Researchers

Sweety Patel

Department of Computer Science, Fairleigh Dickinson University, NJ-07666, USA

Piyush Patel

Department of Computer Science, Rajasthan Technical University, India

Email-

sweetu83patel@yahoo.com

piyushdpatel79@gmail.com

Why Data Profiling And How?

Abstract:

Data profiling is a process of discovering different anomaly in data which Is broken by data field value. Data anomaly breaks the business rules which are required to work application properly so breaking one rules may include thousands of uncorrected data to databases lead data base to computation option of profiling towards wrong direction. It is step by step method in which value of data filed is review as it is not filtered by standard procedure of the business rules apply by the developer may lead wrong data capturing and final report output generated from that become significantly inefficient to make out feature plan depending on that anomaly data.

Keywords: Data Profiling, data discovery, data assessment & data quality.

Introduction:

Profiling of data is also known as discovery of data, data assessment &data quality analysis. It is a process of doing examination on given source of data and collecting statistics and detail into about is done in data profiling, for data quality improvement it is first step.

Some properties are defined below for the data quality of data

It should be correct with real data.

It should be consistent with real data.

It is unambiguous (contains only one meaning).

Consistent –convert ion must be used as once, to display its meaning

Complete –it has no null value or no missing values

Main fundamental task of the data profiling is not only to identifying the data anomalies and different issues that may require cleanup & conforming process.

Data Profiling Process Is Done As Below

- Data profiling process may consist of many steps, such as: initial document of project is created. It consists of deliverable boundaries and time of project. The developer team should be familiar with this process's for necessary time is saved in doing required process on required data, this document also show what are the expectations and requirements.
- Choose appropriate tools analytical and statistical tools which allow us to make out the outlines of quality of the data structure. Those shows

VOLUME 1 NUMBER 4 (July/Aug 2012)
ISSN: 2277–1891

important element of data and also shows the frequency of the given data element

- Data source must be analyzed.
- Scope of data must be determined.
- Must identity verification process for data patterns and formatting of data in the database.
- Find out multiple coding redundant values, duplicates null vales missing values & other anomalies that can be occurred in the source of data.
- Check the relationship between the primary (PK)and foreign key(FK) and find at how this PK & FK relationship influence the data extraction
- Business rules must be analyzed.

Structure of Database Profiling Process

Considering structure of database profiling process may include:

- Analysis of column profiling values which discovers problems with metadata or content quality
- If table structure is single than relationship between the columns must be profiled. Also problem with primary keys and data structure is analyzed.
- Cross table structural profiling, it compares data in between tables, looking out for overlapping values, duplicate values, & analysed foreign keys.
- As a data profiling process is very important and necessary as a fundamental thing of extraction process in ETL. Data should be analysed because it will help to avoid the other problems which can occur later during performing business.
- Decisions analyser 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.

Field Name	NULL	Missing	Actual	Completeness	Cardinality	Uniqueness	Distinctness
Customer ID	0	0	3,338,190	100.00%	3,338,190	100.00%	100.00%
Account Number	0	0	3,338,190	100.00%	3,254,735	97.50%	97.50%
Customer Name 1	50,072	16,690	3,271,428	98.00%	2,997,864	89.81%	91.64%
Customer Name 2	2,450,670	53,077	834,443	25.00%	798,531	23.92%	95.70%
Tax ID	886,703	41,444	2,410,043	72.20%	2,120,837	63.53%	88.00%
Gender Code	1,204,060	50,264	2,083,866	62.43%	8	0.00%	0.00%
Birth Date	627,019	0	2,711,171	81.22%	25,275	0.76%	0.93%
Postal Address Line 1	196,536	5,193	3,136,461	93.96%	2,886,753	86.48%	92.04%
Postal Address Line 2	2,349,569	42,966	945,655	28.33%	875,578	26.23%	92.59%
City Name	171,517	15,171	3,151,502	94.41%	29,876	0.89%	0.95%
State Abbreviation	723,865	0	2,614,325	78.32%	72	0.00%	0.00%
Zip Code	925,591	0	2,412,599	72.27%	48,731	1.46%	2.02%
Country Code	0	0	3,338,190	100.00%	5	0.00%	0.00%
Telephone Number	515,781	0	2,822,409	84.55%	2,624,840	78.63%	93.00%
E-mail Address	1,204,608	0	2,133,582	63.91%	2,037,570	61.04%	95.50%

Figuere1. Counts and percentages for each field that summarize the value of its content characteristics. NULL – count of the number of records with a NULL value for a given query. Missing – number of records with a missing value is counted (i.e., non-NULL absence of data, e.g., character spaces). Actual – number of records count with an actual value (i.e., non-NULL and non-Missing). Completeness –it gives the percentage calculated as Actual divided by the total number of records. Cardinality – it provides the count of the number of distinct actual values. Uniqueness – percentage uniqueness is calculated as Cardinality divided by the total number of records. Distinctness – percentage distinctness calculated as Cardinality divided by Actual.

Other Important Considerable Points While Profiling

- Reusability of the existing data must be identified.
- Resolve the missing value of data field.
- Sort out erroneous values from the database source.
- Data's current state must be identified and data quality issues are determined to develop standards.
- Find out, search out formats of the data.

VOLUME 1 NUMBER 4 (July/Aug 2012)
ISSN: 2277–1891

- Sort out patterns of the data.
- Business rules are revealed.
- Appropriate data values are identity.
- Define transformation for maintaining data validity
- Make out a report for column minimum, maximum, average, median, mean, mode, variance, co-variance, standard derivation & outliers.
- Check whether business rules are applied to entire data base on thoroughly system.
- Make out report results in various formats including .PDF, HTML, XML and CSV.
- Providing data profiling history.

Types of Data Profiling

- String values in a column with distinct length and percentage of rows in the table that each length represent.
- Example: profiling of column of US state codes, which should be always 2 length character but more than that may cause profiling of those data.
- Percentage of null value in the column
- Example: Profiling in postal code /zip code column shows a high number of percentage ratio in the missing value of that field had data profiling as a first step in extraction process of ETS cycle.
- Percentage of regular expression which may be occurred in a column.
- Example: A pattern of phone number column may contain different format patterns like; source pattern [919]999—0000. & other format [919]9990000,[919]999 0000, or [919]999-0000.all above shows same phone number but because of the different pattern of formatting, data is not captured in query if it is used only one format to retrieve data from database. As of that, data must be entered in specific format as an input to any field of table to make, other dependent process easier to work on that.
- There should be some fixed criteria to set a minimum, maximum, average value of any filed of data in database.
- Example: Birth date value must not exceed current date value.
- There should be required to get out each distinct value of each field with how many times it occurs into the database.
- Example: US State filed value must not exceed with 50 different state values.
- Primary key must be chosen in most of any table to make out data non-duplication in records of that table.
- Dependency of values of one table or in more columns of more than two tables
- Example: Two different states may contain same zip code for two different cites.
- Example: Some value in product id column of a sales table have no corresponding value in the product id column of the product table may had digging process to find out proper solution that time profiling is most important criteria to find out deficiency in given data value.

Automated “Bottom -Up” Data Analysis

Complex SQL queries are run by experienced people from technical staff to do profiling. But limitation or condition on query make itself as a limitation on the data profiling .beyond that condition data is profiled by bottom up approach in which each value of each field must be profiled in a systematically way & also stepwise so, no hidden data is remain to profiled and less change to get uncorrected result.

VOLUME 1 NUMBER 4 (July/Aug 2012)
ISSN: 2277–1891

Business Collaboration: With the User Friendly Interfaces

Controlling of the process through user friendly interfaces so anyone can manage at any technical level skill, with this methodology, business & technology professionals can sit at the same table and openly discuss the data issues which are occurred during technical problem solving procedure.

Immediate Cleansing Of Data and Repair

There is no need for delay once data are profiled, cleansing become easy to work on that.

Features of the Data Profiling Task

- Wildcard columns: When profile request is configured the task accept the (*) wildcard in place if column name. This simplified the configuration and makes it easier to discover the characterized of unfamiliar data when task runs, every column of the task is profiled that has an appropriate or concern data type.
- Quick Profile: We can select quick profile to configure a task quick profile to configure a task quickly. A quick profile discovered a data in a table or view by using all the default profiles and default settings.

Conclusion

Data profiling is not only limited to databases but it include different files and also other applications from where data is accessed with data exploring methods ,drill down in to individual data sources and specific records are viewed to make appropriate operation on that data. perform statistical data profiling on organizations data, ranging from simple record counts by category; to analysis of specific text numeric fields .Apply custom business rules to identify a record which cross the threshold value as per the required value or value which fall inside or outside of defined ranges.

References

[1] Nong Ye, The Handbook of Data Mining (Lawrence Erlbaum Associates, Mahwah, NJ. Publication, 2003).

[2] Jiawei Han and Micheline Kamber, Data Mining:Concepts and Techniques (Morgan Kaufmann Publishers, University of Illinois at Urbana-Champaign).

[3] Bharat Bhushan Agarwal and Sumit Prakash Taval, Data Mining and Data Warehousing (Laxmi Publications, New Delhi - 110002, India).

[4] Ralph Kimball,Joe Caserta, Data Warehouse. ETL Toolkit. Practical Techniques for. Extracting, Cleaning,. Conforming, and. Delivering Data (Wisely Publication,Inc).

Author Details:

Sweety Patel
Department of Computer Science, Fairleigh Dickinson University, USA

Piyush Patel
Department of Computer Science, Rajasthan Technical University, India