IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN : 2278-0661, p-ISSN : 2278-8727
PP 67-71
www.iosrjournals.org

Data Warehouse with Data Integration: Problems and Solution

Prof. Sunila Shivtare[1], Prof. Pranjali Shelar[2]
[1](Computer Science, Savitribai Phule University of Pune, India)
[2](Computer Science, Savitribai Phule University of Pune, India)

Abstract: *In the recent years the data is regularity added. This is large challenge for data warehouse. Because large data is collected from different sources & in different structure. Heterogeneous data is managed to converting this different data into a single unified structure. These data must contain historical, current & some real time values from different & most likely from heterogeneous sources. All effective decisions are totally depend on aggregated, calculated, and time-series data values in a data warehouse. Over the period of times many researchers have contributed to the data integration issues, but no researches have collectively gathered all the causes of data integration problems. At all the phases of data warehousing along with their possible solution. Problems regarding data integration are discussed in this paper.*

Keywords: *Data Integration, Data warehouse, SQL-Server Integration Services (SSIS), ETL, T-SQL.*

I. INTRODUCTION

The Data Warehouse is collection of different data, with different structure. So there is need of data integration is important, so the different data having different structure is translated into the single unified form. This collected data is used for decision making purpose, but data integration is the challenge for warehouses. The purpose of the paper is to identify the reasons for data integration problems along with their solutions.
I hoped this is helpful to implementation of warehouse to examined & analyze these issues before a data cleaning.

II. DATA WAREHOUSE

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources.

In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.[11]

This new system environment that users desperately need to obtain strategic information happens to be the new paradigm of data warehousing. Enterprises that are building data warehouses are actually building this new system environment. This new environment is kept separate from the system environment supporting the day-to-day operations. The data warehouse essentially holds the business intelligence for the enterprise to enable strategic decision making. The data warehouse is the only viable solution[3].
The data warehouse is an informational environment that

- Provides an integrated and total view of the enterprise
- Makes the enterprise's current and historical information easily available for decision making
- Makes decision-support transactions possible without hindering operational systems
- Renders the organization's information consistent
- Presents a flexible and interactive source of strategic information

III. DATA INTEGRATION

Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.[2]

Data quality is an increasingly serious issue for organizations large and small. It is central to all data integration initiatives. Before data can be used effectively in a data warehouse, or in customer relationship management, enterprise resource planning or business analytics applications, It need to be analyzed and

cleansed. Understanding the key data quality dimensions is the first step to data quality improvement. To be processable and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. Abundant attempts have been made to define data quality and to identify its dimensions. Dimensions of data quality typically include accuracy, reliability, Importance, consistency, precision, timeliness, fineness, understandability, conciseness and usefulness. Here I have under taken the quality criteria for integration by taking 6 key dimensions as

2.1 Completeness: deals with to ensure is all the requisite information available? Are some data values missing, or in an unusable state?

2.2 Consistency: Do distinct occurrences of the same data instances agree with each other or provide conflicting information. Are values consistent across data sets?

2.3 Validity: refers to the correctness and reasonableness of data

2.4 Conformity: Are there expectations that data values conform to specified formats? If so, do all the values

2.5 Accuracy: Do data objects accurately represent the "real world" values they are expected to model? Incorrect spellings of product or person names, addresses, and even untimely or not current data can impact operational and analytical applications.

2.6 Integrity: What data is missing important relationship linkages? The inability to link related records together may actually introduce duplication

IV. SQL-SERVER INTEGRATION SERVICES (SSIS)

SQL Server Integration Services (SSIS) is a component of the Microsoft SQL Server database software that can be used to perform a broad range of data migration tasks.

SSIS is a platform for data integration and workflow applications. It features a fast and flexible data warehousing tool used for data extraction, transformation, and loading (ETL). The tool may also be used to automate maintenance of SQL Server databases and updates to multidimensional cube data.

V. etl(extract, transform & load)

ETL is a process in data warehousing responsible for fetching data out of the source systems and placing it into a data warehouse. It involves the following

4.1 Extracting the data from source systems, data from different source systems is converted into one consolidated data warehouse format which is ready for transformation processing.

4.2 Transforming the data may involve the following tasks:

1	Applying business rules (so-called derivations, e.g., calculating new measures and dimensions)
2	Cleaning (e.g., mapping NULL to 0 or "Male" to "M" and "Female" to "F" etc.)
3	Filtering (e.g., selecting only certain columns to load)
4	Splitting a column into multiple columns and vice versa
5	Joining together data from multiple sources (e.g., lookup, merge)
6	Transposing rows and columns
7	applying any kind of simple or complex data validation (e.g., if the first 3 columns in a row are empty then reject the row from processing)

Table 1:Transforming Task

1.3 loading the data into a data warehouse or data repository other reporting applications

VI. T-SQL

Transact-SQL (T-SQL) is Microsoft's and Sybase's proprietary extension to SQL. SQL, the acronym for Structured Query Language, is a standardized computer language that was originally developed by IBM for querying, altering and defining relational databases, using declarative statements.

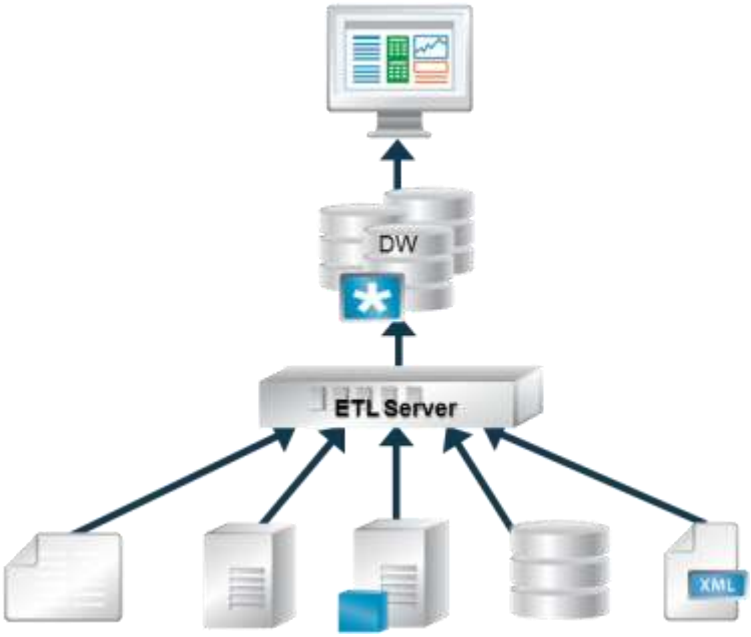

Fig. Data integration

VII. CHALLENGES OF DATA INTEGRATION IN DATA WAREHOUSE

6.1 Converting data types

You need convert between data types as part of an ETL process to ensure that source data types do not cause the process to fail.

6.2 Removing Duplicates from Data

You have data loaded into an SQL Server table – possibly a staging table -and you to remove any duplicate records.

6.3 Deduplicating Data in an ETL Data Flow

You wish to deduplicate data as it flows through an SSIS load package.

6.4 Subsetting column data using T-Sql

You need to break down the contents of a column into smaller, discrete portions.

Table 2: Causes of bad data in data integration

	Causes of bad data in data integration
1	Wrong information entered into source system
2	As time and proximity from the source increase, the chances for getting correct data decrease
3	In adequate knowledge of interdependencies among data sources incorporate DQ problems.
4	Inability to cope with ageing data contribute to data quality problems
5	Varying timeliness of data sources
6	Complex Data warehouse
7	Unexpected changes in source systems cause DQProblems
8	System fields designed to allow free forms (Field not having adequate length). Missing values in data sources
9	Additional columns
10	Use of different representation formats in data sources
11	Non-Compliance of data in data sources with the Standards
12	Failure to update all replicas of data causes DQ Problems.
13	Approximations or surrogates used in data
14	Different encoding formats (ASCII, EBCDIC,....)
15	Lack of business ownership, policy and planning of the entire enterprise data contribute to data quality problems.

VIII. SUGGESTION FOR IMPROVING DATA QUALITY IN DATA INTEGRATION

1. Use the SSIS Data Conversion task to change data types in the data flow and ensure that the destination data types can accept the source data. In SSIS , a data flow task you can use the data conversion task to change a data type. You will be converting to one of the SSIS internal data types. These are described in Appendix A. One a Data Conversion task has been implemented as part of data flow (most often by connecting it to a Data Source),you can select the column(s) whose data you wish to change and then select the destination data type. SSIS will create a second column for each modified data type ,so be sure to map the appropriate column in the destination task.[4]
   At risk of laboring the blindingly obvious, you can also convert data in T-SQL using two main functions:
   - CAST
   - CONVERT
2. Use the ROW_NUMBER() windowing function and a CTE(common table expression) to deduplicate records. When all the data that you load is already perfect, then you have few , if any , problems. Unfortunately , the real world is not always that simple , and one of the essential- indeed often one of the first – things that must done is to remove duplicates from a set of data. Duplicates can not only cause problems for users further down the line (and cause problems whose cause may be hard to trace back to the source), they can cause ETL processes to fail outright. This can become painfully clear if you are using the T-SQL MERGE command , which will choke on duplicate key data for instance.[4]
3. Use the SSIS Aggregate transform to remove duplicates in a data flow. SSIS also allows you to remove duplicates. However you will look in vain for a task with this name, or anything like it in the toolbox. Instead you can do this using the aggregate transform. Setting the operation to group by all columns will deduplicate records.[4]
4. A frequent requirement especially if your data is sourced from an older system, is to break down the contents of a column into smaller, discrete portions. Think of a social security number, where different parts of the data have different significations. Both T-SQL and SSIS can perform this task equally easily and equally fast. The only difference can be how to overcome the problem of variable length substrings. [4]
   In T-SQL there are three, core string manipulation functions that help you separate out the elements of a fixed length string. As they need little or no explanation, it suffices to say that they are, as Left, substring, Right
   Breaking down strings into multiple substrings when the strings are of variable length is only a little harder. It will require judicious use of the following string manipulation functions:

LEN	Returns the number of characters in a string
CHARINDEX	Gives the position of one or more characters in a string
REPLACE	Replaces one or more characters in a string
REVERSE	Reverse a string characters

Table 3: String Manipulation functions

IX. CONCLUSION

The data warehouse is the integration of different resources with having different structure. The data in data warehouse is in textual format, numeric data, some graphs, Tables this data is integrated in the one uniform format so the analyst use this type of data for decision making purpose. Business intelligence is the totally depends on data stored in the warehouse.

But this heterogeneous data is integrated in unified structure is the big challenge for the data integrity in data warehouse. In this paper we discuss some problems and some suggestion for data integration which is helpful for decision making regarding data integration. We discuss some methods for integration in data warehouse.

REFERENCES

Journal Papers:

[1] Razi O. Mohammed and Samani A. Talab, Clinical Data Warehouse Issues and Challenges, International Journal of u-and e-Service, Science and Technology
[2] Rahul Kumar Pandey, Data Quality in Data warehouse: problems and solution, IOSR Journal of Computer Engineering (IOSR-JCE)

Books:

[3] paulraj-ponniah, data-warehousing-fundamentals
[4] ADAM ASPIN, SQL SERVER 2012 DATA INTEGRATION RECIPES: SOLUTIONS FOR INTEGRATION SERVICES ...
[5] Data Warehouse - Basic Concepts - SSDI
[6] The Data Warehouse ETL Toolkit [Wiley 2004].pdf - root
[7] Data Warehousing in the Age of Big Data By Krish Krishnan
[8] DW 2.0: The Architecture for the Next Generation of Data Warehousing: The ...By W.H. Inmon, Derek Strauss, Genia Neushloss
[9] http://www.compositesw.com/
[10] http://datawarehouse4u.info/
[11] docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm
[12] https://en.wikipedia.org