely; the presentation layer, the application layer, and the data management layer. The benefits of the layers are the insulation of software components of one layer from another and better interoperability and higher scalability. This data sharing mechanism is used by a number of universities to help teaching database courses at different locations.

5.2 Clinical Pharmacology and Data Management

Bayer Schering Pharma proposes a data sharing technique that enables sharing of data among

different departments (entities) of clinical pharmacology [6]. It proposes a centrally managed data sharing mechanism known as an exchange area. An exchange area is a static directory tree on a UNIX server with subdirectories. It is used to archive files that had been transferred or shared. All data for one transfer are compressed in a ZIP archive and named according to the direction of data transfer. UNIX drives are mapped within the environment of other operating systems. The transfer from the area to the departments is performed automatically by a cron job in predefined time interval. Cron is a UNIX utility that allows tasks automatically run in the background at regular intervals.

5.3 Hyperion Project

Hyperion project is a system that supports data sharing for a network of independent peer Relational Database Management Systems. In the Hyperion project, each peer includes a database with its own schema and data. Peers can join or leave the network at their own discretion. Moreover, a peer may form an acquaintance with another peer for data sharing purposes. These metadata take the form of mappings, both at the data level and schema level, and they help to bridge semantic and syntactic heterogeneities between peers. Metadata at the data level are expressed as mapping tables [3]. The framework fulfills all the necessary requirements as shown in Table 1 that are not fulfilled in any of the data sharing mechanisms mention above.

Table 1: Comparison of Data Sharing

	Data source Heterogeneity		*Data base type*	*Autonomy*	*Data as a service*
	Schema Matching	*Table mapping*			
NET DB2	Yes	No	Relational	No	Yes
Hyperion	No	No	Relational	YES	No
Clinetics Data Merge	No	No	Any Type	No	No
Proposed Architecture	Yes	Yes	Any Type	Yes	Yes

6. Conclusion

In this paper a framework is proposed which is used to share data seamlessly by autonomous organizations from heterogeneous databases. This approach uses a cloud computing DaaS model in which the data is available as a service. The users get the data wherever there is a browser and a connection regardless of the operating system or the need of other software. An XML database is used to facilitate the provision of data as a service. The proposed framework is demonstrated with a prototype for which any organization can easily share data by keeping their autonomy and with heterogeneous database systems.

References

[1] B. P. Rimal, E. Choi, and I. Lumb, "A Taxonomy and Survey of Cloud Computing Systems," *in 2009 Fifth International Joint Conference on INC, IMS and IDC*, 2009, pp. 44–51.

[2] J. Dyché, "Data as a service explained defined." http://searchdatamanagement.techtarget.com/answer/Data-as-a-service-explained-and-defined, accessed on 12-Nov-2011.

[3] A. Kementsietsidis, "Data Sharing in the Hyperion Peer Database System," on Very Large Data, pp. 1291–1294, 2005.

[4] W. Xiaoli and Y. Yuan, "XML-based Heterogeneous Database Integration System Design and Implementation," *2010 3rd International Conference on Computer Science and Information Technology*, pp. 547–550, Jul. 2010.

[5] H. Hacigumus, B. Iyer, and S. Mehrotra, "Providing Database as a Service," *in Proceedings 18th International Conference on Data Engineering*, 2002, pp. 29–38.

[6] B. V. Gmbh and A. Schwarzer, "Clinetics Data Merge - A Platform for Exchange of Pharmacokinetic Data," pp. 1-5, 2010.