University College of Takestan

Available online at **http://UCTjournals.com**

UCT Journal of Management and Accounting Studies

UCT . J.Educa.Manag .Account. Stud., (UJMAS)

186-191 (2015)

Volume 3,Issue3

ISSN 2382-9748

The Best Data Integration Method by Using SAW Technique in Ministry of Science, Research and Technology (MSRT)

Mahsa Shokri[1] and Ali Reza Pourebrahimi[2]

1 Department of management Information Technology, Electronic Branch, Islamic Azad University, Tehran, Iran
2Associate Professor in Industrial management, Karaj Branch, Islamic Azad University, Karaj, Iran
mahsaashokri@gmail.com
poorebrahimi@gmail.com

ABSTRACT

Many computer systems are usually used by the organizations. The data within different systems should be integrated in order to analyze and provide a report.

The method of saving data and files often focused in the past, the data structure and flow received little attention, therefore, large volume of data and data transfer management carry into account as a big challenge. Since information systems are growing in the organizations, the number and entanglement of interfaces will increase. If the data integration strategy is disregarded, the organization will meet numerous problems.

There are 4 different data integration methods which have their special usage. This article concentrates on the process of data integration and aims to understand the meaning and methods of data integration, to identify variables and effective factors on data integration, to rank the variables of data integration and to provide the best method of data integration by SAW Technique for the Ministry of Science, Research and Technology.

Original Article:

Received 22 June. 2015
Accepted 14 Sep. 2015
Published 30 Sep. 2015

Keywords:
Data integration, Simple Additive Weighting (SAW), MSRT, Ranking

INTRODUCTION

Data integration (DI) by Philip Russom (2011) is a family of techniques and best practices that repurpose data by transforming it as it's moved. ETL (extract, transform, and load) is the most common form of DI found in data warehousing. There are other techniques, including data federation, database replication, data synchronization, and so on. Solutions based on these techniques may be hand coded, based on a vendor's tool, or a mix of both. DI breaks into two broad practice areas. Analytic DI supports business intelligence (BI) and data warehousing (DW), and operational DI is applied outside BI/DW to the migration, consolidation, and synchronization of operational databases, as well as in exchanging data in a business-to-business context. The problem of data integration is universal. It is necessary to use an example in order to elaborate on as described below. You suppose that have multiple sloppy, out of date and opposite information. It is recommended that the data ranked and combined as a method so that a suitable answer can be gained. [2]

1. Goals of Data Integration

The data integration system helps to recite homogenous access to a set of independent and inconsistent data sources. The goals of data integration by AnHai Doan, et.al (2012) are Query, Sources, Number of sources, Heterogeneity and Autonomy which are described in the following paragraphs:

Query: The main goal of most data integration systems is on searching inconsistent data.

Sources: Updating the sources is fairly of interest.

Number of sources: Data integration is already a challenge for a small number of sources (fewer than 10 and often even 2!), but if the number of sources increases the challenges will be serious. Also, when the sources grow intolerably, it is recommended to support Web-scale data Integration.

Heterogeneity: A typical data integration outline involves data sources which were developed independently of each other. At the result, the data sources implement in different systems: some of them are databases, but others may be content management systems or simply files biding in a directory. The sources will have different patterns and references to objects, even when they model the same domains. Some sources may be completely structured (e.g., relational databases), while others may be unstructured or semi-structured (e.g., XML, text).

Autonomy: The sources do not necessarily belong to a single administrative entity, and even when they do, they may be run by different sub organizations. Therefore, it is impossible to suppose that there is complete access to the data in a source or that there is access the data whenever is necessary, and considerable care needs to be given to respecting the privacy of the data when appropriate. Furthermore, the sources can change their data formats and access patterns at any time, without having to notify any central administrative entity. [3]

This article focuses on the data integration methods in order to identify and rank variables and effective factors on data integration and offer the best method of data integration for

the Ministry of Science, Research and Technology (MSRT) by SAW technique.

2. Data Integration Models

The different kinds of data integration models are explained in this section. It is noteworthy that the main challenges for both persistent and mobile data by April Reeve April Reeve (2013) [4] are data access and security management. It is noteworthy that the persistent data security is usually managed in 5 layers: physical, network, server, application, and at the data store. Data motion across the applications and organizations requires more security to protect the data in motion from unauthorized access.

Recovery from failures in processing by April Reeve April Reeve (2013) is very important for both persistent data processing and transient data processing. The techniques for recovery are very different but with some related approaches for the two methods. Every technology may use different solutions to recover the data which maybe suitable for various business and technical solutions. Two following questions play a main role in choosing an appropriate solution:

- How much data can be allowed to be missed in a failure
- How long the systems can be down before recovery must happen.

With persistent data, the model or structure of the data being stored received attention mainly. In managing data in motion, the largest concern is how to associate, map, and transfer data between different systems. There is an important part of the implementation of data integration solutions which involves the modeling of the data in transfer and the use of a central model of the data passing between applications; this is called canonical modeling. [4]

There are 4 data integration methods by AnHai Doan, et.al (2012), [3] which are:

- Batch data integration
- Real-time data integration
- Big data integration
- Data virtualization

3.1. Batch Data Integration

Batch data integration by April Reeve (2013) occurs when data to be transferred from a source to a target is grouped together and sent periodically, such as daily, weekly, or monthly. Most interfaces between systems in the past used to be in the form of passing a large file of data from one system to another on a periodic basis. The contents of the file would be records of consistent layout, and the sending and receiving application systems would agree to and understand the format. The passing of data between two systems whereby a sending system passes data to a target receiving system is called "point to point".

The data file would be processed by the receiving system at some point in time, not necessarily instantaneously; thus the interface would be "asynchronous" because the sending system would not be waiting for an immediate confirmation before the transaction would be considered complete. The "batch" approach to data integration is still appropriate and effective for very large data interactions such as data conversions and loading data snapshots into data warehouses. This type of interface can be tuned to be extremely fast and is appropriate where very large volumes of data need to be loaded as quickly as possible. It is also described as "tightly coupled" because the format of the data file must be agreed to between the systems and can only change successfully if the two systems involved incorporate knowledge of the change simultaneously. [4]

2.2. Real-Time Data Integration

Interfaces that are necessary across systems immediately in order to complete a single business transaction are called "real-time" interfaces by April Reeve (2013), Usually they would involve a much smaller amount of data being passed in the form of a "message." Most real-time interfaces are still point to point between a sending and receiving system and tightly coupled because the sending and receiving systems still have specific agreement as to the format, such that any change must be made to the two systems simultaneously. Real-time interfaces are usually called synchronous because the transaction will wait for the data interface to complete its processing in both the sending and receiving systems. Best practices in real-time data integration solutions break away from the complexity problems of point-to-point and tightly coupled interface design. There are logical design solutions that can be implemented in various technologies. These technologies can be used to implement inefficient data integration as well, if the underlying design concerns are not understood. [4].

2.3. Big Data Integration

The term "big data" by April Reeve (2013) indicates that there are large volumes of data, as well as data of various technologies and types. Taking into account the extra volumes and various formats, data integration of big data may involve distributing the processing of the data to be performed across the source data in parallel and only integrating the results, because consolidating the data first may take too much time and cost too much in extra storage space. Integrating structured and unstructured data involves tying together common information between them, which is probably represented as master data or keys in structured data in databases and as metadata tags or embedded content in unstructured data. [4]

2.4. Data Virtualization

Data virtualization by April Reeve (2013) involves using various data integration techniques to consolidate data real-time from various sources and technologies, not just structured data. "Data warehousing" is a practice in data management whereby data is copied from various operational systems into a persistent data store in a consistent format to be used for analysis and reporting. The practice is used to do analysis across snapshots of historical data, among other things, which is difficult using active operational data. Even when the data required for analysis is only current data, the reporting and analysis architecture usually involves some persistent data store, such as a "data

mart" because the real-time integration and harmonizing of data from various sources has previously been found to be too slow for real-time consumption. However, new data virtualization technologies make real-time data integration for analysis feasible, especially when used in conjunction with data warehousing. Emerging technologies using in-memory data stores and other virtualization approaches allow for very fast data integration solutions that do not have to rely on intermediate persistent data stores, such as data warehouses and data marts. [4]

This paper is organized as follows. The different kinds of data integration models are described in section 2. The data integration methods and their advantages are presented in section 3. In section four, the results of the proposed method are elaborated on using SAW technique. Finally, the paper is concluded in section 5.

3. Data Integration Methods

In this section, the different kinds of data integration methods are described. Several options exist for integrating applications within the enterprise. Sapient (2004) [5] has introduces four data integration methods which are

- Custom Point-to-Point Integrations
- Messaging or EAI (Enterprise Application Integration) Tools
- Web Services
- ETL (Extract, Transform, Load) Tools
-

4.1. Custom Point-to-Point Integrations

Custom Point-to-Point Integrations by Sapient (2004) is a direct point-to-point link is created between applications for each business function. This method has been designed and implemented solely in order to integrating two specific systems directly and can be used for both real time and batch integrations. Furthermore, it is possible to transfer the inconsistent data by special protocol.

Custom Point-to-Point Integrations has custom code for data extraction, business rule processing, data loading and custom data format. [5]

This method [5] has the other advantages such as the following:

- No need to invest in expensive tools up front
- No need for developers to learn new skills and packages
- No extended time frame for developing and deploying enterprise integration strategy

4.2. Messaging or EAI (Enterprise Application Integration) Tools

Messaging or EAI (Enterprise Application Integration) Tools by Sapient (2004) is source systems "publish" enterprise messages to a common bus; application "subscribe" to relevant messages and act on them. This method can wrap each application independently and acts as a dealer between applications. It is possible to store and forward messages by EAI.

EAI not only provides near real-time, guaranteed, once-only delivery but provides an environment in which to define rules. [5]

This method [5] has the other advantages such as the following:

- Systems are integrated but not coupled
- Business rules are centralized in the message broker and transformation engine
- Allows for near real-time integrations which reduced latency
- Solves the n2 problem; as the number of systems increases, the integration effort expands linearly

4.3. Web Services

Web Services by Sapient (2004) is a data integration method which is functionality to be integrated is exposed via XML on an open protocol such as SOAP. Other systems can consume this service if needed. Inputs and outputs to the web services are XML. This method uses a Common language of communication between heterogeneous systems. Web Services has been designed base on standard Internet technologies. The method is been equipped by Self describing and advertising. It is interesting to know that Web Services Supports dynamic discovery & integration. The services of this method are set easily within an overall architectural model. As a final attribute, it is supported widely by major vendors. This method has the advantages such as the following:

- Solves problems similar to those EAI solves
- Need for expensive integration tools
- Use of proprietary integration platforms

4.4. ETL (Extract, Transform, Load) Tools

ETL (Extract, Transform, and Load) Tools by Sapient (2004) [5] is a standard set of tools and processes used to extract, transform and load large volumes of data between systems. It is very useful in populating a data warehouse. This method provides tools for data cleansing which includes correcting misspellings, resolving conflicts (city & zip code incompatibilities), missing elements, parsing elements. It is possible to combine data sources, matching on key values, fuzzy matches on non-key attributes, and textual comparisons to reference tables. ETL Can create surrogate keys: Operational systems and the data warehouse have different assumptions and data requirements thus the data warehouse requires its own set of primary keys.

This method has many considerable facilities. For example: De-duplicate processing: Identifying and eliminating duplicates

Create aggregates to boost performance of common queries in data warehouses and data marts Loading and indexing: For large data warehouses specialized bulk loading processes are required.

This method [5] like the other methods has the advantages such as the following:

- Extremely efficient for moving large volumes of data in short timeframes
- Applies consistent transformations
- Provide or integrate with meta-data for the enterprise data model

4. Experiment and Analysis

Analysis includes the Cognitive interview with experts and the data of the article are collected by questionnaires.
Firstly, the library studies have been done, therefore, the articles have studied which are the result of the researches and implementation of data integration in some organizations which are developed in systematic integration and the effective indexes of integration were extracted. For Example, Many developed countries implement e-government according to the integration concepts. Secondly, the questionnaire was prepared according to the indexes and the ideas of IT experts of Ministry of Science, Research and Technology (MSRT) completed them.
The questionnaire has used Likert scale by considering the objectives and the interpreting of the results in order to identify the indexes and weight them by SPSS Software. The table 2 shows the results of final weight of indexes and their ranking. Secondly, it is necessary to be collected and be studied the data integration methods accompanied by their advantages and disadvantages; finally, this article suggests the best method for MSRT by SAW technique.

5-1. Simple Additive Weighting (SAW)

Simple Additive Weighting (SAW) by Memariani (2009) is one of the oldest methods in MADM. Multi-attribute decision making models (MADM) are selector models and used for evaluating, ranking and selecting the most appropriate alternative among alternatives. [6]
In SAW technique by Hwang (1981) final score of each alternative is calculated as follow and they are ranked. [7]

$$p_i = \sum_{j=1}^{k} w_j . r_{ij} \dots ; i = 1,2,\dots m \qquad [7]$$

It is necessary to have the following information in order to use SAW technique by Faez (2010) [8]:

- Effective indexes in decision making
- Alternatives
- Evaluating each alternative across each index
- The importance of each index in decision making

It seems that the method of SAW by Lazim Abdullah (2014) is the MCDM methods that used for weight determinations and preferences. Churchman et al., (1957) were the first research utilized the SAW method to cope with the portfolio selection problem. This method is also known as a weighted linear combination or scoring method. It is simple and the most often method used multi- attribute decision technique. An evaluation score can be calculated for each alternative by multiplying the scaled value given to the alternative of hat attribute with the weights of relative importance directly assigned by the decision makers followed by summing of the products for all criteria. The advantage of SAW method is that it is a proportional linear transformation of the raw data. It means that the relative order of magnitude of the standardized scores remains equal. [9-10]

Table 1. Final weight of indexes and their ranking

Rank	Weight	Index
1	0.243	Communication and information technology capacity
2	0.231	Database technology
3	0.214	The method of access How to access
4	0.199	Available software capacity
5	0.19	Management capacities (Saving and monitoring)
6	0.176	documentation
7	0.165	The level of awareness and acceptance of internal users
8	0.159	The level of complexity
9	0.155	The level of awareness and acceptance of external users
10	0.147	Variety and multiplicity of systems
11	0.132	Financial advantages of data integration
12	0.131	The degree of dependence between programs
13	0.13	Implementation manager
14	0.126	The time value in the present processes
15	0.121	Diversity of sectors and users
16	0.093	comprehensiveness

Table2. Influence of each index on the data integration methods

Index / Method	Communication and information technology capacity	Database technology	The method of access	Available software capacity	Management capacities	documentation	The level of complexity	The level of complexity	The level of awareness and acceptance of external users	Variety and multiplicity of systems	Financial advantages of data integration	The degree of dependence between programs	Implementation manager	The time value in the present processes	Diversity of sectors and users	comprehensiveness
W_j	0.24	0.23	0.25	0.11	0.2	0.18	0.17	0.16	0.16	0.15	0.13	0.13	0.13	0.13	0.12	0.93
ETL	0.6	1	1	0.5	1	0.2	0.1	0.8	0.1	0.9	1	0.1	0.3	0.5	1	0.6
web service	1	0	1	1	0	1	0.1	0.5	0.1	0.2	0	1	0.8	0.5	0.1	0.5
Point to Point	0.5	0	1	0.2	0	0	0.1	0.3	0.1	0.5	0	0.5	0.5	0.5	0	0
EAI	0.5	0	1	0.3	0	0.2	0.1	1	0.1	0.9	0	1	0.3	0.5	1	1

Table3. Final score and ranking of data integration methods by SAW

Data integration method	Decision score	Ranking
ETL	2.29	1
web service	1.81	3
Point to Point	0.74	4
EAI	2.06	2

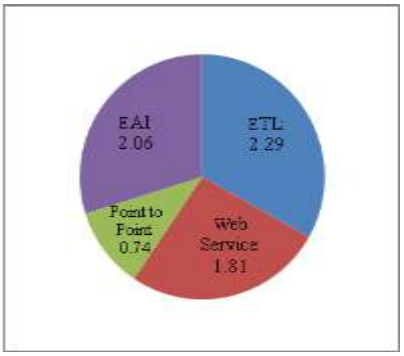

Figure1. The ranking of data integration methods in according to the indexes

5. Conclusion

It can be concluded that the best method for data integration is ETL by considering SAW method. Since it is difficult to justify the financial managers of an organization for allocating the financial sources for a long-term data integration project, therefore, ETL is the best way for implementing data integration.

Since it isn`t necessary to update systems in real time, it is recommended to avoid expensive methods such as EAI. But the above method isn`t the best method and lead to compound reports. It is necessary to suggest a suitable proposal and identifying the requirements of academic-information system as the main objective of each ministry such as MSRT. The data integration methods respectively ranked as the following:

- ETL: 2.29
- EAI: 2.06
- Web Service: 1.81
- Point to Point: 0.74

6. References

[1] Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246). ACM.

[2] Philip Russom(2011 May). What Works in Data Integration, Volume 31.

[3] Doan, A., Halevy, A., & Ives, Z. (2012). Principles of data integration. Elsevier.

[4] Reeve, A. (2013). Managing Data in Motion: Data Integration Best Practice Techniques and Technologies. Newnes.

[5] Sapient, (2004). (http://web.mit.edu/itag/eag-0.1/EnterpriseIntegrationOpts.pdf)

[6] Memariani, A., Amini, A., & Alinezhad, A. (2009). Sensitivity analysis of simple additive weighting method (SAW): the results of change in the weight of one attribute on the final ranking of alternatives. Journal of Industrial Engineering, 4, 13-18.

[7] Yoon, K. P., & Hwang, C. L. (1995). Multiple attribute decision making: an introduction (Vol. 104). Sage publications.

[8] F.Faez., SAW Method.(2010).(http://www.faez.ir/ CourseFile/ SAW.pdf)

[9] Abdullah, L., & Adawiyah, C. R. (2014). Simple additive weighting methods of multi criteria decision making and applications: A decade review. International Journal of Information Processing and Management, 5(1), 39.

[10] Churchman, C. W., Ackoff, R. L., & Arnoff, E. L. (1957). Introduction to operations research.