Name - Nihar Sudhanshu Limaye

Program – MCM

Email id – nihar.limaye3@mail.dcu.ie

Project title - Developing Data Value Analytics of a RDB for an Online Retailer

Student number - 18210876

Supervisor name – Dr. Rob Brennan

Developing Data Value Analytics of a RDB for an Online Retailer

Nihar Sudhanshu Limaye
School of Computing,
Dublin City University,
Dublin, Ireland
nihar.limaye3@mail.dcu.ie

***Abstract*— This research paper defines a new approach to assessing the business value of data in a relational database. Data in today's world is an asset and yet to assign a value to it is a challenge. Many previous papers on data value have created awareness about it, also they have highlighted some methods to find data value like the use of data value dimensions and metrics, use of a data value capability maturity model (CMM) but most of the papers failed to explain the technical approaches to analysis of relational databases for value. The research presented here includes an automated system which considers the results from surveys, personal interviews and compares them with results obtained from the relational database (RDB) and provides a data value assessment. There are very few pieces of literature available today on data value assessment and the research given in this paper uses RDB data for value assessment makes it distinctive compared to other data value assessment methods.**

Keywords—RDB, information systems, CMM (keywords**)**

I. INTRODUCTION

In today's world, data volume is increasing day by day due to the introduction of multimedia and social media. The daily data generation is huge as compare to earlier years hence to manage such kind of data is very crucial for everyone. Some kind of governance must be present to decide which data to keep and which to discard. The data value plays deciding role regarding data storages as it can rank the data depending upon its importance and then the decision on data deletion, curation or improvement can be made easily.

The research discussed in this paper mainly focuses on finding data value. For this, a case study of MyVolts an online retailer is taken under consideration. The literature available today regarding data value propose some of the methods like using a metric-based approach, using surveys and personal interviews, etc. but most of them failed to mention the actual practice to get a value. The research is done in this paper mainly focuses on three research questions as follows,

1. To what extent can the data value of a relational database (RDB) data source of an organization be accurately measured using a metrics-based approach?
2. Which predefined data value metrics are most effective at predicting a value for the MyVolts online retailer use case?
3. Which specifically tailored metrics will add the most to the characterization of data value the MyVolts online retailer use case?

This research uses some techniques already available in literature like the use of metric-based approach, use of surveys. The metric-based approach is a quantitative method that develops an algorithm to decide the value of the corresponding data. As this research with data value is new only a few dimensions and metrics are present and those are loss of information, Market value, Time, Usage[2][4] in addition to that this research uses some tailored dimensions like security, quality, volume[7] to ensure the correctness of data value.

What are Dimensions?
It is a category designed while considering user-oriented views towards data for finding a data value.

How Dimensions and Metrics finds a value?
In this research one of the Dimension is Volume. The volume gives rise to metrics such as more data more value and more joins more value if these two things are satisfied by a database table then its value is high compared to other tables.

The system has main components such as data value surveys stored in Google forms, data cleaning tool (Only to remove junk data) to make data more readable, a relational database (RDB) to store data obtained from MyVolts, visualization tool (Tableau) to display the results and on top of that a heart of the system which is a python script communicates and manages all other tools discussed above.

The research includes an automated system which considers the results obtained from surveys, personal interviews, and RDB results and compares them to provide data value.

The rest of the paper consists of a literature review (Section II) which comments about previous studies done to find data value, a use case of the research (Section III), design of the system(Section IV) with an explanation to every block present in it. An evaluation (Section V) mentions working of a system and a final summary of this paper in the conclusion section and Future enhancements.

II. LITERATURE REVIEW

This section provides an analysis of existing approaches for assessing data value in databases and summaries relevant parts of pre-existing work that are relevant to the research. Further sections explain the gaps present in the existing

literature and evaluate the possible measures to resolve it with a new methodology.

To evaluate data governance value-driven methods are useful [3]. The data available today is massive but to decide which one is useful some data governance techniques will be useful. The data value is a part of data governance. Brennan et al.2018 created awareness about value-driven data and sets a benchmark for getting a data value of a database.

The data is getting created day by day due to the introduction of multimedia, social media, etc. But to decide which data is valuable very few techniques are available till date. In practice, information has a notional value only, people think it is valuable but they can't put a number on it [4]. Moody et.1999 identifies "Laws of information" to make understand data value. But they didn't provide any concrete metrics to calculate a quantifiable data value for a specific data asset. Their research only includes a single approach towards data valuation rather than the inclusion of two or more approaches to get concrete results and only considers surveys, questions but didn't look into the technical aspect to balance the results. The system designed in this research is a perfect mixture of non-technical methods as well as technical aspects to generate data value.

The use of Capability Maturity Model (CMM) enables an organization to identify and measure the current state of their data value monitoring processes and shows how to take steps to enhance value monitoring to exploit the full data value potential in their organization [9]. This paper [9] includes a CMM that tries to identify specific metrics but uses open data to analyze it. The research takes some defined metrics mentioned in this paper and tries them to fit in a relational DB environment.

The main method to get a data value is finding dimensions of information values [2]. Sajko, Mario & Rabuzin, Kornelije & Bača, Miroslav. (2006) gives an idea of dimensions and their corresponding metrics to be used for value assessment. The main dimension discussed in Sajko 2006. Are Rebuilding, Quality, Legislative, Time and Market value but these metrics can vary according to use case given. The research appends some more dimensions and provides concrete evidence of data value assessment.

To conclude this section, there are very few literatures available today to find a data value. The research includes an automated system which considers the results of both surveys and results obtained from the relational database (RDB) and compares them to provide data value. Below table consists of dimensions taken from literature to assess data value.

Referenced paper	Dimensions	Metrics
[5]	Usage	Recency, Degree of usage
[6]	Rebuilding	Similarity, Amount of new information
[4]	Cost, Volume, Utility	cost of acquisition, cost of selling, increase revenue
[7]	Volume	database with highest no. of records
[8]	Security	Encrypted data, Less access to a data
[2]	Rebuilding,Usage,Legislative,Time,Market Value	Amount of new info, highest records, Unique data, Shelf life

Fig 1. Dimension/Metrics from literature

III. USE CASE & REQUIREMENTS

MyVolts is a small Irish company, based in Dublin, focuses on providing modern and innovative power solutions. It has one main database and the Amazon seller info is another data source.
For this case study the main DB splits into sections based on groups of related tables and we examine two of them Pricing, Sales along with Amazon DB.

Why MyVolts is considered for this research?

The MyVolts is a Small Medium Enterprise (SME) and an online retailer so they process huge amount of data. But they have some constraints regarding the storage of a data and this research will give them a quick evaluation of their data and help them to manage their resources towards crucial data.
As mentioned in the literature review following are the key requirements of this research:

It requires multiple datasets (with at least a thousand records in it). In this case, MyVolts Amazon DB, Pricing DB, and Sales DB.

To store data obtained from MyVolts a relational database tool(MySQL) running on a local windows machine is required. The MyVolts also uses a RDB hence to simulate the working environment similar kind of database is required in this research.

The dimensions already present to measure a data value such as market value, time, usage is needed in this use case also some additional dimensions as well as their corresponding metrics will help to determine a value.

Also, basis metrics a series of manual data value assessment surveys and interviews have to be recorded with the help of Google forms to generate a CSV file which will act as an input to a RDB. The data obtained from MyVolts has some junk (unknown characters, distorted values) so to clean it

open-source data cleaning tool is required such as OpenRefine.

To display the results of the research a tool is required. The tableau desktop is the tool which will be used. As it is compatible with both python and MySQL. It will display the data value of a database table.

The most important part is Data value analysis tool, it is a tool which is a python script designed to manage all the blocks of the designed system such as Google forms, database, and visualization tool. It compares the output generated from database queries to survey results and provide value and ranks database tables.

The following are the main characters in the research:

Domain experts: He/she is an expert from MyVolts organization and helps to answer survey questions and provide insights about the business process within MyVolts.
DB expert: This person shares a database and helps the user to understand the database and also responsible to answer technical surveys.
User: The person responsible to gather all inputs and runs the system to generate data value.

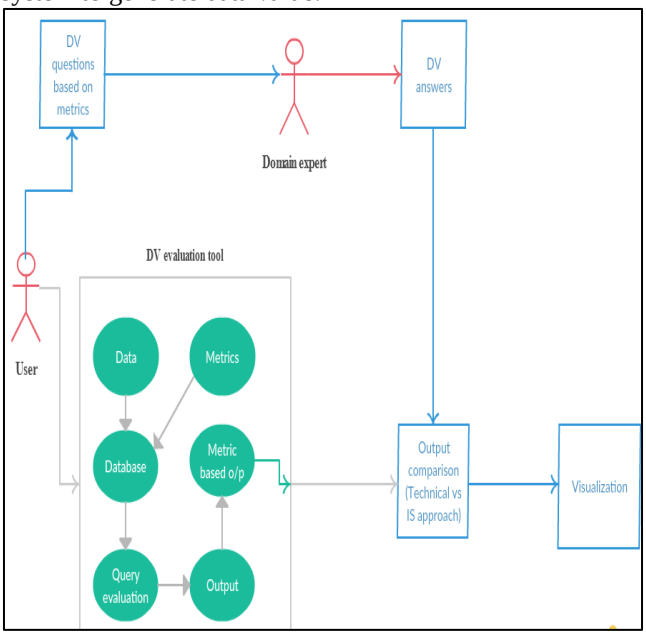

Fig 2: Use case of the research

IV. DESIGN

As mentioned in requirement multiple tools including RDB, Python script and Visualization tool are needed to design a system.
The tableau and MySQL work well with each other[1]. As multiple articles are present today explaining MySQL and Tableau connections.

Also python is one of the famous scripting language available in market. It communicates well with MySQL to fetch and insert records from its database[2].
All the tools used in this research are compatible with each other to produce good results.

To find a data value only SQL language is not sufficient as it has multiple restrictions like it can't produce some good analysis like comparing two results and incrementing the count for a particular value, allocating variable in run-time etc.so use of such language was not good for getting automation.

Why the system is hosted on local rather than on Cloud?
All the components are hosted on a local machine. So it will only run on a single machine. Hence to make available data value analytics tool to the public it must be hosted on a cloud. With tools like MySQL and Python can be hosted on a cloud, not tableau. So to maintain the same components on the cloud environment was a challenge hence it is deployed on a local machine. This is a small challenge encountered while designing a system.

The system in this research comprises of six parts:

1. Metric design:
Metrics is a measure related to the corresponding dimension and helps to determine the value of a data. The metrics are taken from literature where it was used earlier to measure data value (Refer to Fig.1).
A list of metrics are designed based on dimension present to measure value from literature and on this basis questions in surveys and queries of SQL are designed.

2. Data cleaning tool:
It is useful to remove unwanted, unreadable data and avoid distortion in a report.

3. Google forms:
It records surveys and generates CSV file which is input to a MySQL database. The questions asked for domain experts are as follows along with metrics and dimensions.

Dimension	Metrics	Possible questions generated
Rebuilding	Reproduction of new data	Which table has a capacity to produce new data?
Security	Encryption	Which database table contains encrypted values?
Volume	High Record Count	Which database table stores the highest records?
	Interconnection	Which database table can join with most of the other data?
Usage	Number of writes in a day	Which database table is used more frequently in a day?

[1] https://help.tableau.com/current/pro/desktop/en-us/examples_mysql.htm

[2] https://dev.mysql.com/doc/connector-python/en/connector-python-example-connecting.html

Quality	Duplicate records	Which table has least duplicates?
	Null records	Which database table has least null values?
Legislative	Confidential data	Which data loss will create legal consequences if organisation losses it?
Time	Shelf life reduces	Which database table information value falls in course of time?
Market value	Unique data	What database is unique for your organization?

Fig 3. Survey Design

4. Relational Database:
A tool stores data provided by MyVolts in tables mainly Amazon DB, Sales and Product along with file generated by Google forms.

5. Data visualization tool:
The tableau report displays names of databases along with its rank compared to other databases along with its value.

6. Data value analysis tool (Python Script):
The script communicates with the database and visualization tool to produce data value.

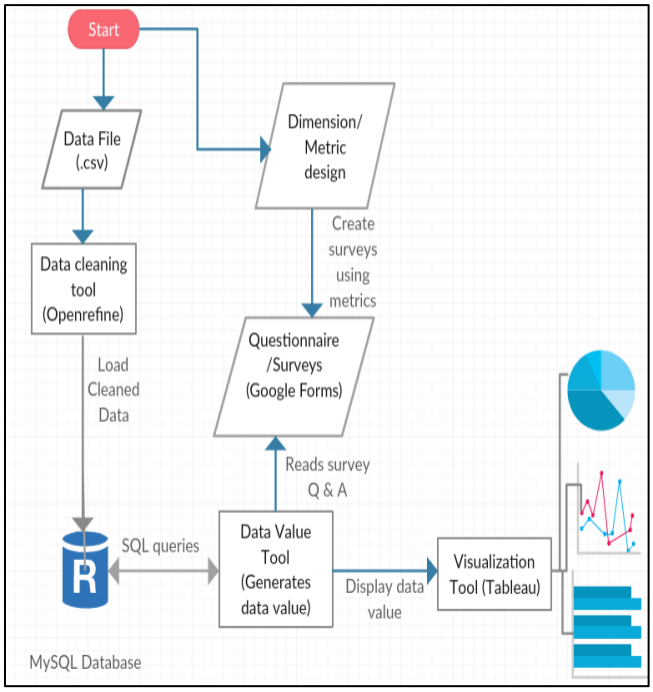

Fig 4. System design

Tools used:

Database: MySQL workbench 8.0CE[3]
It is open-source RDB similar with which is in use in MyVolts. Also the interface is easy to understand and runs smoothly on a windows machine.

Data cleaning: OpenRefine[4]
It is developed by Google and compare to other data cleaning tool it gives interface and suggestions while cleaning data. It is very simple to use.

Visualization tool: Tableau desktop 2019.2[5]
It is used to designed reports on the data value. Also, it is compatible with RDB MySQL and connects with it easily. It is one of the powerful visualization tool available in the market.

Script: Python code running on Jupyter notebook[6]
It is an interface which provides a platform to execute python. It is compatible with both Tableau and MySQL also it requires less memory to run on a local windows machine.

V. Evaluation

This section includes a complete evaluation process for this research.

The hypothesis is considered for this research is "Whether the metrics selected can identify most valuable data for the organization?"

As mentioned in the literature some previous methods or principles to determine data value consists of approaches using an information system, use of surveys, defining dimensions and corresponding metrics but most of them only considered the one side of a coin i.e. Non-technical approach, most of them failed to provide automated approach towards finding a data value. This research uses an automated approach which includes a combination of surveys as well as technical methods to get a data value.

A) Procedure/Method to get a data value used in research

This research mainly focuses on the automated system which compares survey results to database output and ranks a database table after finding its value. To get the value multiple steps needs to be performed starting from data gathering till data value display. The system is a combination of multiple tools such as data cleaning, data value tool, database, visualization tool and they all work gradually to get a data value.

B) Datasets

The dataset in this research is data acquired from MyVolts. They shared data to the user with a categorization like Amazon DB, Sales DB, and Pricing DB. These are three separate tables each contains a thousand records. Before

[3] https://www.mysql.com/products/enterprise/database/
[4] http://openrefine.org/
[5] https://www.tableau.com/products/desktop
[6] https://jupyter.org/

loading the data in a MySQL DB some small amount of cleaning is done using a cleaning tool such as open refine.

Amazon Database- It is a table with table name as amazonuk_listings and it includes electric products displayed on the Amazon website for sale by MyVolts. It gives a price, a quantity of a product, last updated price of an item, etc.

Pricing Database- The table contains pricing of items available with MyVolts. The organization has a strategy of lowering a product price if it is not sold for a longer duration. So it contains price data in addition to that discount percentage to get a new price.

Sales Database- MyVolts also has its website available to sell products. The sales from their website stored in this table with personal information of a customer.

All the tables mentioned above are considered in this research for finding a data value of each of them and finally rank them basis their value.

C) Metrics and Dimensions

As mentioned earlier in literature the research includes some of the dimension already present to get data value. Also new dimensions and metrics are generated specifically for this case study. Sajko, Mario & Rabuzin (2006) suggests some of the dimensions like Rebuilding, Legislative and Usage.
But also with it, three more dimensions are included for the case study such as Volume, Security and Quality.

D) Analysis methods

After getting dimension and corresponding metrics a set of question are designed with multiple-choice options to select a database table name (refer fig 3). The Google form link is shared with domain experts of MyVolts to record the answers. After the survey gets completed the questions and answers are downloaded and saved it in.csv file.

The python script is a data value tool designed for the MyVolts use case. This script has four main jobs.

i) Load the Google form questions and answers into a new database table.
The script loads the CSV file in a google response database table with columns such as a question, answer, metrics, and dimension.

ii) Query the database according to dimension and metrics
The database already contains tables of Amazon, Sales, and Price which is obtained from MyVolts for analysis. Script fires some SQL queries specifically designed in line with survey questions to get a result.
E.g. Survey Question- Which database stores the highest records?
Survey Answers- Amazon DB
SQL query: Query to get the highest record count of database table.

Likewise multiple queries are designed according to metrics and the results of comparison are given below:

Dimension	Possible questions generated	Survey Ans	Tool result	Match
Rebuilding	Which table has a capacity to produce new data?	Pricing DB	Pricing DB	Yes
Security	Which database table contains encrypted values?	Sales DB	Sales DB	Yes
Volume	Which database table stores the highest records?	Amazon DB	Amazon DB	Yes
	Which database table can join with most of the other data?	Pricing DB	Pricing DB	Yes
Usage	Which database table is used more frequently in a day?	Amazon DB	Amazon DB	Yes
Quality	Which table has least duplicates?	Pricing DB	Pricing DB	Yes
	Which database table has least null values?	Pricing DB	Pricing DB	Yes
Legislative	Which data loss will create legal consequences if organisation losses it?	Sales DB	Sales DB	Yes

Time	Which database table information value falls in course of time?	Amazon DB	NA	NA
Market value	What database is unique for your organization?	Pricing DB	NA	NA

Fig 5. Data value tool results

iii) Give points to the database after each successful output and rank them accordingly
If both Survey answers and tool result matches then the script add points against the database table. (refer fig 5)
E.g. The survey answer, as well as DB query, says amazon DB has highest record count. Hence the value of Amazon DB table is now 1. If database tables satisfy more metrics more points are added against it. Finally, the highest point database table is most valued.

iv) Connect with a visualization tool to display database ranks and value.
The visualization tool is connected to a MySQL database via a python script. It displays the results which can be easily understandable. The report display is as follows:

The report shows the value of each database table and now clearly the Pricing DB has more points compared to the other two. Hence we can conclude that Pricing data is more valuable than the other two database tables.

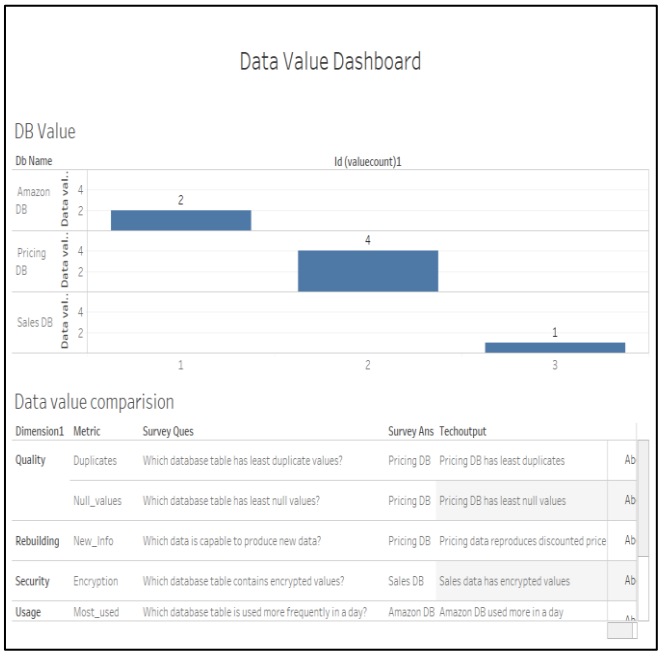

Fig 6. Data value report

VI. Conclusion

The available literature with data value analytics is limited and hence to find a value is a bit challenging task but this research used some of the literature and tried to find data value using an automated approach.

The research satisfies the first research question as "To what extent can the data value of a relational database (RDB) data source of an organization be accurately measured using a metrics-based approach? " nearly 70 % of metrics used to measure data value gives positive results and it can be said that research was successful for capturing data value for a RDB data source. Metrics of Dimensions like Rebuilding, Security, Volume, Usage, Quality, and Legislative are successful for measuring data value. However metrics of dimensions like Time, Market value failed to measure data value for an RDB data source because Market value needs contextual information and Time can't be measured without proper data. It would have been a successful attempt for time dimension to measure if data contained specific information to measure it.

Also, research answers the second research question “Which predefined data value metrics are most effective at predicting a value for the MyVolts online retailer use case?” The predefined metrics are five (Rebuilding, Usage, Legislative, Market value, Time) amongst eight discussed in this research and from them, two of them perform badly with RDB i.e. Market value and Time. Rest three (Volume, Security, Quality) tailored metrics designed specifically for MyVolts use case gave a perfect match between survey answers and tool output and produced data value for RDB data source.

In the end, the value of a relational database (RDB) data source accurately measured for a given use case of MyVolts with predefined dimensions/metrics like Rebuilding, Legislative and Usage and with tailored dimensions like Security, Volume and Quality.

VII. Future Enhancement

The research used limited tools and technology and hence the future enhancement can be the inclusion of more tools. Some of the possible enhancement are mentioned below:

The Relational DB is used in this research hence the future enhancement will be to use the data value tool with other databases like a data warehouse, cloud database, and non-RDB.

Only eight dimension and metrics are considered for this research, more can be designed and integrate with the system to get concrete results.

References

[1] Even, Adir & Shankaranarayanan, Ganesan. (2005). Value-Driven Data Quality Assessment..

[2] Sajko, Mario & Rabuzin, Kornelije & Bača, Miroslav. (2006). How to calculate information value for effective

security risk assessment. Journal of Information and Organizational Sciences. 30.

[3] Attard, J. and Brennan, R. (2018). Challenges in Value-Driven Data Governance.

[4] Moody, Daniel, and Peter Walsh. “Measuring The Value Of Information: An Asset Valuation Approach.” Seventh European Conference on Information Systems (ECIS’99) (1999): 1–17. Web. 12 July 2017.

[5] Chen, Ying. “Information Valuation for Information Lifecycle Management.” Second International Conference on Autonomic Computing (ICAC’05). N.p., 2005. 135–146. Web.

[6] al-Saffar, S, and G L Heileman. “Semantics-Based Information Valuation.” 2008 4th International IEEE Conference Intelligent Systems. Vol. 1. N.p., 2008. 6–58. Web.

[7] Wijnhoven, Fons, Chintan Amrit, and Pim Dietz. “Value-Based File Retention.” Journal of Data and Information Quality 4.4 (2014): 1–17. Web. 12 July 2017.

[8] Tuemmler, B. (2019, 05 23). *The Impact of Information Governance on Cybersecurity*. Retrieved from www.nuix.com: https://www.nuix.com/blog/impact-information-governance-cybersecurityhttps://www.nuix.com/blog/impact-information-governance-cybersecurity

[9] Rob Brennan, Judie Attard, and Markus Helfurt, Management of Data Value Chains, a Value Monitoring Capability Maturity Model, 20th International Conference on Enterprise Information Systems (ICEIS), Portugal, 21-24 March 2018, Olivier Camp, Joaquim Filipe, 2018. http://www.tara.tcd.ie/handle/2262/82277