Global Journal of Computer Science and Technology
Volume 11 Issue 13 Version 1.0 August 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Transformation of Flat File into Data Warehouse

By Muhammad Inayat Ullah, Muhammad Zeeshan, Mahwish Kundi

Gomal University,Dera Ismail Khan, Pakistan

Abstract - A Flat file (Semi Structured) Data comes from different sources or operational systems for storage in the data warehouse. Extraction, transformation and loading of the data could be necessary. Moreover, input flat file data must be transformed into a uniform format which could be more suitable for analytical purposes. Aim of this research is to analysis the delimiters of the flat file, to transform flat file into uniform format and suggest a suitable algorithm for implementation such type of algorithm could be solve the problem of transformation of the flat file data and such algorithm could be useful for extraction, transformation and loading of huge amount of flat file data into data warehouse.

GJCST Classification : *H.2.7*

Transformation of Flat File into Data Warehouse

Muhammad Inayat Ullah[α], Muhammad Zeeshan[Ω], Mahwish Kundi[β]

Abstract - A Flat file (Semi Structured) Data comes from different sources or operational systems for storage in the data warehouse. Extraction, transformation and loading of the data could be necessary. Moreover, input flat file data must be transformed into a uniform format which could be more suitable for analytical purposes. Aim of this research is to analysis the delimiters of the flat file, to transform flat file into uniform format and suggest a suitable algorithm for implementation such type of algorithm could be solve the problem of transformation of the flat file data and such algorithm could be useful for extraction, transformation and loading of huge amount of flat file data into data warehouse.

I. Introduction

A flat file is also called plain text file it is semi structured file. A flat file consisted of data which is separated with commas, white spaces, tabs, tube (|) any many other characters.

Al-Dubaee et al. (2010) stated that a flat file is a plain text or mixed text and binary file which usually contains one record per line or physical record (example on disc or tape). Within such a record, the single fields can be separated by delimiters, e.g. commas, or have a fixed length. In the latter case, padding may be needed to achieve this length. Extra formatting may be needed to avoid delimiter collision. There are no structural relationships between the records.

Mathew (2005) stated that a flat file database should consist of nothing but data and, if records vary in length, delimiters. More broadly, the term refers to any database which exists in a single file in the form of rows and columns, with no relationships or links between records and fields except the table structure.

Pratnortis (2005) stated that the advantage of a flat file is that it takes up less space than a structured file. However, it requires the application to have knowledge of how the data is organized within the file. By using Structure Query language and a database (rather than a collection of files in a file system), a user or an application is free from having to understand the location and layout of data (for example, the length of each item of data, its type of data, and its relationship to other data items). Another form of flat file is one in which table data is gathered in lines of ASCII text with the value from each table cell separated by a comma and each row represented with a new line. This type of flat file is also known as a comma-separated values file.

II. Different Terms Related to Research

a) Flat file

Flat file is a single lined plain text file containing delimiters comma, tabs, tube (|) and many other characters. It is also called mixed text file. The main advantage of the flat file is that it takes less space as compare to the structured file.

b) Data warehouse

Data ware house is a large data base which contains historical data in the form of cubes (Aggregates) and decision maker use this data for future decisions.

III. Steps for Solving the Problem

The main steps in conducting the research work are given as under:

1. Read and review the research paper related to the data warehousing, flat file (Semi structured) data.
2. Review the research critically.
3. Design the effective algorithm to transform the flat file (semi structured) Data into structured data.
4. Implement the algorithm and check output.

IV. Solution to the Problem and Methodology

Steps for the solution to mentioned problem are:

1) Take a flat file.
2) With the help of proposed algorithm convert it into html form.
3) Import html data into the database.
4) Manage (updation/deletion/insertion) the flat file data in the data warehouse.

V. Algorithm

```
Input : Flat file containing single line plain text data.
Output : Data transform into uniform format.
Begin
    1.  Enter Input flat file
    2.  Flat file transform to uniform format
    3.  Transform uniform data to html form
    4.  If flat file is transformed as a html then
    5.  Import html data into database.
    6.  Extract stored data from the database.
    7.  Manage (updation/deletion/insertion) of the flat
        file data in the data warehouse.
End
```

Author [α] : Muhammad Inayat ullah, University of Engineering & Technology, Peshawar, pakistan
Telephone : +923339977807, E-mail : inayatbinzeb@gmail.com
Author [Ω] : Zeeshan, ICMS, Hayatabad phase-V, Peshawar, Pakistan.
Telephone: +923139300775 E-mail : zeese2010@gmail.com
Author [β] : Mahwish Kundi,Gomal University,Dera Ismail Khan, Pakistan.
Telephone: +923459774168 E-mail : kundikhan84@gmail.com

VI. Implementation of Algorithm

In the first step take two flat files first file is containing the following contents with following delimiters. Model for transformation of flat file (Semi Structured data file) is show in the figure 1.

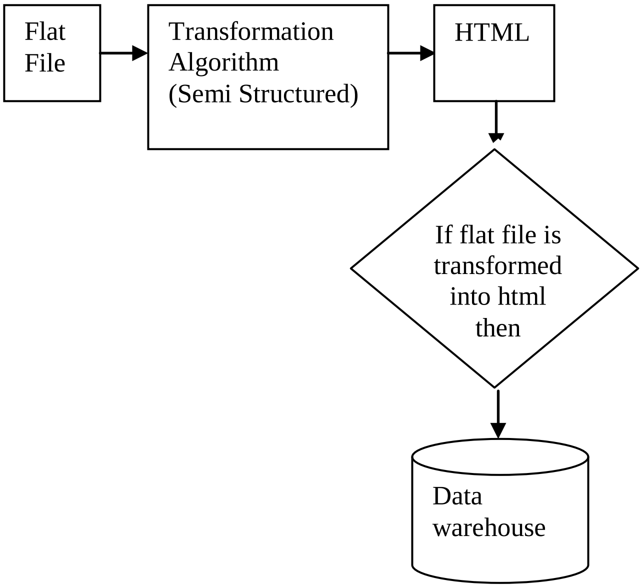

Flat File
Transformation Algorithm (Semi Structured)
HTML
If flat file is transformed into html then
Data warehouse

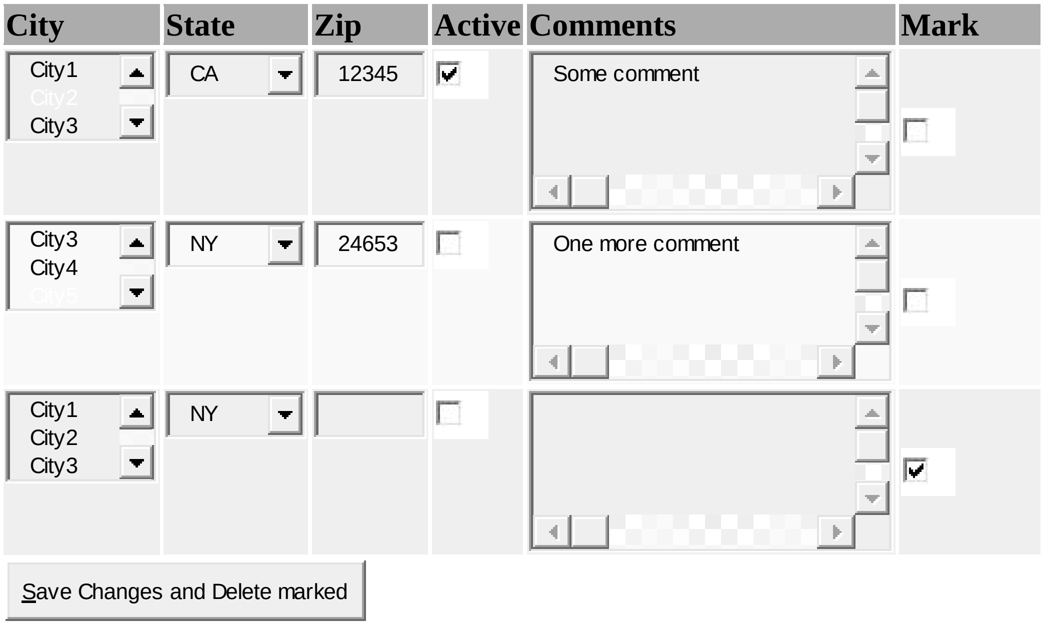
Figure.1 : Transformation Model of Flat File

```
City,LIST,3,City1:City2:City3:City4:City5
 State,LIST,1,NY:CA:LA
 Zip,STRING,8
 Active,LOGICAL,1,Y:N
 Comments,TEXT,30:2
```

Above text file is separated with colon and semi colon delimiters. And the second flat file is containing the following contents with following delimiters.

```
City2|CA|12345|Y|Some comment
City5|NY|24653|N|One more comment
```

Above text file is separated with tube delimiter.In the second step an algorithm is used for the transformation of the flat file into the html format. Flat file transformation into html format is shown bellow.



City	State	Zip	Active	Comments	Mark
City1 City2 City3	CA	12345	☑	Some comment	□
City3 City4 City5	NY	24653	□	One more comment	□
City1 City2 City3	NY		□		☑

Figure. 2 : Transformation of the proposed algorithm in html format

Output of the algorithm is in html format so with the help of suitable query html data is transformed into structured data. When data is completely transformed into structured format and then management of the flat file (updatation/deletion/insertion) is easy. In this research PHP and MySQL are used for the implementation of the proposed algorithm.

VII. Conclusion

In data warehouse data comes from different operational systems and flat file is one of the semi structured plain text file that comes form source and to store it in data warehouse flat file doesn't store directly in the data warehouse it need some transformation in

ETL (Extract, Transform and load) of the data warehouse which is the core component of the data warehouse. In this paper a flat file containing delimiters comma, colon and tube (|) characters and an algorithm is used for transformation of the flat file into uniform format means in html form and then transformed in to structured data. Flat file management (updation/deletion/insertion) is easy when transform it into structured format. Proposed algorithm is implemented in PHP and MySQL tools.

References References Referencias

1. Al-Dubaee, S. A. and Ahmad, N. 2010. "Multilingual Lossy Text Compression Using Wavelet Transform". First International Conference on Integrated Intelligent Computing. (ICIIC). pp. 39-44.
2. Bettina Berendt, Bamshad Mobasher, Miki Nakagawa, and Myra Spiliopoulou, The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis, Proceeding of the WEBKDD 2002 Workshop, Edmonton, Canada.
3. Cortes, K. Fisher, D. Pregibon, A. Rogers, and F. Smith. Hancock, A Language for Extracting Signatures from Data Streams, in: Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining, 2000, pp. 9–17.
4. Eliashberg, Jehoshua, Jedid J. Jonker, Mohanbir S. Sawhney, and Bernard Wierenga (2000), "MOVIEMOD:An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures," MarketingScience, 19 (3), 226-43.
5. Godes, David, Dina Mayzlin, Yubo Chen, Sanjiv Das, Chrysanthos Dellarocas, Bruce Pfeiffer, Barak Libai,Subrata Sen, Mengze Shi, and Peeter Verlegh (2005), "The Firm's Management of Social Interactions,"Marketing Letters, 16 (3), 415-28.
6. Liu, Yong (2006), “Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue,” Journal of Marketing, 70 (3).
7. Mayzlin, Dina (2006), "Promotional Chat on the Internet," Marketing Science, 25 (2), 155-63.
8. McCallum, Andrew and Ben Wellner (2005), "Conditional models of identity uncertainty with application to noun co reference," Advances in Neural Information Processing Systems, 17, 905-12.
9. http://www.mysql.com (retrieved on May 18, 2011)
10. http://www.php.net (retrieved on May18, 2011).