Data Doctor: An Efficient Data Profiling and Quality Improvement Tool

Shruti Sarkate
B.Tech Final Year CST
UMIT, SNDT University
SantaCruz, Mumbai

Kaveri Tare
B.Tech Final Year CST
UMIT, SNDT University
SantaCruz, Mumbai

Ruchi Kamble
B.Tech Final Year CST
UMIT, SNDT University
SantaCruz, Mumbai

Narendra Gawai
Asst Professor
Dept Of CST
UMIT,SNDT University

Abstract— Many business and IT managers face the same problem: the data that serves as the foundation for their business applications is inconsistent, inaccurate and unreliable. Data profiling is the solution to this problem and, as such, is a fundamental step that should begin every data-driven initiative. In this paper we have implemented the technique of data profiling such as Column Analysis, Frequency Analysis, Null Rule Analysis, Constant Analysis, Empty Column Analysis and Unique Analysis.

*****

I. INTRODUCTION

Data profiling is a process for analyzing of data. It will discover anomalies and will help in better understanding of data.

Data profiling is an analysis of the candidate data sources for a data warehouse to clarify the structure, content, relationships and derivation rules of the data. Profiling helps to understand anomalies and to assess data quality, but also to discover, register, and assess enterprise metadata. Thus the purpose of data profiling is both to validate metadata when it is available and to discover metadata when it is not. The result of the analysis is used both strategically, to determine suitability of the candidate source systems and give the basis for an early go/no-go decision, and tactically, to identify problems for later solution design, and to level sponsors' expectations. The benefits of data profiling is to improve data quality, shorten the implementation cycle of major projects, and improve understanding of data for the users. Discovering business knowledge embedded in data itself is one of the significant benefits derived from data profiling. Data profiling is one of the most effective technologies for improving data accuracy in corporate databases. Although data profiling is effective, then do remember to find a suitable balance and do not slip in to "analysis paralysis". Standard data profiling automatically compiles statistics and other summary information about the data records. It includes analysis by field for minimum and maximum values and other basic statistics, frequency counts for fields, data type and patterns/formats, and conformity to expected values. Other advanced profiling techniques also perform analysis about the relationships between fields, such as dependencies between fields in a single set and between fields in separate data sets.

Data profiling is the process of examining the data available in an existing data source and collecting statistics and information about that data. Profiling data is an important and frequent activity of any IT professional and researcher.

Data profiling essentially consists of three aspects of analysis:

- **Column analysis,** deals with the analysis of values in terms of ranges, data type, size,and number of occurrences of the values. It also counts the null values.
- **Column profiling :** This will help the user in discovering total number of records, null **percent**, unique percent, minimum, and maximum value in column, documented data type etc
- **Frequency Analysis:** This profiling will help the user in **finding** total number of distinct values in the columns.
- **Null Rule Analysis:** This will help the user in finding all the columns in the table which has 100% null values.
- **Constant Analysis:** This **will** help the user in discovering those columns which has less than 4 and greater than 0 distinct values.
- **Unique Analysis :** This will help the user in finding all the columns in table which has 100%

uniqueness.

- **Empty Column Analysis:** This will help the user to find all the columns which has all the null values.
- **Dependency analysis** deals with the association of columns and find out the relationships between the data column. It is mostly useful for building the data warehouse. It is mainly used to find out referential integrity constraints between the tables. Dependency Analysis is used to find out parent child relationship among the tables.
- **Redundancy analysis,** deals with the redundant columns in the different **column** name with same value or same column name with different data values. Data analyst can use this information for making conclusion. [11]

II. EXISTING SYSTEM AND TOOLS

Data integration projects often run into difficulties because the exact nature and quality of the underlying data is not known up-front. When data-related issues are not fully discovered until late in the project, they typically cannot be easily quantified in terms of project effort impact and, therefore, cause unexpected de-livery delays and budget overruns. To overcome the data related issues and find out the effectiveness and usefulness of data there is a need for such a tool which will provide all these data issues and helps business users, data analyst to find a way to improve the data and helps in Data Modeling activities. The new application that will be developed must have following modules:

- Displaying the available servers
- Displaying the list of all the databases on the selected server
- Displaying the list of all the tables for a selected server and a database
- Displaying list of all the columns for a selected server, database and table
- Options for performing different types of profiling Application must support following types of profiling:

1. Column Profiling
2. Frequency Analysis
3. Constant Analysis
4. Null Rule Analysis
5. Empty Column Analysis
6. Unique Column Analysis

- In addition to above mentioned profiling activities, the application must also support Single Table Structural Profiling
- A feature must be provided to export the profiling results to an Excel file
- Data Quality (Sufficiency and uniqueness) indicators must be provided for the selected column.

Initially the data Profiling activities used to be done by writing complicated SQL queries**.** This would be comfortable for analyst or user who knows to write SQL queries**.** Many of us do not know the proper syntax and format for writing SQL queries**.** To overcome this, Data Profiling tools were introduced**.** Data Profiling Tools, to a some extent overcome the limitations for writing complex queries**.** All types of profiling activities were not supported by the tools**.** User has to understand and learn how to use the tool.

Data Cleaner

Data Cleaner is a data quality and data profiling tool for data validation, and comparison. It has limited database connection capability and has limitation in generating reports.

Limitations: The license cost of tool is very high and hence not affordable for small size business. [12]

Talend Open Profiler

Talend Open Studio is used for data integration and is based upon Eclipse RCP. It can generate code using java or perl data transformation scripts. The GUI comprises of metadata repository and a graphical designer.

Limitations: Limited features, not complete data profiling and data quality tool. [12]

Trillium

Trillium is leading data quality tool in the market having great capability to improve quality of data.
Limitation: It offers a limited number of direct database connectors and it relies heavily on ODBC. Its mostly data quality tool, limited data profiling.

Microsoft

Microsoft SQL Server 2012 has inbuilt Data Quality Services which incorporates Data profiling, as a part of the product.

Limitation: It is optimized specifically for Microsoft SQL Server Sources (SQL Server, excel and CSV Files.). Other database cannot be profiled using this tool.

III. III. PROPOSED SYSTEM

As recent as a few years ago, the area of data analysis- understanding the quality and structure of data assets within an application- was a relatively ill-defined area within an organizations IT strategy. Traditional approaches to data analysis are usually dependent upon a combination of inputs-documentation, individual knowledge, and ad-hoc data base query tools- which are used to select aspects of a data source. Such approaches are often time consuming and incomplete, as analysis tends to be concentrated in known areas of the data. Data problems abound in most organizations. Data problems can include data inconsistencies, anomalies, missing data, duplicated data, data that does not meet business rules, and orphaned data just to name a few. These problems can limit or even ruin your data initiatives. Because organizations rely on data that is inconsistent, inaccurate, and unreliable, large-scale implementations are ripe for failure or cost overruns. More disturbing, the organizations usually do not understand the magnitude of the problem or the impact that the problems have on their bottom line. Data problems within an organization can lead to lost sales and wasted money, poor decisions, sub-standard customer relations, and ultimately, failed businesses. Data profiling is the first step toward diagnosing and fixing problematic data. Data Explorer transforms the way companies think about data, enabling business analysts, data stewards, and IT developers to work together to profile all data, for all projects and all applications. A set of unified, role-based data profiling and discovery tools so that the business can be more self-sufficient and IT more productive. The main idea for developing this project is to implement a user friendly graphical interface for handling database functionality without using a direct query to retrieve data from database. This application will be useful for software programmers who use oracle data or SQL server base regularly, while developing application involving huge amount of data. Database is the place where information is stored in the form of tables so in order to organize database information users should log into oracle or SQL server database using console and execute queries which will be a time taking task. In order to overcome this problem a web based front end application is implemented. Using this application, programmers can retrieve data from database without passing direct queries. The sole objective of this project is to take strategic and analytical decisions in favor of users and small-scale organizations. Therefore the first request is to make visible the list of available servers present in the system. Once all the servers are displayed, one has to select a particular server and establish connection with the same. After successfully establishing connection with the server, all the databases present in it are retrieved. A database is selected and all the tables in the database are retrieved. And data profiling has to be performed on the selected table. The results obtained by data profiling can be used for statistical analysis in the organizations. The results obtained can also be directly mailed.

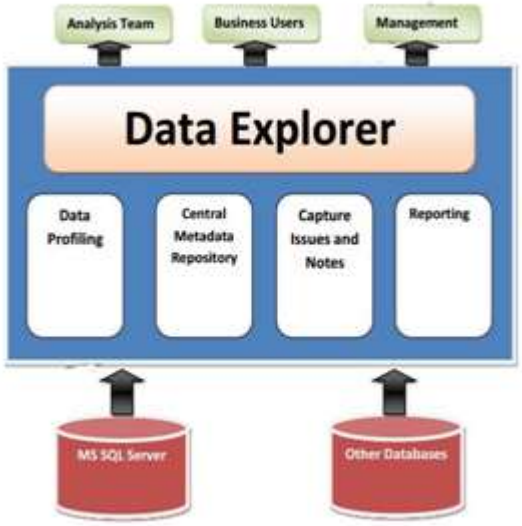

Fig.1 System Architecture

IV. IMPLEMENTATION DETAILS:

We have implemented the system by using .Net Framework and MS sql server, Oracle, Flat Files. The input Files may be from different databases such as MS sql server, Oracle, Flat Files

A sample interface for performing profiling is shown below:

V. SYSTEM ARCHITECTURE

At the very high level there will be analysis, business users and management who would be using Data Explorer. There will be Data Profiling module, central metadata repository, capture issues and notes and reporting. For data profiling activity data explorer will connect to MS sql server or oracle or any other database for which profiling has to be done, there is central metadata repository which will store the result. Capture issues and notes is a suggestion.

ISSN: 2321-8169
2292 - 2297

Fig.2 Main Interface: this is the main screen which gives 4 options to the user.

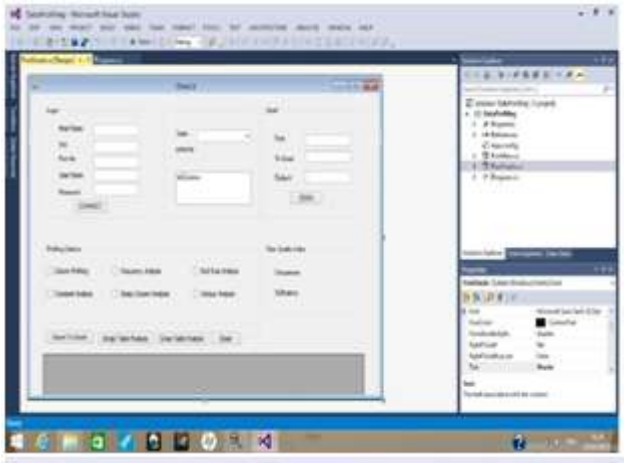

Fig.3 User Interface

Fig.4 Column Profiling: It gives total number of Records, Minimum and maximum values, unique percentage etc

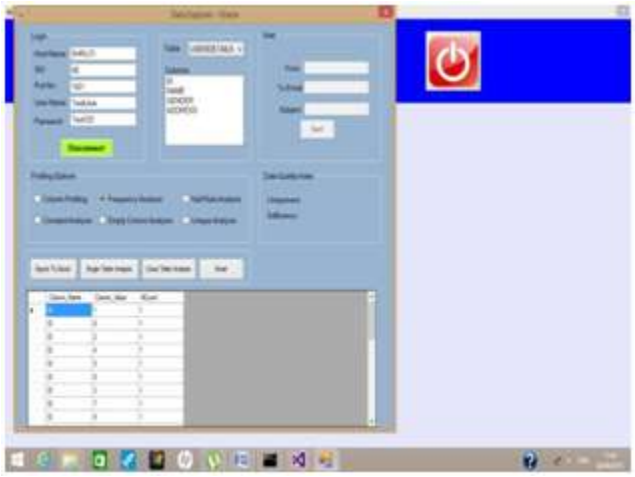

Fig.5 Frequency analysis: It gives us total number of distinct values in the column.

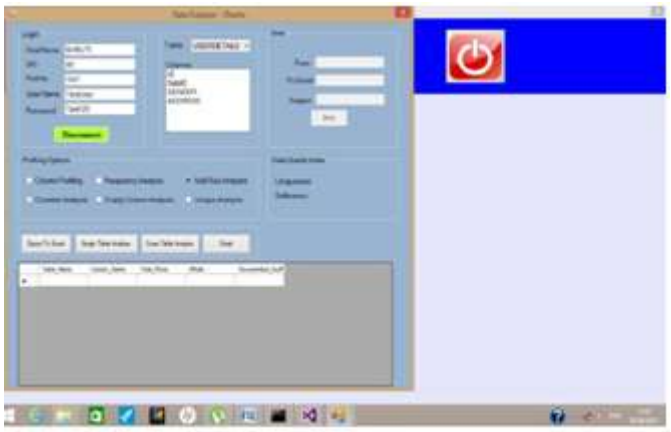

Fig.6 Null Rule Analysis: It gives us those columns which has 100% null values

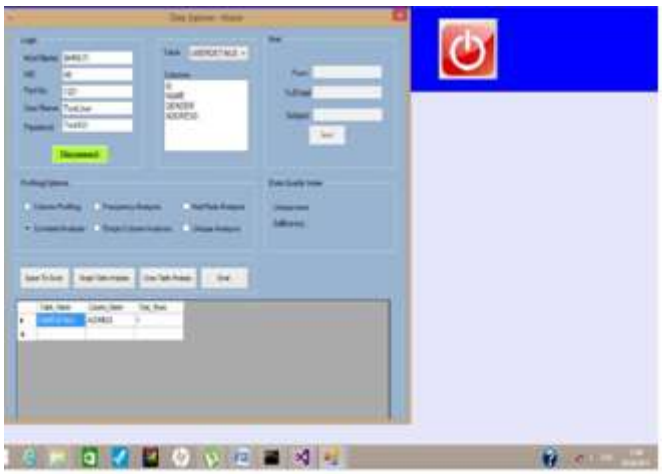

Fig.7 Constant Analysis: It gives us columns which has values greater than 4 and less than 0

 ISSN: 2321-8169
2292 - 2297

Fig.8 Empty column analysis: It gives us columns which has 100% null values.

Fig.9 Unique analysis: It gives us all the columns which has 100%uniqueness

Fig.10 Single table structure analysis: It will the user to find out primary key, composit primary key within the table

Fig.11 Cross table analysis: It will help the user the user to find out integrity constraint.

Here is the architecture diagram. Now at the very high level there will be Analysis, Business Users and management who would be using data explorer. So there will not be directly interaction interactively with our tool. Our tool will have different modules. Data profiling module, Central metadata repository, Capture Issues and Notes and Reporting. Now for data profiling activities, DP will connect to MS SQL server or oracle or any other database for which profiling has to be done, then there is central metadata Repository which will store the result. Capture issues and notes, this is again a suggestion but right now it is not implemented in current project. Bur we can have capture issues and notes means we can make DE more intelligent so that it itself identifies data problems better than human identifies it. So we can say it is further enhancement to the project and then we can have Reporting. It is also not implemented in the current scope but we can include it means we can develop the reporting on the basis of profiling enhancement. Right now we have reports in tabular format but we can build report in graphical format. We can build a trend report. Data quality can be updated everyday. So the trend report will show the difference between previous week data and current week data. So that is the trend how a DQ increases or decreases. And then again at the base level we have all the databases.

VI. VI. CONCLUSION AND FUTURE SCOPE

We proposed an efficient Data Profiling Technique. We have described the functionality of data profiling tool and the technique to come up with the rules. So, you want to make data a strategic asset at our organization. You understand that data must be consistent, accurate and reliable if you want your organization to be a leader. The

most effective approach to consistent, accurate and reliable data is to begin with data profiling. And the most effective approach to data profiling is to use a tool that will automate the discovery process. We are going to find out a technique for automatic discovery process. This tool further can be enhanced for unstructured data such as audios, videos and images.

REFERENCES

[1] Pitney Bowes, Data Profiling: Underpinning Data QualityManagement,2007
[2] Maunendra Sankar Desarkar, Data Profiling for ETL Processes, IIT Kanpur. 2008
[3] Data Profiling Revisited Felix Naumann_Qatar Computing Research Institute (QCRI), Doha, Qatar fnaumann
[4] J.Euzenat and P. Shvaiko. Ontolog Matching. Springer Verlag, Berlin Heidelberg New York, 2007.
[5] M. J. Cafarella, A. Halevy, and J. Madhavan. Structured data on the web. Communications of the ACM, 2011.
[6] F. Chiang and R. J. Miller. Discovering data quality rules. Proceedings of the VLDB Endowment, 2008.
[7] F. D. Marchi, S. Lopes, and J.-M. Petit. Unary and n ary inclusion dependency discovery in relational databases. Journal of Intelligent Information Systems, 2009.
[8] D. J. Abadi. Column stores for wide and sparse data. In Proceedings of the Conference on Innovative Data Systems Research(CIDR) Asilomar, CA, 2007.
[9] W. Wu, B. Reinwald, Y. Sismanis, and R. Manjrekar. Discovering topical structures of databases. In Proceedings of the International Conference on Management of Data (SIGMOD), Canada, 2008.
[10] A. Rostin, O. Albrecht, J. Bauckmann, F. Naumann, and U. Leser. A machine learning approach to foreign key discovery. In Proceedings of the ACM SIGMOD Workshop on the Web and Databases, 2009.
[11] David Loshin, ’Data Profiling Techniques for your BI program’, Nov 2004.
[12] Bhavin Shah, Dipti Darade,Nidhee Rathod, Rohini Sawant,’Data Explorer’, International Journal for Scientific Research & Data cleaner and Development, 2015.