Original Research Paper Computer Science

Data Transformation Complexity and Proposed Solution

Mitul Tripathi Student, Symbiosis Institute of Computer Studies and Research (SICSR), A Constitute of Symbiosis International University, Pune-411016, Maharashtra, INDIA

Dr. Pravin Metkewar Associate Professor, Symbiosis Institute of Computer Studies and Research (SICSR), A Constitute of Symbiosis International University Pune- 411016, Maharashtra, INDIA

ABSTRACT

In Today's Scenario, as lot of complexities emerge in business, as different approaches companies or individual are adopting to store the data in data warehouse as the approach is different it leads to lot of problem while integrating these different type of data and bring them into common platform, So that they can be on same page and can be understood and evaluated. Transformation includes various mathematical algorithms and process to transform the data into some predefined structure or format. Here the Integration part comes. Integration, is a process of amalgamation of two structure following common attribute.

Before Moving to Actual Transformation we need to know, what is the Architecture of Data-warehouse, how data get transformed? Data get fetched from different data sources which is situated in different location and store into the staging area which is also known as ETL layer where all the logic and algorithm is implemented on data and all the transformed data get store in the Operation Data-Source. There can be two different scenario. First, Data which is store in Operational Data Source, Data mart will be created based on this through ETL and then through Data Mart, Data Warehouse should be created.

Here I am going to propose solution for removing processing time and reduce complexity while transforming data from one base language/ format to another.

KEYWORDS : Data migration, data wrangling, PBD, FIPS, SWYN, KARMA

1. Introduction

Data Transformation means identify the base language of data from its source and get it to convert into language which is required into the remote /end user system. We need to conversion of program from one computer language to another computer language in order to process requisite data. In Real life scenario we have seen lot of example. For example a +

typical Microsoft office product which run on windows OS, to read the data having format of xls ,doc,ppt,accdb etc in android OS we need to program same product in android platform to read and write same type of file. Lot of tools available which can convert java code which is Object oriented language to VB code which is Event Driven or BASIC language. But Data Transformation and Application migration is totally different aspects.

1.1 Transforming Data

A variety of data transformation may be needed throughout data wrangling, which includes reformatting, extraction, outlier correction, type conversion, and schema mapping.

1.2 Data formatting, extraction, and conversion

The challenge with data wrangling is to prioritize data and split it for further diagnose and integration. Here we have FIPS1(Federal Information Processing Standards) this standards used in integration of tables.

One more method PBD (Programing by demonstration) assist in specific cleaning tasks. PBD converts data according to user's desired format via direct selection. Likewise SWYN helps to select data and represent preview of data after its transformation. Others method and concepts also exist such as Potluck, Karma, Vegemite which play vital role in extraction and transformation of data. The most famous is Ajax which comes in hand with data transformation and entity resolution. These tools enables data variety of data transformation, which further extends to data reshaping and data reformatting.

Data transformation sometimes not face only technical issues, sometimes semantic issues also induce complexities, such as currency convertor, if we are lacking data of current value of currency , though it can be converted but it would be converted in old value which doesn't have any meaning. Likewise data wrangling also face in conversion complexities in case of converting zip code into longitude and latitude centroids.

Sometimes it is observed during transformation phase due to large amount of data in same staging area, cause to data crash or leads to missing of valuable of data. It must ensure that objective of data conversion. Why it is converted in particular format. How to optimize system during data conversion. Because so many data leads to slow down the system, and consume so much CPU. If we talk about parallel transformation, it's fast but leads complexity and huge effort to code and generate architecture.

1.3 Fixing Erroneous values

After data transformation sometimes it is found a large gap between expected output and the output given after transformation, a typical case of crime records if we take outlier detection of crime records where standard deviation is greater than 3 from the mean which ideally is not preferable. As it is weird output analyst has two option first one he should accept such scenario because of some sudden enhancement of crime records or he has to set limit on standard deviation, he has to do this process in iterative modem, so that he comes into some conclusion.

One example is Google Refine,2 which leverages freebase 2 to enable entity resolution and discrepancy detection. Another example is D-Dupe system,3 which helps users to perform entity resolution. Human input is used to improve the system's suggestions via active learning.

2. Proposed Solution

We can transform data with two method first one include extraction of data, data mapping, load and verification, and program which will transform data, these code generation program is used to write middle layer application such as java and xslt. This solution surely would impact on not only transformation but also minimize the RAM consumption as in this model a virtual data base exist which hold clean data coming from staging area, which will be reside in virtual database in chunked form. It would be Application responsibility to get chunk of particular data format integrated and convert these into desire format but it would be priority wise and have time set to each transformation.

This architecture would include all the phase where transformation program will play indispensable role where all the request of data from virtual database would be taken via thread basis (thread is instance of processor which divided in sub process responsible to accomplish their respective jobs in respective limit). And transformation would be done accordingly. Each thread would be allotted time to transform data in respective format if they fail. Thread would get interrupted and thread having second priority would be active to perform their job. Once it would accomplish it task token would be allocated to previous thread i.e. thread 1. It not only perform job better, in prompt way but also leads to no data loss, during transformation.

Data Transformation model

3. Conclusion

Proposed solution which is defined in the solution section will handle the data transformation complexity by using using java thread, where each thread get assigned each unformatted data to be transformed in respective format, each thread would be responsible to fetch and handle the data from different data source and make the raw or unformatted data available in hidden layer(middle layer) which is process layer, where each data processed and get transformed and again respective thread will be allocated to carry processed data and load into the data warehouse.

During the process, session would be allocated to each thread, when thread will be executed, its session will start, by the time this thread is working on particular block of data warehouse, no other thread will have access to that block. Once thread job is over that block again unlock to load the formatted data by another thread.

Data transformations are the application of a mathematical modification to the values of a variable. There are a great variety of possible data transformations, from adding constants to multiplying, squaring or raising to a power, converting to logarithmic scales, inverting and reflecting, taking the square root of the values, and even applying trigonometric transformations such as sine wave transformations. The goal of this paper is to begin a discussion of some of the issues involved in data transformation as an aid to researchers who do not have extensive mathematical backgrounds, or who have not had extensive exposure to this issue before, particularly focusing on the use of data transformation for normalization of variables

References:

[1]. kandel et al, sagepub.co.uk

[2]. Roberston GG, Czerwinski Mp and Churchill JE visualization of mappings between schemas.

[3]. Kang H, Getoor L, shneiderman B, Bilgic M and Licamele L, Interactive entity resolution in relational data: A visual analytic tool and its evaluation. IEEE trans visual comput Graph 2008 14:999-1014;