Validating A Big Data Environment Using Various Data Profiling Analysis

K. Makesh Babu, Dr. K. Mohan Kumar

Abstract: Online activities such as e-governance and social networking are producing large volume of transactional data. Due to inconsistent, inaccurate, unreliable and missing of data in government organization, they cannot retrieve the accurate data for current and future use. The organization finds their data as confusing and disorienting. Data profiling is the only solution to this problem. Data profiling is the technique that can process and correct the data from any data source. It provides the most effective way to achieve consistent, accurate and reliable data. In Data profiling multiple analysis are done to correct the data. Such as Single Column analysis, Multi-Column analysis, Multi table analysis and Data rule validation analysis. Various data profiling analysis used to overcome this challenge of inconsistent data along with much needed data quality for analytic results within bounded execution time. Generally, these analyses performed to overcome the causes of data quality issues at Data profiling. These techniques can quickly validate whether the field data is consistent across the data source, and whether the information is consistent with the expectations. This data profiling techniques enables to create a valid data to generate reports for future analysis. This articles focuses on various data profiling analysis to give better data quality.

Index Terms: Big data, Data Profiling, Data Rule Validation, Data Quality, e-Governance, Multi Column, Multi Table, Single Column.

1 INTRODUCTION

1.1 e-Governance
E- governance involves the application of Information and Communication Technologies by government agencies for information and service delivery to citizens, business and government employees. Governments deal with large amount of data. To ensure that such data is put to an effective use in facilitating decision-making, a data warehouse is constructed over the historical data. It permits several types of queries requiring complex analysis on data to be addressed by decision-makers. [1] e - Governance applications awareness is growing among common citizens. It generates large volume of data named Big data. This constant growth of e-Governance data size is referring e-Governance projects as a potential candidate for doing Big Data Analysis.[2] The main objective of E-Governance is to change organization into e-organization. An e-organization needs to focus on the following things.[3]

1. Develop client orientation
2. Improve customer focus
3. Manage customer relationships
4. Business processes improvement
5. Organizing information systems
6. Making better decisions.

1.2 Big data
E - Governance is the primary sources of Big Data are a large volume of both structured and unstructured data. Due to increase of large volume of data it is difficult to store and process in an available data source. The difficulties includes in big data are capturing, restoring, searching, sharing, transferring, analyzing and maintain data quality. The major sources of Big data applications are Health care, Transport, e-
Objective of data profiling
1. To create a valid data
2. To understand the metadata characteristics of the data under purview.
3. To have an enterprise view of the data for the purpose of Master Data Management and Data Governance
4. Helps in identifying the right candidates for Source-Target mapping.
5. Ensure data fits for the intended purpose. It helps to identify the Data issues and quantify them.

This paper explores various data profiling techniques and how their efficiency are measured at various stages.

3 LITERATURE REVIEWED
Preet Navdeep, et al., (2016), focused on role of big data analytics in e-Government projects. This paper deals with different data analytics techniques and their outcomes. Also it explains the major benefits and challenges of using Big data analytics with e-Governance.[2] Ranjith Singh, et al., (2010), presented the classification of Data Quality issues. The purpose of this classification will be helpful for data quality community. Also it explains the causes of data quality problems and issues at various stages. The important issues are data sources and data profiling stages.[3] Wei Dai, et al., (2016), the paper introduced a method for building data quality metrics, and showed how to calculate data quality. In this paper a new data profiling definition and tasks were presented. Also this paper proposed a data profiling tool framework for big data.[5] Felix Naumann, (2013) pointed out the potentials and the needs of modern data profiling. Data profiling tasks compare various techniques both in their abilities and their efficiency. This data profiling techniques are important for the fast growing areas of Big data.[6]

4 METHODOLOGY
Data profiling methodology has two different approaches. A bottom-up approach starts at the most atomic level of the data and moves to progressively higher levels of structure over the data. By doing this, problems at lower levels are found and can be factored into the analysis at the higher level. In top-down approach data inaccuracies at the lower levels may confuse the process and make it difficult to establish the true data rules. This is why a top-down approach will not work effectively in the face of data inaccuracies. Many analysts will correct data inaccuracies at each level before moving to a higher level. This is done to make the higher levels of data profiling more successful.

4.1 Types of Data Analysis
The following are the analysis of data profiling

4.1.1 Structure Analysis
It is also known as structure discovery. It helps to validate the data for consistent and formatted correctly. There are different processes to validate the data. Validating with metadata, Pattern matching and Collecting statistical information process helps to validate the dataset for consistent.

4.1.2 Content Analysis
In this analysis individual elements are looked closely to check data quality in terms of Completeness, Uniqueness, Values Distribution, Range and Pattern.

4.1.3 Relationship Analysis
This analysis is very essential for gaining a better understanding of the connection between data sets. Meta data analysis helps to determine key relationship across different sources of data. [8], [9]

5 DATA PROFILING TECHNIQUES
In data profiling technique both approaches are applied. In general there are four data profiling techniques help to accomplish better data quality: Single column profiling, Multi-column profiling, Multi-table profiling and Data rule validation.

5.1 Single Column Profiling
It scans through a table and counts the number of times each value shows up within each column. This method can be useful to find frequency distribution and patterns within a column of data. It is a general form of data profiling. There are different methods used in this techniques are given below

1. Cardinalities: It identifies number of distinct values that appear with in column
   1) Number of values / Distinct values.
   2) Number of Nulls/ Uniqueness.
   3) Min, Max, Average and Median values.
2. Values Distribution
   1) Detect whether data follows some well-known distribution.
   2) Frequency distribution for value intervals.
3. Data types and Value patterns
   1) String vs. number vs. date vs. complex.
   2) Categorical vs. Continuous.
   3) Semantic domains
   4) Regular expressions
4. Assumption
   1) All values are of same type.
   2) All values have some common properties. [7]

5.2 Multiple Column Profiling
It is made up of two processes: key analysis and dependency analysis. Key analysis used to locate possible primary key in a database. Dependency analysis is a more complex process that determines whether there are relationships or structures embedded in a data set. Both techniques help analyze dependencies among data attributes within the same table. It is used to identify relationship, structure and joint properties embedded in dataset. Also it involves validating data.
integration, indexing, optimization and schema design. The following methods of this analysis is given below

1. Uniqueness
   1) Unique column – only unique values
   2) Key discovery – No Null values
2. Inclusion dependencies
   1) Foreign key
   2) All values of in A are also present in B.
3. Functional dependencies:
   1) Normalization
   2) Two records have same X values; they also have same Y values.
4. Partial dependencies
   1) No perfect hold – clearing only 10% or 1% of the records.
   2) Useful for data cleansing.
5. Conditional dependencies
   1) Useful for cross-source integration

5.3 Multiple Table Profiling
It is also called Inter-table analysis. It uses foreign key analysis, which is the identification of orphaned records and determination of semantic and syntactic differences, to examine the relationships of column sets in different tables. This can help cut down on redundancy but also identify data value sets that could be mapped together. The following are the methods of Multiple Table profiling
1. Data overlap
   1) Duplicate detection – detect multiple representation of the same real-world entity.
   2) Record linkage
2. Orphan records detection.
3. Identify data values sets that could be mapped together.

5.4 Data rule validation
Finally, data rule validation uses data profiling in a proactive manner to verify that data instances and data sets conform to predefined rules. This process helps find ways to improve data quality and can be achieved either through batch validation or an ongoing validation service. Verify that data and data sets conform to the predefined rules/business rules. [6]

6 DATA PROFILING PROCESS
Data profiling process may consist of two phases: In first phase defines the project description and the data structure. The data profiling techniques are executed in phase II.

Phase I:
1. Create the project initiation document. It contains the outline and time of the project.
2. Clearly understanding the document to avoid wasting time on performing unnecessary tasks.
3. Documents shows expectations and requirement of the task
4. Select Analytical tool and statistical tool which allows you to outline the quality of data structure.

Phase II:
1. Analysis of column profiling values which discovers problems with metadata or content quality

2. If table structure is single than relationship between the columns must be profiled. Problem with primary keys and data structure is also been analyzed.
3. Multi table structural profiling, it compares data in between tables, looking out for overlapping values, duplicate values, & analyzed foreign keys.
4. As a data profiling process is very important and necessary as a fundamental thing of extraction process in ETL.
5. Analyzing business rules, decisions analyze team cannot make proper decision in the unclean data which is only cleaned by profiling its value as its first step of cleansing of data.

7 RESULTS AND DISCUSSION
This stage is considered to be more vulnerable to data quality problems as the data is called from various heterogeneous environments. The following Table1 shows the types of data profiling and their methods. Data profiling is the technical analysis of data to describe its content, consistency and structure. At each and every stage of process some percentage of data are cleaned.

<table>
<thead>
<tr>
<th>S.No</th>
<th>Profiling Technique</th>
<th>Methods</th>
<th>% of Records Cleaned</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Single column</td>
<td>Cardinalities</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Value Distribution</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data types and value Patterns</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assumption</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Multi column</td>
<td>Uniqueness</td>
<td>18%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inclusion Dependencies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Functional dependencies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Partial Dependencies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Conditional dependencies</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Multi table</td>
<td>Data overlap</td>
<td>12%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orphan records Detection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mapping</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Data rule Validation</td>
<td>Predefined rules</td>
<td>36%</td>
</tr>
</tbody>
</table>

The following Figure 2. Shows percentage of records cleaned when data profiling process is executed in a data profiling tool using some sample data. It gives only 86% of cleaned data through the above process. The aim of various data profiling analysis is to improve the percentage of data records to be cleaned.
8 CONCLUSION

Data profiling is the crucial first step in data quality. It is important to validate the e-government data for reliability to public. An approach of this data profiling analyzes the data is reviewed valid and complete for e-governance. Government deals with large amount of data to ensure that, such data is put an effective use in facilitating decision making. This data profiling techniques enables to create a valid data to generate report for future analysis. Once the desired results are achieved, the same can be replicated in other sectors of the governments and other organizations.

REFERENCES


