Data Quality Issues and Current Approaches to Data Cleaning Process in Data Warehousing

Jaya Bajpai
MBA-IT Student
SICSR, affiliated to Symbiosis International University (SIU), Pune, and Maharashtra, India

Pravin S. Metkewar
Associate Professor
SICSR, affiliated to Symbiosis International University (SIU), Pune, and Maharashtra, India

Abstract

In this paper we have discussed the problems of data quality which are addressed during data cleaning phase. Data cleaning is one of the important processes during ETL. Data cleaning is especially required when integrating heterogeneous data sources. This problem should be addresses together with schema related data transformation. At the end we have also discussed the Current tool which supports data cleaning.

Keywords- Data cleaning, Data warehousing, Data cleaning tools, ETL, Data quality

I. INTRODUCTION

Data cleaning, also called data scrubbing, deals with detection and removing of errors and inconsistencies from data in order to improve the quality of data. Data quality problem are present in single data collection, such as files and databases. Problems can arises due to misspelling which take place during data entry process, missing information or other valid data. The need of data cleaning increases rapidly when multiple data stream are need to be integrated e.g. in data warehouses, federated database system or global web-based information system. This is because the sources often contain redundant data in different data in different representation. The main aim is to provide access to accurate and consistent data, consolidation of different data representation and to eliminate the duplicity of data.

II. DATA CLEANING PROBLEM

This section classifies the major data quality problem to be solves by data cleaning and data transformation. Data transformations are needed to support any changes in the structure, repetition and content of data. These transformation becomes necessary in many situations e.g. of deal with schema evolution, migrating a legacy system to a new information system, or when multiple data sources are to be integrated.

We roughly distinguish between single source and multi-source problems and between schema and instance related problem. Schema level problems are also reflected in instances .They can be addresses at the schema level by an improved schema design, schema translation and schema integration. Instance level problem refers to the errors and the inconsistencies in the real data content which are not visible at schema level. They are primary focus on data cleaning .The single source problems also occur in multisource cases.

A. Issues Caused due to Single Source of Data

In single source data problem arises at 2 levels. One is schema level whereas other is instance level

1) Problems at Schema Level

There are various single source problems that occur at schema level. These problems occur due to lack of appropriate model specific or application specific integrity constraint. Sometimes there is limitation in data model or the design of schema is poor. Sometimes only few integrity constraints are defined to limit the overhead of integrity control.

2) Problem at Instance Level

The problems which arises at instance level i.e. errors and inconsistencies, cannot be prevented at the schema level example misspelling.

<table>
<thead>
<tr>
<th>Scope</th>
<th>Dirty Data</th>
<th>Remark/Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>Illegal Values</td>
<td>Bday: 23.14.90</td>
</tr>
<tr>
<td>Record</td>
<td>Violation of Attribute Dependencies</td>
<td>Age=10; bday=20-08-1950</td>
</tr>
<tr>
<td>Record Type</td>
<td>Uniqueness Violation</td>
<td>Cust 1 = &quot;mohan&quot;,adhar no:124</td>
</tr>
<tr>
<td>Sources</td>
<td>Referential Integrity Violation</td>
<td>Cust name=Sumit dept=409</td>
</tr>
</tbody>
</table>
The reduce the cleaning problem; we should prevent dirty data to enter the data source. If dirty data enters the data sources, the cleaning of the data sources become extremely difficult and expensive. To solve this problem there should be appropriate design of database schema, integrity constraint as well as of data entry applications.

B. Issues Caused due to Multi Source Data
The problems of dirty data which is present in single source become enlarged when multiple sources need to be integrated. The data in the sources file may have been represented differently; the data may be overlapped or contradict. This happen because the data sources are present at different location and they are deployed and maintained independently to serve specific needs.

Due to this large amount of heterogeneity take place in data management system, data models, schema design and the actual data.

At schema level, the data model and the schema design difference are to be addressed by the steps of translation of schema and integration of schema respectively.Naming and structural conflict take place at schema design. Naming conflict arise when different names are used for same object or same name is used for different object.

The conflicts in structure occurs in many variations and it refers to the different representations of the same object in different sources e.g. attribute representation vs table representation of the data, different component structure, different data type, different integrity constraints.

In addition to schema level, many conflicts arrive at instance level also. These are data conflicts. All problems from the single source case can occur with different representation in different sources. This includes duplicated records, contradiction records, different value representation for gender, different characters for marital status, different measurement of values (some may be in rupees, some in dollars, some in euro). Apart from this there are various aggregation level also(sales per product vs sales per product group) as refer to different points in time.

The main problem in the process of cleaning the data from various streams is to identify the overlapping of data, in particular the matching records refers to the same real-world entity. This problem is also referred as merge/purge problem, duplicate elimination or object identity problem.

To solve the problem of multiple source, duplicate information should be purged out there should be consolidation and merger in order to achieve a consistent view of real world entities.

<p>| Table 2: Issues caused due to multisource problem |</p>
<table>
<thead>
<tr>
<th>Cno</th>
<th>Name</th>
<th>Street</th>
<th>City</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Mili Kare</td>
<td>22,H road</td>
<td>Kolkata,WB</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>Harish Kare</td>
<td>H street 2</td>
<td>delhi</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No</th>
<th>Lname</th>
<th>Fname</th>
<th>Gender</th>
<th>Street</th>
<th>City</th>
<th>State</th>
<th>Phone</th>
<th>Cid</th>
<th>Cno</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>kare</td>
<td>mili</td>
<td>f</td>
<td>H road,22</td>
<td>kolkata</td>
<td>WB</td>
<td>7866574876</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>kare</td>
<td>harsh</td>
<td>m</td>
<td>H street 2</td>
<td>delhi</td>
<td>delhi</td>
<td>78685976854</td>
<td>79</td>
<td>79</td>
</tr>
<tr>
<td>3</td>
<td>kare</td>
<td>sumit</td>
<td>m</td>
<td>HN cross</td>
<td>pune</td>
<td></td>
<td>78685976854</td>
<td>79</td>
<td>79</td>
</tr>
</tbody>
</table>

III. DATA CLEANING APPROACHES

There are several phases evolved in data cleaning

A. Data Analysis
A detailed data analysis is required to detect which kind of errors and inconstancies are to be removed. Manual inspection, data sample are used for analysis of data. In spite of these, analysis programs should also be used to gain metadata about data properties and to detect the problems that are related to data quality.

B. Definition of Transformation Work Flow and Mapping Rules
A large number of data transformation and cleaning step have to be executed depending upon the data sources, their degree of heterogeneity and dirtiness of data. Early data cleaning step can correct the single source integration problem and prepare data for integration. In later steps, data integration/schema and cleaning multisource instance problem e.g. duplicates can be dealt with.

There should be cleaning steps and declarative query in the schema-related data transformation to enable automatic generation of the transformation code. We can also use user-written cleaning code and during transformation a special purpose tool can also be used. There are some data instance which do not have any cleaning logic, thus transformation steps may request user feedback for them.
C. Verification
The transformation workflow and transformation definition should be tested and evaluated for correctness and effectiveness. It may be needed to improve the sample or source data or improve the definition if necessary. Multiple repetitions of analysis, verification and design steps are needed. E.g. some errors only become important after applying some transformation.

D. Transformation
Execution of transformational steps is required either by running the ETL workflow for loading and by refreshing the data in a data warehouse or during the period of answering the queries on multiple sources.

E. Back Flow Process of Cleaned Data
After the removal of single source errors, the cleaned data should also replace the dirty data in the original source in order to give all legacy application the improved data. This will avoid the redoing the cleaning work for future data extraction. For data warehousing cleaned data is available from the data staging area.

1) Data Analysis Stage in Data Cleaning
If there is few integrity constraints, then meta data reflected in the schema is typically insufficient to assess the data quality of source. It is thus important to analyses the actual instances to obtain real meta data on data characteristics or unusual value pattern.

This metadata helps finding data quality problem. There are 2 focused approaches for data analysis, data profiling problems and data mining.

Data profiling focus on the instance analysis of the individual attribute. It derive information such as the type of data, the length, value, discrete values, range and their frequency, variance, uniqueness, occurrence of null values, typical string pattern (phone number) etc. providing an exact view of various quality aspects of the attribute.

2) Problem of Conflict Resolution in Data Cleaning
A set of transformation steps has to be specified and executed to resolve the various schema level data and instance level data quality problems. In order to deal with single-source problems various transformation are need to be performed and to prepare for integration with other sources. There are various steps which are as follows:

<table>
<thead>
<tr>
<th>Extraction of values from the free form of Attributes</th>
<th>Validation and correction</th>
<th>Standardization</th>
</tr>
</thead>
</table>
| 1) Free-form of attributes captures the multiple individual values that should be extracted to achieve a precise representation and support the next step of instance matching and duplication of elimination. | 1) This step examines each source instance for data entry errors and to correct the error automatically as far as possible. 
2) This step examines each source instance for data entry errors and tries to correct them automatically as far as possible. 
3) 3. Attribute dependencies (date of birth to age, total price depends on unit price / quantity, city is dependent on phone area code that can be used to detect problems and substitution of missing values or correct wrong values. | 1) To facilitate instance matching and integration, and conversion of attribute values should be done in consistent and uniform format. |

Example: name and address

Example:
Name, string data should be converted to upper or lower case.

IV. TOOL SUPPORT
A large variety of tools is available on the market to support data transformation and data cleaning tasks, in particular for data warehousing.

1) Some tools concentrate on a specific domain, such as cleaning name and address data, or a specific cleaning phase, such as data analysis or duplicate elimination.

2) Due to their restricted domain, specialized tools typically perform very well but must be complemented by other tools to address the broad spectrum of transformation and cleaning problems.

3) Other tools, e.g., ETL tools, provide comprehensive transformation and workflow capabilities to cover a large part of the data transformation and cleaning process.

4) A general problem of ETL tools is their limited interoperability due to proprietary application programming interfaces (API) and proprietary metadata formats making it difficult to combine the functionality of several tools.
A. Specialized Cleaning Tools

Features of Specialized Data Tools are:

1) Specialized cleaning tools address a particular domain, mostly name and address data, and focus on redundant data. The transformations are to be given either beforehand in the form of a rule library or via interaction by the user. A number of commercial tools, e.g., IDCENTRIC (FirstLogic), PUREINTEGRATE (Oracle), QUICKADDRESS (QASSystems), REUNION (PinneyBowes), and TRILLIUM (TrilliumSoftware), works on cleaning this kind of data. They provide methods like extracting and transforming name and address information into individual atomic elements, validation of street names, cities, and zip codes, in combination with a matching facility based on the cleaned data. They have huge library of pre-specified rules dealing with the problems which occur in processing this data. For example, TRILLIUM’s extraction (parser) and matcher module contains over 200,000 business rules. The tools also provide facilities to custom or extend the rule library with the user-defined rules for specific needs.

2) Duplicate elimination: Sample tools for duplicate identification and elimination include DATACLEANSE (EDD), MERGE/PURGELIBRARY (Sagent/QMSoftware), MATCHIT (HelpITSystems), and MASTERMERGE (PinneyBowes). Usually, they require the data sources already be cleaned for matching. Several approaches for matching attribute values are supported; tools such as DATACLEANSE and MERGE/PURGE LIBRARY also allow user-specified matching rules to be integrated.

B. List of ETL Tools used for Data Cleaning Process

An extensive number about business devices backing those ETL procedure to information warehouses over a thorough way, e.g., COPYMANAGER (Information Builders), DATASTAGE (Informix/Ardent), extricate (ETI), POWERMART (Informatica), DECISIONBASE (CA/Platinum), DATATRANSFORMATIONSERVICE (Microsoft), METASUITE (Minerva/Carleton), SAGENTSOLUTIONPLATFORM (Sagent), Also WAREHOUSEADMINISTRATOR (SAS). They use repossess fabricated on a dbms will wrst bindings every one metadata around those information sources, target schemas, mappings, script programs, and so on. Previously, a uniform manner. Schemas What's more information are concentrated from operational information sources by means of both local record Also dbms gateways and additionally standard interfaces for example, such that ODBC Also EDA. Information transformations need aid characterized for a easy-to-use graphical interface. To point out individual mapping steps, a proprietary principle dialect also a thorough library of predefined transformation capacities are regularly pro- vided. The instruments Additionally help reusing existing conversion solutions, for example, outside C/C++ routines, by giving work to a interface with incorporated them under the inside change library. Conversion preparing may be conveyed crazy Possibly Eventually Tom's perusing a motor that interprets those specified transformations at runtime, alternately by aggregated code. Every last bit engine-based devices (e.g., COPYMANAGER, DECISIONBASE, POWERMART, DATASTAGE, WAREHOUSEADMINISTRATOR), have An scheduler and backing workflows for complex execution dependencies Around mapping occupations. An workflow might additionally conjure outside tools, e.g., to specific cleaning errands for example, such that name/address cleaning or copy disposal. ETL devices regularly have minimal inherent information cleaning abilities in any case permit the client with define cleaning purpose through a proprietary API. There may be as a rule no information examination backing to naturally identify information errors also inconsistencies. However, clients could actualize all the rationale with the metadata looked after Also Toward deter- mining substance aspects for those help about amassed works (sum, count, min, max, median, variance, deviation). The furnished change library blankets Numerous information conversion Also cleaning needs, for example, such that information sort conversions (e.g., date reformattting), string works (e.g., split, merge, replace, sub-string search), arithmetic, experimental What's more measurable functions, and so on. Extraction from claiming qualities starting with free-form qualities may be not totally programmed yet the client need will define the delimiters dividing sub-values. Those lead dialects regularly spread if-then Furthermore body of evidence constructs that assistance taking care of exceptions done information values, for example, misspellings, abbreviations, lost or enigmatic values, What's more qualities outside of go. These issues could additionally make tended to toward utilizing a table lookup build Furthermore join.

V. CONCLUSIONS

In this paper we have tried to classify problems of data quality in data source .The difference between single and multi -source and between schema and instance level problem is highlighted. We have also provided an overview of commercial data cleaning tools more work is needed on design and implementation of language for supporting both schema and data transformation. Data
cleaning is also used for query processing on heterogeneous data sources. E.g. in web based information system. Data cleaning is an important part in data warehousing and it is important to understand tools which are required for data cleaning.

REFERENCES