International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 11 | Nov -2016 www.irjet.net p-ISSN: 2395-0072

Survey on Data Profiling and Data Quality Assessment for Business Intelligence

Vivek Teegalapally[1], Kiran Dhote[2], Vamsi S. Krishna[3], Shubham Rao[4]
[1]*Student, Computer Science, SKN Sinhagad Institute of Technology, Maharashtra, India*
[2]*Student, Computer Science, SKN Sinhagad Institute of Technology, Maharashtra, India*
[3]*Student, Computer Science, SKN Sinhagad Institute of Technology, Maharashtra, India*
[4]*Student, Computer Science, SKN Sinhagad Institute of Technology, Maharashtra, India*

Abstract - *The amount of data generated every day in the modern world is tremendous. This provides an opportunity for businesses to make informed decisions to expand their growth and productivity using data mining techniques. In this paper, two loosely coupled systems are designed to aid the data mining processes required for Business Intelligence. The two systems are Data profiling, Data quality assessment. These systems are developed as plug-ins which could be used by specific Business Intelligence application systems/ projects/ solutions. All of them are designed based on RESTful API architecture using pure HTTP/POST calls.*

Key Words: *Data Mining, RESTful API, Business Intelligence, Data Profiling, Data Quality Assessment*

1. INTRODUCTION

In the modern world, the amount of data generated every day is huge. It is estimated that 2.5 Quintillion bytes of data is generated every day. This provides a big window of opportunity for business to analyse the data and make informed business decisions. Because of the size of the data, businesses have turned to data mining techniques and tool to perform Business Intelligence.

There is a saying in the field of data science that 70% of a data scientist's time is spent cleaning and preparing the data. These tasks aren't part of the core Business Intelligence but are crucial to achieve accurate results.

Technologically, business intelligence systems consist of data warehouses, data marts, OLAP and data mining techniques as its components. The data in the warehouses is accessed remotely with the web service architectures.

SOAP, which stands for Simple Object Access Protocol, has been traditionally used to communicate with web services. It relies exclusively on XML to provide messaging services. But it has some problems such as the XML used to make requests and receive responses in SOAP can become extremely complex.

REST is a lightweight alternative to the problem. It communicates based on pure HTTP calls. Also, the data to be received can be represented in many different formats, such as JSON, XML, and YAML.

2. Data Profiling

Data profiling is the process in which the data is examined in an existing data source (e.g. a file or a database) and obtaining summaries and statistics about the data. [1]

Data profiling uses different kinds of strategies which are descriptive such as Mean, Minimum, Maximum, Standard deviation, Frequency and Variation. The metadata information includes Data types, Length, Discrete values, occurrence of null values and uniqueness and abstract type recognition. [2] This data can be used to discover problems like misspelling or wrong value representation and duplicates.

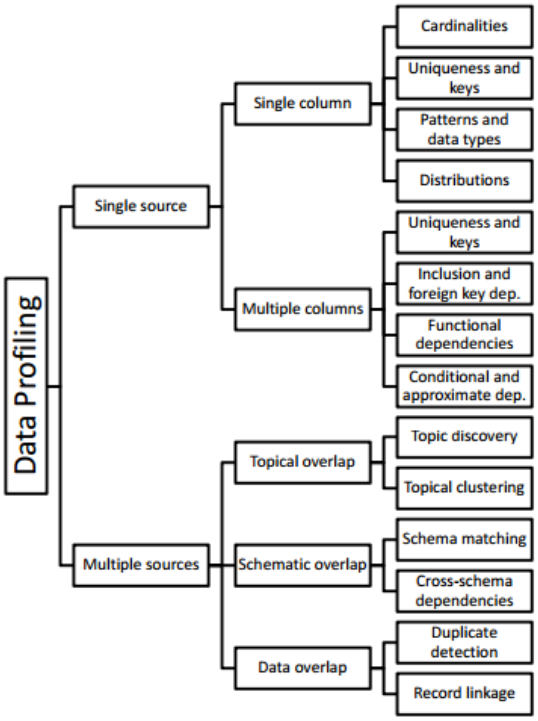

Fig -1: Classification of Data Profiling tasks

Data Profiling provides various benefits such as improving the quality of data, shortening the process cycle of big projects and improving the understanding of data for users. Data profiling helps the user to find data quality rules and requirements that will support a more thorough data quality assessment in a later step.

The use cases of Data Profiling are:

- Accuracy and reliability
- Data cleaning
- Data integration
- Scientific data management
- Data analytics

3. Data Quality Assessment

Data quality assessment is the scientific process and statistical evaluation of data to determine if they possess the quality for business purposes and they are of usable quality.

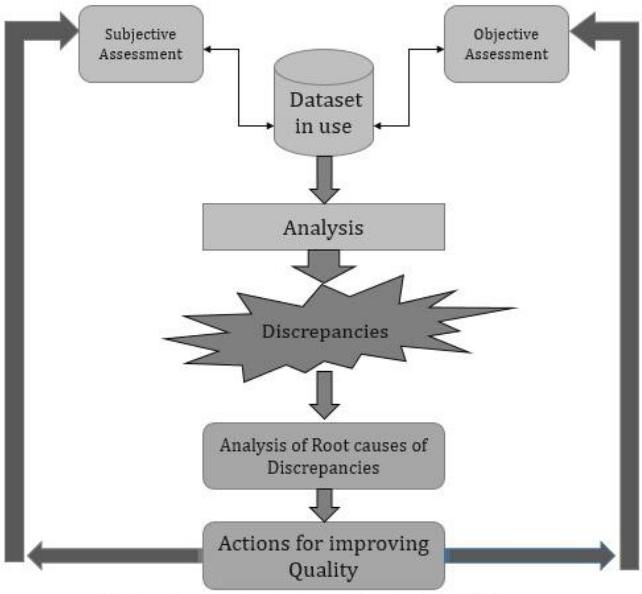

Fig -2: Data quality assessment in practice

Data quality assessment displays the issues with the data which allow the business company to plan for data cleaning and enrichment. This maintains the quality standards of the data. The quality issues such as inconsistent data, missing data and inconsistent data are easy to identify and correct, but much more issues should be looked upon with much more defined processes.

Data Quality Assessment processes are associated with some practices and prerequisites. The five points of data quality:

- Reliable and Accurate
- Serviceability
- Accessibility
- Methodological soundness
- Integrity

To use the subjective as well as objective methods to increase the data quality requires the following steps:

- Subjective and objective tests on the quality assessments of data and Serviceability.
- Obtaining the discrepancies and determining their causes.
- Taking steps to improve the data and enrich it.

The use and analysis of data must be based on precise and high-quality data, which is a necessary condition for generating value from the input data.

Many business issues can be related directly, to a situation where quality of the data doesn't meet the expectations. With the understanding of the data, information, and the ways in which information value decreases when data does not meet the basic expectations, Different types of business impacts related with poor information quality can be seen, And discuss ways about the impact on costs due to poor data quality.

4. RESTful API

The Representational State Transfer (REST) is the abstraction of the elements of the distributed hypermedia system architecture. REST does not consider the details of implementation and protocols in order to maintain the roles of the parts, the constraints on their interactions with other parts and their representation of data elements. It imposes constrains on the parts, connectors and data that define the web architecture, and thus its behavior as a network-based application. [3]

REST makes it possible to make highly scalable network applications. Companies in the league of Google, Facebook and Twitter make use of this to provide services with which users can take advantage of the company's established architecture without the need to know the inner working. E.g. Google Maps and Facebook graph API.

Typically RESTful APIs are run on top of HTTP protocol.

For an application to be called RESTful, it should adhere to the following constraints:

- Client-Server model
- Stateless
- Cacheable
- Uniform Interface
- Layered System
- Code-on-Demand

Resources in REST architecture represent the entities in the application's domain. Each resource or collection of resource has a unique URL. Clients work with the resource representation. Client sends a HTTP request to server and the server sends a HTTP request back.

5. Proposed System

Fig -3: Architecture of the proposed system

The proposed system is divided into three parts: Client side which consists of the GUI of module, REST API server which handles the HTTP request from client and backend server.

All the components of the system are going to be written using Python libraries, Pandas and scikit-learn for data processing, matplotlib for data visualization and Flask as a REST API framework. The database will be a relational database.

The client side consists of a GUI for each of the module. The RESTful server takes request from the client side. The backend server consists of the python code processing the data and the database holding the data.

5.1 Functional overview

To use the services of the Business Intelligence module, the user logs into the system via a client side user interface. After the authentication is done, the user is presented with a visual representation of the connected database and various actions which can be performed on it.

The client side interacts with the REST API server sending requests to access resources. The interaction between them happens purely based on HTTP calls as defined in the RESTful API paper. The client sends an HTTP request to the API. The API parses the request and tells the data processing module to process and give back the required data. The REST API then sends back an HTTP response containing the information requested by the client. The client then represents the data in a rich visual form.

5. CONCLUSION

Data profiling and Data quality assessment are important steps of any Business Intelligence process. The proposed system enables any person to make use of the data profiling and data quality assessment modules. The system is based on RESTful API architecture and the heavy processing work is off-loaded to the server.

The modules are going to be open-sourced.

REFERENCES

[1] Theodore Johnson (2009), "Data Profiling", in Encyclopedia of Database Systems, Springer, Heidelberg

[2] David Loshin (2009), "Master Data Management", Morgan Kaufmann Publishers, ISBN 9780123742254

[3] Roy Thomas Fielding (2000), "Architectural Styles and the Design of Network-based Software Architectures", University of California, Irvine