International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

XML-Based Data Integration

Sachin Samrat Medavarapu

Abstract: ***The purpose of the data integration system is to provide a unified global view to the user over the heterogeneous sources. Such data integration raises main challenges such as Schema creation and schema mapping. With the advent of improvement of Internet Technologies there has been a greater demand on the integration of data from diverse sources, especially in e-business. Schema Integration is done in two ways binary array and n-array. This paper discusses about n-array schema Integration.***

Keywords: data integration, schema mapping, heterogeneous sources, e-business, Internet technologies

1. Introduction

With the growth of large computer applications in recent years, the necessity of Data Integration came forward. Data Integration is the process of combining data from more than one data source. Modern scientific applications require integrating heterogeneous data [1]. With the exponential raise of web applications there is an exigency of combining different management systems and data storage systems. Besides, there is a necessity to link web applications and data sources.

Due to rapid development of massive application system, now a days it became problem to maintain huge data sources. The integration of heterogeneous data sources also become the big problem of modern computing. Daily over the network there are lot of persons tries to access the data available. So to maintain the data over the network the heterogeneous data integration technology is to be adopted [2].

Schemas and schema mappings are two important elements that are to be considered during Data Integration. Schemas define the structure and also semantics of data in data sources, whereas schema mapping is used to transform data between two different schemas and this is also used to transform queries from one source to another source [3]. During schema integration schemas of the sources are combined, but data is not touched.

The main approaches of heterogeneous data integration shown as follows [4]:

a) Federated Database:
A federated data base is the controlled and coordinated operation of DBMS component database management system (FDBMS). It has three characteristics like distribution, heterogeneity, autonomy. It is also one technology which is derived from the database; it is one kind of data organization and storage technology which is different from the database.

b) Wrapper Mediator:
Mediator receives request from the user and the request is transformed to child query using this process mediator optimizes the transformation to decrease the response time and increase the hit ratio. It integrates a particular data source and transforms it into the general data model which is to be accessed by the user, after the request transformed to wrappers, wrapper will interact with corresponding data source and execute the query.

2. Literature Review

XML: In the recent times XML (Extensible Markup Language) is highly used in developing web pages and it plays a major role in internet web pages. With this drastic change storage, XML documents became important.
Implementation Of Heterogeneous Data Integration Based On XML:

The heterogeneous data is integrated from various data sources with XML. The implementation takes place in different phases; It is divided into three main categories [5]:
1) XML schema integration phase
2) Query phase
3) Result integration

XML Schema Integration:
XML Schema allows machines to carry out the rules and expresses the vocabularies. It also provides support for defining structure, content and semantics of XML document. In current day's most of the applications developed using RDBMS like Oracle, DB2, Sybase, SQL Server. Every RDBMS stores the data in the form of rows and columns.

X Query:
It is also a query language which is flexible and concise and easily understood. It is derived from the other XML Query language called QUILT OQL, SQL, and XML-QL. It has special feature FLOWR, which is nothing but the, for, let, order by, where, return of sql language. It is divided into two steps they are; Schema creation and Schema Integration.

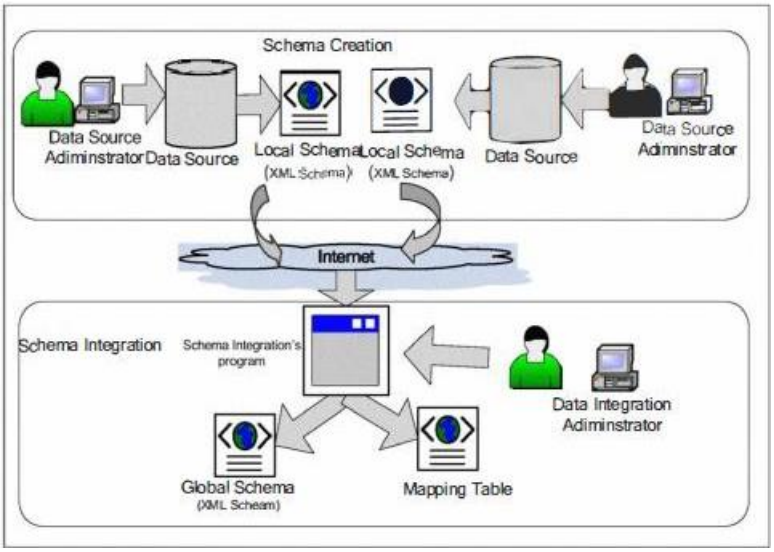

Figure 1: Schema Creation and Integration [1]

Schema Creation:
In the first process, administrator will generate the local schema described by XML schema using the tools. In this phase local XML schema is created.

Local Schema:

```
<?xml version="1.0"?>
<xsd: schema attributeFormDefault="qualified" elementFormDefault="qualified" xmlns: xsd="local host/cars">
<xsd: element name="Automobile">
<xsd: complex Type>
<xsd: sequence>
<xsd: element maxOccurs="unbounded" minOccurs="O" name="car" type="santro: car Type"/>
</xsd: sequence>
</xsd: complexType>
<xsd: key name="blue cars">
<xsd: selector xpath="santro: car"/>
<xsd: xpath xpath="car: color"/>
</xsd: key>
</xsd: element>
<xsd: complexType name="carType">
<xsd: sequence>
</xsd: sequence>
</xsd: complexType>
</xsd: schema
```

The above code is a sample example for car searching for a Santro with blue color. This is the local schema.

Global Schema:

```
<?xml version="1.0" encoding="UTF-S"?>
< xmlns: xsd="localhost/~car">
<xsd: element name="database">
<xsd: complexType>
<xsd: sequence>
<xsd: element maxOccurs="unbounded" minOccurs="O" ref="Automobile"/>
</xsd: sequence>
</xsd: complexType>
</xsd: element>
<xsd: element name="Automobile>
<xsd: complexType>
<xsd: sequence>
<xsd: element name="car">
<xsd: complexType>
<xsd: sequence>
<xsd: element name="vehicle type" type="xsd: int"/>
<xsd: element name="year" type="xsd: string"/>
<xsd: element name="tolerance" type="xsd: string"/>
</xsd: sequence>
</xsd: complexType>
</xsd: element>
</xsd: sequence>
</xsd: complexType>
</xsd: element>
</xsd: schema> Global Schema
```

The above code is an example for car searching for a Santro with global variables year, tolerance and vehicle type. This is the Global schema.

Schema Integration
Second process is the integrated search using the defined ontology. In generally administrator will make the mappings between the local schema and the global schema.

When two schemas are integrated, the data integration process is binary whereas when more than two schemas are integrated; it is an n-array [6]. Again, binary process is classified into two types, they are:

1) Ladder Style: In this style two schemas can be integrated at a time. It can also integrate the old schema with a new schema.
2) Balanced Style: In this style until the global schema is created, data integration of initial schemas is done in pair.

N-array is classified into two types, they are:

1) One shot style: In this type all the schemas are integrated at the same time.
2) Iterative style: It picks group of arbitrary schemas and integrate them. It is similar to ladder fashion in binary array.

This paper is narrowed down to one shot style where all the

schemas are integrated immediately. Here in XML schema, data must be updated in regular intervals. In this process, initial pre integration must be done and it is classified into three steps. They are: the first step is, obtaining of data type definitions, attributes and elements by parsing the xml file. The second step is elements that are found in the previous step are compared as well as merged. In the final step, the merged schemas are transformed in to a global xml schema.

Node list must be produced during the data integration of heterogeneous data. This node list is implemented in arrays in Java. Each local schema file is represented by each node in the node list. Each intermediate node will have information of its child node. Based on its node type, child integration policy is done and it is done starting from the root node. This method is continued till all the terminal nodes are integrated.

Figure 2 shows the xml-based integration, where xml document is designed with a Global schema and that is connected to a local schema and the local schema is connected different data sources. The data is retrieved from the data sources. Initially, query is sent from global schema to local schema and the query is redirected to the data source. The result obtained in the data source is sent to the local schema and the obtained result is finally received at the global schema end. Finally, data from different local sources are extracted at the Global XML schema. For integration using xml we use apache Xercers parser and Java Document Object Model (JDOM) technology.

Initially query is sent to local schema and then query is processed to next stage data source. The result obtained is sent to local schema and the result finally arrives to Global schema.

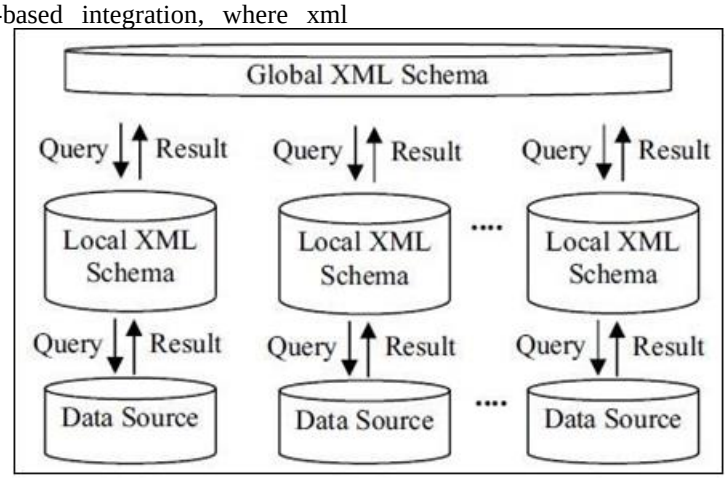

Figure 2: Integration of XML Schemas [3]

Proposal:
In one shot style all schemas are integrated at the same time but, it takes a plenty of time. The proposed idea of this paper is to design schema integration which not only integrates simultaneously but also in a quick time.

3. Conclusion

By analyzing and researching the methodologies this framework the heterogeneous data Integration is solved by the global XML schema which maps local XML schema to show an integrated view of massive sources.

References

[1] Chong-Shan Ran; Ma-Chuan Wang, "An XML Schema-based data integration, " *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol.7, no., pp.100-102, 9-11 July 2010
[2] Dongkwang Kim, Karpjoo Jeong, et al. An XML Schema-based Semantic Data Integration. IEEE.2006
[3] Xiong Fengguang; Han Xie; Kuang Liqun; , "Research and implementation of heterogeneous data integration based on XML, " *Electronic Measurement & Instruments, 2009. ICEMI '09.9th International Conference on*, vol., no., pp.4-711-4-715, 16-19 Aug.2009
[4] Yongzheng Lin; , "Study and technological realization about heterogeneous data integration based on XML Schema, " *Test and Measurement, 2009. ICTM '09. International Conference on*, vol.2, no., pp.394-397, 5-6 Dec.2009
[5] Madhavan and A. Halevy. Composing mapping among data sources. In VVLDB, 2003
[6] Sanjay Madria, Kalpdrum Passi, Sourav Bhowmick. An XML Schema integration and query mechanism system, Data & Knowledge Engineering 2008.