Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-9, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Data Profiling Issues in Data Warehousing

Nikeeta N Shevgan[1] &Pravin S Metkewar[2]
[1]MBA-IT, SICSR, Affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India.
[2]Assoc. Professor, SICSR, Affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India.

Abstract: *Data profiling is used for data governance. It is process of discovering anomalies in data which breaks the business rules of the system required to work it properly. Thus the data profiling is used on structured, semi-structured and unstructured data to check data patterns, rules, and validations to make decisions on data. In this paper pattern matching issue of data is going to be discussed and solution for the same is proposed with the help of methodologies.*

Keywords*: Data profiling, Data governance, Data patterns.*

1. Introduction

Data profiling is a particular sort of information examination used to find and describe vital elements of data sets. Profiling gives a photo of data structure, substance, guidelines and connections by applying measurable systems to give back an arrangement of standard attributes about data - field lengths and cardinality of segments, granularity, esteem sets, position designs, content examples, inferred principles, and cross-section and cross-record data connections and cardinality of those connections. In data profiling there are numerous issues in a matter of seconds data stockroom is confronting like Insufficient data profiling of data sources is in charge of data quality issues; Manually inferred Data about the data substance in operational frameworks spreads poor data quality; Inappropriate determination of Automated profiling apparatus cause data quality issues and some more.

Data profiling uses various types of distinct measurements, for example, least, most extreme, mean, mode, percentile, standard deviation, recurrence, and variety and in addition different totals, for example, number and aggregate. Extra metadata data got amid data profiling could be data sort, length, discrete qualities, and uniqueness, event of invalid qualities, run of the mill string examples, and dynamic sort acknowledgment. The metadata can then be utilized to find issues, for example, incorrect spelling, missing qualities, shifting worth representation, and copies. Different analyses are performed on different structures.

2. History Research

A study by Gartner estimated that "more than 25 percent of critical data within Fortune 1000 enterprises" to be flawed. This means data quality is the problem in most organisations.

Data profiling procedure may comprise of numerous strides, for example, beginning record of task is made. It comprises of deliverable limits and time of venture. The designer group ought to be acquainted with this present procedure's for vital time is spared in doing required procedure on required information, this record moreover show what are the desires and necessities

A. Data Profiling The Old Way

Manual Approach

Traditionally, data profiling required a skilled specialized asset who could manually query the data source using Structured Query Language (SQL). There is frequently a no communication between the business expert who realizes what the information should to be, and the specialized technical programmer who knows SQL

B. Data Profiling The New Way

Benefits of using data profiling

i. Increased Speed:

Industry gauges for manual data profiling are roughly 3-5 hours for each characteristic; by using data profiling tools, this can be diminished to 15-30 minutes for each attribute. Test ROI, expecting 1500 qualities: $281,250 short the expense of data profiling software

ii. More Thorough Analysis:

With a manual methodology, by and large just a subset of the characteristics and the lines

are tried; with data profiling tools, thorough assessment/analysis of the data can be performed.

iii. Normal Repository:

Data profiling tool give a typical vault to putting away data profile results and other key metadata, for example, notes made during analysis.

- Information profile data is concentrated
- Whole group can share and leverage the data

There are so many tools available in the market to improve data quality by data profiling.

3. Issues In Data Profiling

There are numerous problems with data profile which degrade the quality of the data like:

- Insufficient data profiling of the data sources
- Manually derived information about the data contents in operational data stores.
- Inappropriate selection of data profiling tool.
- Lack of identification of missing data relationship, dependencies, etc.
- Pattern matching problem in data profiling.

These are some of the problems data profiling and in this paper we are going to resolve the problem of pattern matching.

There is a metadata file along with the data profile which has business rules defines in it. Accordingly data profile must match with the metadata and improves the quality of the data. But the problem in pattern matching is that if there is a business rule stating that email id of employee must have one special character then all the email id should match the metadata if it does not match then the data profiling tool will analyze the data show issue with the concerned email id.

Data profiling is just a analyzing tool for the data but when the issue in data Is found then it is no more concern of data profiling tool; here I am broadening the scope of the data profiling tool by analyzing the data as well as giving the solution for the data which is not profiled as per the business rules.

4. Proposed Solution Towards The Problem

The fig A shows the solution of the pattern matching issue of the data profiling.

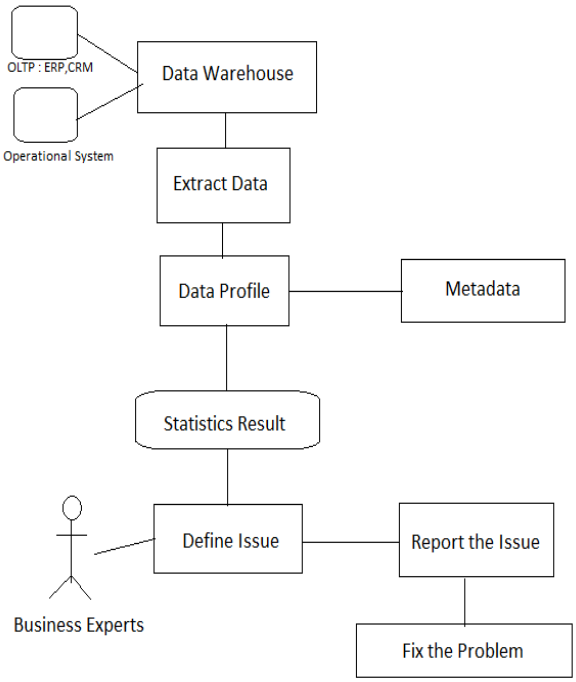

Figure 1: Data Profiling

It explains:

- First data is collected in the organizations data warehouse.
- Information or file on which profiling has to be is extracted information or file data selected to be profiled; metadata file (having all the business rules defined in it) for the chosen data is selected.
- Then profiling is started and results of profiling is generated which statistics of the data showing number of fields having problem and fields which are accurate as per the metadata.

Up to this step the normal data profiling is done now am expanding the scope of it and solving the fields having problem with the metadata.

- Here problem will be stated by the business stewards and accordingly severity of the problem will also be stated.

Ex : if employee's name is not spelled correctly, email id is also not as per the business rules but the phone number of the employees are correct and math with the given metadata. So here business stewards will take look to the statistics

results of the data and prioritized the problem as per the severity.

Like as in this case business stewards will mark the issue according to the severity of data.

- Employee's phone number is correct so organisation can contact them through the phone calls so no severe issue email id for communication.
- Employee's name is the key thing which must be correct because misspelled names can arise many problems within an organization so misspelled names must be corrected.

So business experts (stewards) will classify the issue and mark it with priority colours:

Red : severe Issue.
Green : not an issue
Blue : requires additional review

After this step business analyst or expert will send report to the technical programmer who will fix the problem as per stated by the analysts

5. Conclusion

Hereby I conclude by this paper that data profiling is not just analysing tool for the data related issues it can also be provide solution to the problem to improve data quality. Defining a problem with priority and fixing it accordingly will improve the data and make it more useful in taking business decisions.

6. References

[1] Brett Dorr, Pat Herbert,"Data Profiling: Designing the Blueprint for Improved Data Quality", DataFlux Corporation, Cary, NC,USA.

[2]Shankar Ganeshh R, Sathish Kumar Srinivasan, Subramanyam B S, "Data Profiling – A Quick Primer on the What and the Why of Data Integration", Architecture and Technology Services HCL Technologies, Chennai, August 2008.

[3] Michael Anderson, ". Data Profiling: The First Step in Data Quality", on January 23rd, 2012

[4] Felix Naumann," Data Profiling Revisited", Qatar Computing Research Institute (QCRI), Doha, Qatar.

[5] Sweety Patel, Piyush Patel, "Data Profiling", *Fairleigh Dickinson University, NJ- 07666, USA , Rajasthan Technical University, India,* Original Research Article,2012.

[6] https://en.wikipedia.org/wiki/Data_profiling

[7] http://searchdatamanagement.techtarget.com/definition/data -profiling.

[8] Erhard Rahm and Hong Hai Do (2000), "Data Cleaning: Problems and Current Approaches" in "Release of the Technical Committee on Data Engineering", IEEE Computer Society, Vol. 23, No. 4, December 2000.

[9] Ranjit Singh, Dr. Kawaljeet Singh," A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing",University College of En ineering (UCoE), Punjabi University, Patiala (Punjab), INDIA, May 2010.

[10] Rahul Kumar Pandey," Data Quality in Data warehouse: problems and solution", *PhD Scholar, Surguja university (Chhattisgarh), India,2014.*