CO-XML/DW: CONVERTING XML SOURCES INTO DATA WAREHOUSE DESIGN

[1]THIDAR WIN, [2]KHIN NWENI TUN

University of Computer Studies, Yangon
E-mail: thidarwinster@gmail.com, knntun@gmail.com

Abstract- XML (eXtensible Markup Language) and schema language are data exchange format in web based system. Many organizations widely use XML for e-commerce and internet based applications. On the other hand, data warehouse maintains historical data and analyze data which is used by organizations to enhance decision making on the business process. XML data was insufficient to analyze on the Web. So it is required to convert a given XML schema into data warehouse schemas and to systematically analyze. This paper proposes a hierarchical design framework conversion of XML/DW to provide a suitable multidimensional data model for relational online analytical processing (ROLAP). We present an automatic approach to support this schema conversion process based on the source of data that are XML schema and conforming XML document. We get more than one data warehouse schemas from the given XML schema using Schema Graph and Extended Entity-Relationship Model. The conversion algorithm comes into being where both correct structures and hidden integrity constraints are translated.

Index Terms- XML, XML Schema, Data Warehouse, Star, Snowflake, Fact Constellation Schema, ROLAP, Schema Graph, Extended Entity-Relationship.

I. INTRODUCTION

In recent years, the web has been developing incredibly and has become main information interchange among various organizations. XML is also emerging as a worldwide standard for representing and exchanging data among applications over the web. Number of organizations are beginning to store data in XML format and using XML as an intermediate format for publication. XML data is associated with Document Type Definition (DTD) or XML schema [6].

DTD which describes precisely what elements could appear as document and what the contents of the elements and attributes are. However, DTD does not have any built-in data types, user-derived data types and allows only limited control over cardinality. Nevertheless, XML schema is more powerful to represent XML document structure and overcome the limitations of DTD. XML schema could work as a fact repository [6].

Data warehouse is subject-oriented, non-volatile, integrated, time variant collection of data which helps in developing strategic decision. Data warehouse also store huge amount of information about multiple data sources which are used for analysis. Data warehouse is based on multidimensional model which is presented by the fact table and the dimension tables. The association between these tables are generally represented through three data warehouse schemas namely star schema, snowflake schema and fact constellation [7].

The increasing use of the web and the requirement of integration data from different data sources furthered the research in conversion of XML/DW. In this paper, we propose how to perform schema exchange tasks and how data warehouse schemas are generated from the conceptual model.

This paper is organized as follows. Section 4 proposes hierarchical design framework to convert XML/DW. Section 5 presents experiment result on multidimensional schemas quality. Finally, section 6 concludes this paper.

II. RELATED WORKS

This section discusses some related works of conversion XML schema into data warehouse schema. Many articles propose different conceptual models for XML design, but there is no standard format for conceptual design of XML data. S.Sen, R.Ghosh, D.Paul, N.Chaki proposed "Integrating Related XML Data into Multiple Data Warehouse Schemas". The paper [1] proposed Schema Graph with level wise approach but this graph which less semantic constraint for XML schema. However all these generated XML schema is converted to star, snowflake and fact constellation schema in multidimensional structure. But it cannot define structure of XML schema and conforming XML documents.

S.Dasgupta, S.Sen and N.Chaki proposed "A Framework to Convert XML Schema to ROLAP". The paper [2] converted star schema or snowflake schema from the given single XML schema. Entity-Relationship is not a new formalism for conceptual modeling, but an adapted mix of the tree and ER support semi-structured data representation as hierarchy. The tree did not consider XML schema which more than one root the element. Fact table and dimension tables have identified from ER diagram to provide a suitable multidimensional data model.

Fact constellation schema could not be identified because the connections between different root elements are out of knowledge. The possible sharing of the dimension is out of scale. This leads to failure of identifying some of the business processes which are

actually related. ER lack the ability to express the semantic and special characteristics of XML features such as complex structure elements, ordering, schema-less content and mixed content.

Parimala N. and P.Pahwa [3] proposed "From XML Schema to Cube". The paper described a semi-automatic approach for designing multiple cubes of multidimensional models from the XML schema. An attribute tree has arrived by parsing the XML schema. Some nodes in the attribute tree may not carry additional information. Grafting and pruning is used to remove these unnecessary nodes. Facts, dimensions and measures, which are used to describe the multidimensional cubes, are identified. The choice construct of the XML schema is shown to give multiple cubes in both the cases. It will not be a quite a correct relationship between the fact instance and the dimension instances. The system did not consider extensions of types and the cardinalities in depth.

III. THEORY BACKGROUND

In this section we are presented different terminologies which have been used in this paper.

A. XML

XML has been accepted as a data exchange format with the rapid development of web services. XML allows the flexible development of user-defined document type and makes easier to transfer to database system. Exchanging data as XML greatly reduces time complexity, since the data can be read by different incompatible applications. The content of XML document is encapsulated within elements that are defined by tags. Tags may contain additional descriptive information, namely attributes. Elements are hierarchically organized in a tree-like structure.

```
<?xml version="1.0" encoding="utf-8" ?>
<BookCatalogue xmlns=http://www.publishing.org
mln:xsi"http://www.w3.org/2001/XMLSchema-instance"xsi:schema
Location=BookCatalogue.xsd">
<Book>
   <Title>Yogasana Vijnana: the Science of Yoga</Title>
   <Author>Dhirendra Brahmachari</Author>
   <Date>1966</Date>
   <ISBN>81-40-34319-4</ISBN>
   <Publisher>Dhirendra Yoga Publications</Publisher>
   <Cost currency="INR">11.50</Cost>
</Book>
<Book>
   <Title>The First and Last Freedom</Title>
   <Author>J. Krishnamurti</Author>
   <Date>1954</Date>
   <ISBN>0-06-064831-7</ISBN>
   <Publisher>Harper &amp;amp; Row</Publisher>
   <Cost currency="USD">2.95</Cost>
</Book>
</BookCatalogue>
```

Fig. 1. XML Document of BookCatalogue

B. XML Schema

XML schema typically expresses in terms of constraints on the structure and content of XML documents. XML schema file provides (i) elements (ii) attributes (iii) data types and the ability to create additional types and (iv)relationships between elements [11]. The basic syntactical constraints are imposed on XML itself. There are two types of XML schema design. Salami design corresponds to having all of the elements defined within the global namespace and then referencing the elements. Russian Doll design contains only one single global element. All the other elements are local. Only the root element defines within the global namespace [1].

C. Conceptual Models

a) Schema Graph: In a graph, data structure is identified and data model is used to describe conceptual data models by providing graphical notations. Schema Graph shows that the relationships of elements and enable users to fully see the relationships between each elements.

b) Extended Entity-Relationship Model: has become the most commonly used expression tool for conceptual model. EER is a high-level incorporating extension to the original ER model, used in the design of databases. Additionally it includes the concepts of a subclass and superclass, along with the concepts of specialization and generalization, the concept of a union type of objects of different entity types. EER is the ability to express the integrity constraints, semantic constraints and special characteristics of XML.

D. Data Warehouse Schemas

Data warehouse allows XML schema to be processed analytically based on ROLAP. The popular way of describing data warehouse is multidimensional model that is organized by fact table and dimension tables. Fact Table contains measures, values that are to be analyzed and keys are related dimensions tables. Dimension tables represent the perspectives of analyzing data using hierarchically organized dimension attributes. The relationships between these tables are generally represented through data warehouse schemas. If XML schema is found as disconnected, more than one data warehouse schema is identified [1].

a) Star schema: a fact table in the middle which is connected to a set of dimension tables. This schema is simplest schema, easy to understand and easy to navigate between the tables due to less number of joins and most suitable for query processing.

b) Snowflake schema: a refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables. It is easier to maintain and saves the storage space.

c) Fact Constellation schema: multiple fact tables that share dimension tables, viewed as a collection of stars, called galaxy schema.

IV. FRAMEWORK TO CONVERT XML/DW

In schema conversion, a source schema s1 is given and the goal is to find a target schema t1. T1 is equivalent to s1, while they belong to different models in the schema integration (e.g., data warehouse and XML). Integrating schema between different models has been extensively investigated. The conversion framework of XML/DW is shown in Fig. 2.

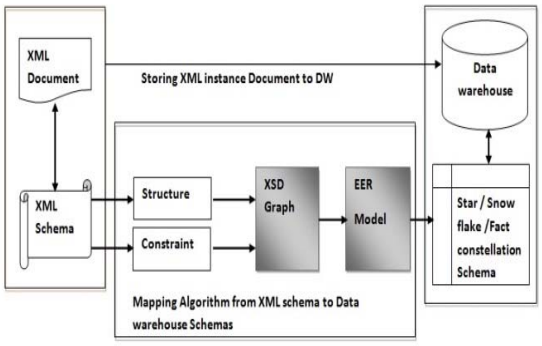

Fig.2. Overview of Framework for Co-XML/DW

XML schema for Russian Doll and Salami design are used as input. We construct Schema Graph as a conceptual model for the conversion process that data could be extracted from the XML schema according to the proposed methodology and these data is transformed into EER schema under the conversion rule. Finally, data warehouse schemas which could be identified from EER schema.

E. Example of XML Schema

XML schema of the given problem definition is shown in Fig.3. At first the proposed XML schema is applied on it to generate the schema graph.

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified"
elementFormDefault="qualified" targetNamespace
="http://www.publishing.org"  xmlns:xs
="http://www.w3.org/2001/XMLSchema">
<xs:element name="BookCatalogue">
<xs:complexType>
<xs:sequence>
<xs:element name="Book"
maxOccurs="unbounded"minOccurs="0">
<xs:complexType>
<xs:sequence>
<xs:element type="xs:string" name="Title"/>
  <xs:element type="xs:string" name="Author"/>
  <xs:element type="xs:short" name="Date"/>
  <xs:element type="xs:string" name="ISBN"/>
  <xs:element type="xs:string" name="Publisher"/>
<xs:element name="Cost" maxOccurs="unbounded"
minOccurs="0">
<xs:complexType>
<xs:simpleContent>
   <xs:extension base="xs:float">
   <xs:attribute type="xs:string" name="currency"
use="optional"/>
</xs:extension>
</xs:simpleContent>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>
```

Fig. 3. XML Schema of BookCatalogue

F. Properties of Schema Graph

Schema Graph allows the entire semi-structured to be viewed as a Graph (V, E) in level organization.

a) Schema Graph represents hierarchical structure as vertices correspond to elements, attributes in XML schema.

b) The name of vertices of Schema Graph would be same as the name of elements of XML schema.

c) It is a level wise separable graph. Elements and attributes are not distinguished in this graph.

Procedure of XML Schema-to-Schema Graph

Begin

Input: XSD: XML schema file with {'E' = {E1, E2,… En} set of elements, 'A' = {A1,A2,..Am} where A is set of attributes in XSD}

Output: G: Schema Graph

1. Find out 'E' in XSD file
2. If XSD has one 'E' that does not have the predecessor, then
3. define 'E' as root vertex in the Level-1 of G.
4. Repeat step 5 to 13 for all 'E' in XSD starting from root 'E'.
5. If 'E' is simple type, then
6. map as child of parent node.
7. If 'E' is complex type which has sequence of 'E' within the same type, then
8. 'E' will become the parent of sub elements as level-2 in G.
9. If sub element is complex type, then
10. sequence of 'E' within the same type will become top element at the next level and so on.
11. If E has 'A', then
12. it would be mapped as leaf node of current 'E'.
13. Cardinalities (min and max Occurrence fields in the XML schema) are also mapped as properties of Schema Graph construct.
14. End

XML schema is converted to Schema Graph by applying propose method "XML Schema-to-Schema Graph". The element BookCatalogue has no predecessors. It would be identified as root vertex. BookCatalogue have one sub element Book. The attributes of the every element are connected with corresponding elements. Book would be connected

with title, author, date, ISBN,Cost, publisher. Book have only sub element Cost. Cost have one attribute namely currency. XML schema of Fig. 3 is converted to Schema Graph as shown in Fig.4.

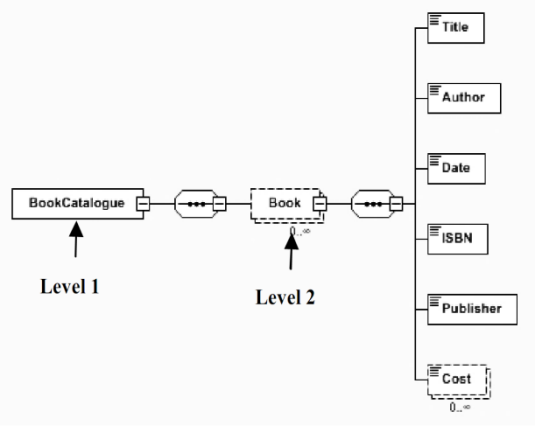

Fig. 4. Schema Graph of Book Catalogue

G. Rules for converting XML Schema Structure to EER Model

The selected root element and relevant elements of Schema Graph which can be mapped into entities of EER model.

a) An element with simple data type such as integer becomes an attribute of parent entity if it is a part of a sequence within a complex element.

b) If the simple element is a standalone element then it's modeled as the attribute of the root entity.

c) A set of attribute production rule of Schema Graph are converted to entities' attributes while converting XML schema into EER schema. Data types are not displayed in EER.

d) An element with complex type can be one of the empty, elements only and mixed content. The rules for converting mixed content and elements only types are similar to the EER entity. The empty element is modeled as an empty EER entity.

e) Build up relationships between entities according to the set of element production rules of Schema Graph (e.g. rule expressing type assignments, parent-child relationships etc.).

f) A subclass with more than one super class is called a shared subclass. In case of multiple inheritances, attributes are creating in the same superclass inherited more than once via different paths.

g) EER model takes care to avoid element redundancy in XML logical schema and to keep the connections among concepts of EER schema.

Procedure of Schema Graph-to-EER

Begin Input: G: Schema Graph with {'Rx' root element, 'E'= {E1,E2,…En}set of elements, 'A' ={A1,A2,…Am}' set of attributes}

Output: EER with {'Re' root entity, 'A' = {A1, A2,…Am}' set of attributes}

1. THE 'RX' IN G IS CONSIDERED AS THE 'RE' IN EER MODEL.
2. REPEAT STEP 3 TO 14 FOR BUILDING ALL ENTITIES FROM 'RE' TO ALL LEAF NODE.
3. DETERMINE ALL NODES UNDER 'RE' NODE IN G.
4. If it is 'A', then

(i) It is represented as 'A' of parent entity in EER schema.

(ii) The 'A' has a unique parent entity.

(iii) The 'A' processes no sub-entity.

5. Else if it is 'E' with simple type , then
6. It begin 'A' of parent entity.
7. Else if it is 'E' with complex type , then
8. Complex 'E' at top-level are modeled as entity.
9. Else if it is 'E', then
10. check all the siblings nodes at same level
11. If siblings nodes are E, then
12. check they consist of same 'A' or not

(i) If they consist of same 'A', then

(ii) these 'A' represent as the specialization of the parent node which is represented as relationship in EER schema.

(iii) Else the entity having relationship with its parent.

13. CARDINALITY CONSTRAINTS DEFINED MIN AND MAX NUMBER OF TIME WHICH 'E' MAY PARTICIPATE IN THE RELATIONSHIP TYPE.
14. INTEGRITY CONSTRAINTS DEFINED ALL THE ENTITIES TO ADD A PRIMARY KEY. THIS KEY IS OBTAINED BY ADDING AN ATTRIBUTE WITH THE ENTITY: NAME OF ENTITY +ID.

End

The following Fig. 5 has BookCatalogue, Book and Cost entities. Book has relationship with Cost entity. BookCatalog is considered as fact table and Book and Cost are considered as dimension tables. During the mapping, both data structures and semantic constrains are correctly preserved to data warehouse schemas.

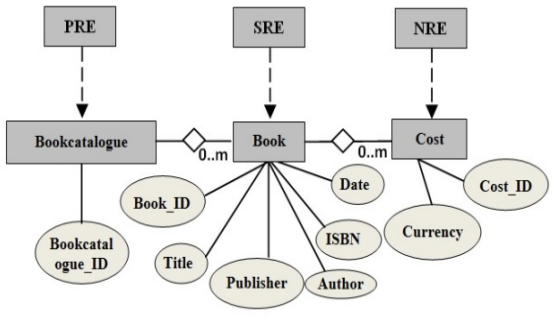

Fig. 5. EER Model of BookCatalogue

H. Identification of Fact and Dimension Tables

We define the particular entity from EER model, which corresponds to the root element of XML schema.

a) An entity has no predecessor entity. It has parent-child relationship which has one-to-many relationships which is termed as Primary Root Entity (PRE). The name of PRE entity is defined name of fact

table. Fact table is constructed by taking the primary key of PRE.
b) Primary keys of all the other entities which are connected directly to PRE as Secondary Root Entity (SRE).
c) Other entities are neither PRE nor SRE are designated as Non Root Entity (NRE).

I. Building Data Warehouse Schemas

a) Star Schema: if EER consists of PRE and SRE's only, it is identified with star schema. Fact table is named as the name of PRE + "Fact". The primary keys of SRE are placed in PRE as the foreign keys and the SRE's are acting as dimension tables in the star schema.
b) Snowflake schema: if EER contains of PRE, SRE and NRE provided that PRE's is not connected, it is recognized with snowflake schema. The primary keys of SRE are retained in PRE and SREs act as dimension tables. Similarly SREs would contain the primary key of other tables that is represented by NRE. At least one NRE should exist in SRE to be identified as snowflake schema.
c) Fact-Constellation Schema: Fact constellation schema is identified if PRE's share dimension tables of SRE.

Procedure for Snowflake Schema
Begin
Input: EER Model with {'PRE' Primary Root Entity, 'SRE 'Secondary Root Entity, 'NRE' Non Root Entity}
Output: Snowflake Schema
Read EER model
1. If PRE then {
Create Fact Table "PRE.Name";
Define PRE_ID;
}
2. For (i=1 to n do) {
3. PRE find SREs;
4. If SRE then {
Create Dimension Table "SRE.Name";
Define SRE_ID;
SRE find Parent_PRE;
Assign: Parent_PRE:= SRE_ID;
}
5. For (j=1 to m do) {
6. SRE find NREs;
7. If NRE then {
Create Dimension Table "NRE.Name";
Define NRE_N_ID;
NRE_N find Parent_SRE;
Assign: Parent_SRE:= RE _N_ID;
}
8. For (k=1 to p do) {
9. NRE find next NRE
10. If NRE then {
Create Dimension Table "NRE.Name";
Define NRE _N+1_ID;
NRE _N+1 find Parent_NRE _N+1-1;
Assign: Parent_NRE_ N+1-1 : = NRE_ N+1_ID;
}
Repeat step 9
}}}
End

V. EXPERIMENTAL RESULT

XML schema allows constraints to be specified on the elements and attributes. When mapping an XML schema to data warehouse schema, XML schema constraints are mapped to appropriate constraints on the tables and columns.
a) Unique element specified the uniqueness constraint on an element or attribute using the unique element. In the process of translating an XML schema into data warehouse schema, the unique constraint in the XML schema is mapped to a unique constraint in the Table.
b) Key constraint specified using the key element that the values of the bookID child element of the books element must have unique values and cannot have null values.
c) Keyref constraint specified using the keyref element which allows relationship between elements. The element is converted to a corresponding foreign key constraint on the columns in the tables, if a schema specifies the keyref element, during the schema mapping process. The mapping process converts these schema constraints into appropriate constraints on the tables created.

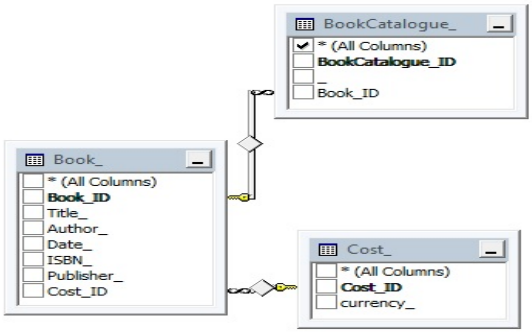

Fig.6. Snowflake Schema

This system shows the snowflake schema in Fig.6. It has defined one fact table namely BookCatalogue fact and dimension tables namely Book and Cost according to Fig.5. Fact table is constructed by taking the primary key of PRE namely BookCatalog. SRE Book which is connected with PRE directly by some relationship. NREs Cost is connected to directly with SRE Book.

J. Performance Evaluation

A.Gosain and J.Singh proposed "the quality measurement metric method "to measure the quality of output data warehouse schema [9]. Quality depends on several factors and characteristics such as functionality, reliability, usability, understandability, correctness, readability, simplicity.
The proposed method reduces XML components by using specialization and generalization on EER. The

complexity of the generated dimension is smaller than dimension without using specialization.

The values of metric are shown in Table 1. The metric definition is only one step in the complete process of obtaining a set of valid and useful metrics. We tried to obtain a set of metrics for measuring the complexity of data warehouse models that help designers to improve the quality of data warehouse, thereby affecting understandability of this model. This metric can be applied schema level for multidimensional model.

Table 1. Quality measurement metric value

Quality Metric	Meaning	Value
NFT	Number of fact tables	1
NDT	Number of dimension tables	2
NSDT	Number of shared dimension tables	0
NAFT	Number of attributes and foreign keys of fact tables	2
NADT	Number of attributes and foreign keys of dimension tables	9
NASDT	Number of attributes of shared dimension tables	0
NHD	Number of hierarchy dimensions	1
RScA	Ration of Schema Attributes	9/2

CONCLUSION

Co-XML/DW was provided conceptual presentation for XML. This research was proposed a hierarchical design framework to convert XML/DW. As a result, we got multiple data warehouse schema definitions from the related XML schema. We were proposed an automatic approach to support this conversion processes and our approach could be applied to semi-structured data to convert them suitable structured data in data warehouse. We have implemented to reduce conversion time of schema and to enhance decision making for the most of the business organization. Organization must find ways to efficiently analyze and manipulate XML data within their data warehouse and also to take the advantage of using web environment.

REFERENCES

[1] Soumya, S., Ranak G., Debanjali, P. and Nabendu, C. ”Integrating Related XML Data into Multiple Data Warehouse Schemas” University of Calcutta Kolkata -700009 West Bengal, India, JSE-2012, CS & IT 04, pp. 357–367.

[2] S.Dasgupta, S.Sen and N.Chaki, "A Framework To Convert XML Schema to ROLAP"; Proc. of the Second International Conference on Emerging Applications of Information Technology (EAIT 2011), Kolkata, India, 2011 ISBN : 978-1-4244-9683-9.

[3] Parimala N and Payel, P., “From XML schema to cube” International Journal of Computer Theory and Engineering; Vol. 1, No 3 August 2009.

[4] Arijit, S. and Erik, W.2006. “The case for conceptual modeling for xml” , Technical report, Computer Engineering and Networks Laboratory, ETH Zurich.

[5] Haitao, C. and Husheng, L., “A Survey to Conceptual Modeling for XML”, College of Computer Sciences Beijing University of Technology Beijing, China, 978-1-4244-5539-3/©2010 IEEE.

[6] M. Jensen, T. Miller, and T.B. Pedersen, “Specifying OLAP Cubes on XML Data”, Journal of Intelligent Information Systems, 2001.

[7] Soumya, S., Ranak G., Debanjali, P. and Nabendu, C. 2012, “Integrating Related XML Data into Multiple ROLAP Warehouse Schemas”, International Journal of Software Engineering & Applications (IJSEA).

[8] Gunjan, C., “Designing of Star Schemas from XML Document”, International Journal of Multidisciplinary and Current Research, 2014.

[9] Matteo, G., Stefano, R. and Boris, V.,”Data warehouse design from XML sources”, Proceedings of the 4th ACM international workshop, 2001.

[10] A.Gosain and J.Singh “Quality Metrics for Data Warehouse Multidimensional Models with Focus on Dimension Hierarchy Sharing” Advances in Intelligent Systems and Computing 320, DOI: 10.1007/978-3-319-11218-3_39 Springer International Publishing Switzerland 2015.

[11] G. Chandwani, “Designing of Star Schemas from XML Document”, Vol.2 13 January 2014.

★★★