

Data Explorer

Bhavin Shah¹ Dipti Darade² Nidhee Rathod³ Rohini Sawant⁴

^{1,2,3,4}Department of Information Technology

^{1,2,3,4}Mumbai University, Padmabhushan Vasantdada Patil Pratishthan's College of Engineering
Sion, Chunabhatti, Mumbai – 400022

Abstract— Data profiling is the process of examining the data available in an existing data source (e.g. a database or a file) and collecting statistics and information about that 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.[1]

Key words: Data profiling, CRM, ERP

I. INTRODUCTION

Organizations around the world are looking for ways to turn data into a strategic asset. However, before data can be used as the foundation for high-level business intelligence efforts, an organization must address the quality problems that are endemic to the data that's available on customers, products, inventory, assets, or finances. The most effective way to achieve consistent, accurate, and reliable data is to begin with data profiling. Many business and IT managers face the same problem: the data that serves as the foundation for their business applications (including customer relationship management (CRM) programs, enterprise resource planning (ERP) tools, and data warehouses) 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. Business intelligence (BI) mainly refers to computer-based techniques used in identifying and analyzing business data. [1]

Our project Data Explorer is based on these above mentioned concept i.e. Data Profiling and Business Intelligence. Data Explorer will transform the way any company thinks about data. With Data Explorer, business analysts, data stewards, and IT developers can work together to profile all data for all projects and all applications.

With Data Explorer, business analysts and data stewards can easily profile data themselves and monitor data issues on an ongoing basis using browser-based tools designed especially for them. IT developers can automatically discover and analyze data using prebuilt rules and a single, unified development environment to reuse data profiling results across projects, boosting productivity and eliminating errors. With Data Explorer, a business analyst will be able to discover and analyze all data anomalies across all data sources, find hidden data problems that put projects at risk, pinpoint structural issues that prevent data quality issues before they become a more of a problem.

II. LITERATURE SURVEY

Initially the data Profiling[5] 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 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. Using SQL Queries is a traditional approach where development time required is more. It also needs results to be exported to excel or notepad for analysis. The existing tools have complex user interface and limited functionality. The setup and installation cost is more. The license cost for a single platform is also high.

Traditional approaches to data analysis are usually dependent upon on a combination of inputs – documentation, individual knowledge and ad hoc data base query tools – which are used to selected 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 profiling tool sets, like BDQ Analysis, allow organizations to accurately and efficiently analyze and diagnose the quality of their data. By completing a process of analyzing complete data sources as one process, organizations capture a complete understanding of their data assets.

DataCleaner[6] is a data quality software application that is used for data profiling, validation, and comparison. It has limited connection capability and limited report capability.

Talend OpenStudio is an Open Source project for data integration and is based upon Eclipse RCP. Talend Open Studio operates as a code generator which allows data transformation scripts and programs to be generated using Java or Perl. The graphical user interface is comprised of a metadata repository and a graphical designer.

IBM InfoSphere Information Analyzer analyzes the structure, content and quality of data sources to uncover missing, inaccurate, and inconsistent data early in the data integration lifecycle. Informatica acquired Similarity ATHANOR product and has renamed it to Informatica Data Quality in January of 2006. Informatica acquired Evoke software to the data quality suite and has renamed Evoke Axio to Information Data Explorer in January of 2006.

III. PROPOSED SYSTEM

Data Explorer is based on concept of Business Intelligence. Business intelligence (BI) mainly refers to computer-based techniques used in identifying and analyzing business data. Our project Data Explorer is based on these above mentioned concept i.e. Data Profiling and Business

Intelligence. Data Explorer will transform the way any company thinks about data. With Data Explorer, business analysts, data stewards, and IT developers can work together to profile all data for all projects and all applications.

With Data Explorer, business analysts and data stewards can easily profile data themselves and monitor data issues on an ongoing basis using browser-based tools designed especially for them. IT developers can automatically discover and analyze data using prebuilt rules and a single, unified development environment to reuse data profiling results across projects, boosting productivity and eliminating errors. With Data Explorer, a business analyst will be able to:

- Discover and analyze all data anomalies across all data sources.
- Find hidden data problems that put projects at risk.
- Pinpoint structural issues that prevent data quality issues before they become a more of a problem.

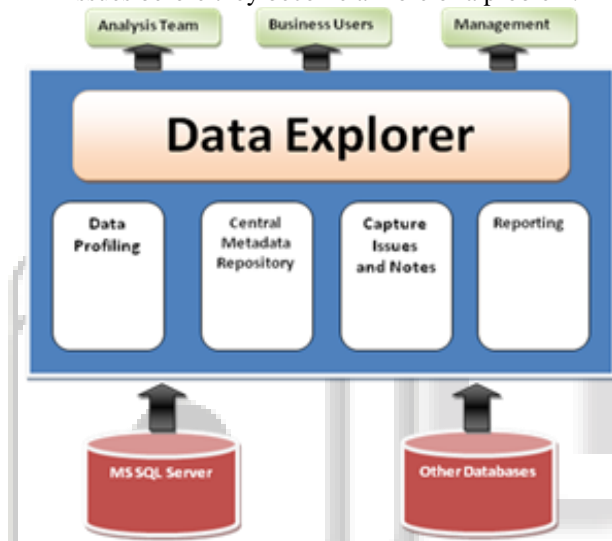


Fig. 1: Architecture diagram of Data Explorer
Data Explorer enables quick discovery of data issues.

The most evident feature of Data Explorer is no more writing of queries to profile data. It is time efficient as it shortens the implementation cycle of major projects. Due to simple graphical user interface it improves understanding of data for the naive users. Due to discovering business knowledge it improves data accuracy in corporate databases. It allows to analyze the data values to find areas that are incomplete, inaccurate or ambiguous. It can also verify relationships across columns and tables. Supporting multiple Databases like Oracle 10g, Oracle 11g, MS SQL Server 2005, MS SQL Server 2008, My SQL etc. Apart from the functionalities, it will also be useful for both Small and Big Industries as the licensing cost will be cheaper than the existing ones. Also, it will provide an integrated framework for the different types of Database and handling Data Quality. It will also provide support for unstructured data in flat files to carry out profiling activities. In short, it will be complete end to end package for both profiling and data quality.[2]

It supports different types of profiling like Column Profiling, Constant Analysis, Unique Analysis, Null Rule Analysis, Frequency Analysis, Empty Column Analysis, Primary / Composite Key Analysis.

The application allows the user to select a set of columns to analyze. It also allows the user to select the rules he wants to check. Then it comes up with the rules by looking at the entries present in those columns.

Column Profiling it is relatively straight forward and is used to get statistical properties of each column. It scans through the entire table and gets total no of records, null percentage, unique percentage, minimum and maximum value in the column, documented data type etc.

Constant Analysis it calculates the ratio of the number of the distinct elements and the total number of elements in the table. If the ratio is very low, it means that average frequency of the elements is very high and the column can be a categorical field. When this ratio falls below some predefined threshold, the application declares it as a categorical field. The domain of acceptable values is the set of the distinct elements in the column. For example it helps in discovering those columns which has less than 4 and greater than 0 distinct values.

Null Rule Analysis is used in finding all the columns in a table which has 100% NULL values. For checking nullability it finds out the number of null entries in that column. If there is at least one null entry, it means the column allows null values, since we assume the data source to be clean. If there is no null value, then it reports the column under consideration as a non-nullable column.

Unique Analysis is used in finding all the columns in table which has 100% uniqueness. For checking uniqueness of the field, the application finds out the number of distinct elements in the column. If number of distinct elements is same as the number of entries present in the column, it means that all the elements are distinct.

Primary Key / Composite Key Analysis helps us to find out the possible primary or composite key columns which can be have unique combination. Finding this relationship is done only if both the columns are non-null and of the type Integer or Smallint. First it tries to find out if the range of the first column is totally contained in the range of the second column. If it is not, then the second column cannot be a foreign key. Otherwise it checks for the inclusion dependency of the first column in the second column. If this checking gives positive result, the application declares the second column as a foreign key of the first column.

Frequency Analysis helps in finding the no. of distinct values in the columns and the no. of time the value is repeated in a column. Domain Analysis helps in finding the columns which does not satisfy the List of Values. For Example Column: Gender, it can have two values either Male or Female. If there is any other value other than these twos, Domain Analysis will find out this.

Single Table Structural Analysis will help user to find out primary key, composite primary key within the table. It will find out all possible combinations of column which can act as a primary key.[5]

IV. CONCLUSION

We have described the functionality of our data profiling tool Data Explorer and the technique to come up with the rules. These rules can be used to eliminate some type of dirtiness or inconsistency. However, coming up with more sophisticated rules can ensure more quality data to be sent to

the target data repository. However this scheme is completely new to the users and the proposed authentication techniques should be verified extensive for usability and effectiveness. This paper explores how data profiling can help determine the structure and completeness of data and, ultimately, improve data quality. The paper also covers the types of analysis that data profiling can provide as well as how data profiling fits into an overall data management strategy.

V. FUTURE WORKS

Data Explorer can be further extended to support unstructured or semi-structured data like flat files, .csv files. It can also be extended to support more platforms. It can also be extended to support other relation data bases like MS Access, MySQL, Sybase etc Time efficient. It can also be enhanced by including Data Quality reports on top of Data Quality Results. There can be mechanism to store the profiling results so that it can be used or referred later at any point of time.

REFERNCES

- [1] (1. Won Y. Kim, 2003) A taxonomy of dirty data. *Data Min. Knowl. Discov.*, 7(1):81–99, 2003.
- [2] Wang R.Y., Lee Y.W., Pipino L.L., Funk J.D. *Journey to Data Quality* The MIT Press 2006
- [3] F. Chiang and R. Miller, “Discovering data quality rules,” in *VLDB*, 2008.
- [4] Data Cleaning: Problems and Current Approaches, Erhard Rahm* Hong Hai Do University of Leipzig, Germany.
- [5] Data Profiling for ETL Processes, Maunendra Sankar Desarkar, IIT Kanpur.
- [6] Hernandez, M.A.; Stolfo, S.J.: Data Cleansing and the Merge/Purge Problem. *Data Mining and Knowledge Discovery* 2(1):9-37, 1998.
- [7] Srikant, R.; Agrawal, R.: Mining Generalized Association Rules. Proc. 21st VLDB conf.