From Data Quality to Predictive Big Data Quality

Rhemat, I., Chountas, P. and Dr Francois Roubert

This is a copy of the author’s accepted version of a paper subsequently published in the proceedings of the 12th IEEE International Conference on Intelligent Systems - IS'24. Varna, Bulgaria 31 Jul - 29 Aug 2024 IEEE.

The final published version will be available online at:

https://ieeexplore.ieee.org/Xplore/home.jsp

Proceedings of 2024 IEEE 12th International Conference on Intelligent Systems (IS)

On Predictive Big Data Quality

Ismael Rhemat,[1,2]
[1] School of Computer Science & Engineering, University of Westminster
115 New Cavendish Street, London W1W 6UW
[2] Mellitah Oil & Gas Co., Tripoli-Libya
Dahra Kabira St., PO Box 314, Tripoli, Libya
e-mail: Ismael.Irhemat@mellitahog.ly

Panagiotis Chountas[1], Francois Roubert[1]
[1] School of Computer Science & Engineering, University of Westminster
115 New Cavendish Street, London W1W 6UW
e-mail: {p.i.chountas, F.Roubert}@westminster.ac.uk

***Abstract*— Businesses process use data to produce services and or goods, and most importantly for developing business strategies as part of data driven change management approach. The goodness of the data used is an important factor regarding process lead time, customer satisfaction and service quality. The main contribution of this work is to define a big data quality conceptual framework related to business process performance as part of an online analytical, predictive mining framework. Here we consider the domain of Mellitah Oil & Gas (MOG), in Libya.**

Keywords- Big Data Quality, Business Processing, Predictive mining

I. INTRODUCTION

The Quality of a Data object may be understood as the degree to which data satisfy the requirements [1], defined by the main enterprise. Several conceptual models in the context of data quality are available. We concentrate the state- of-the-art on model quality. Conceptual Models (CM) are the abstraction of the enterprise domain under consideration [2]. Generally, the following three objectives are associated with the CMs [2] [3],[4]:

- Meet the stakeholders requirements;
- Provide a conceptual representation of the entrerprise doemain, and used as the design for the implementation and evolution of digital business systems.

The task of requirements specification corresponds to a formalization task by means of an appropriate modelling formalism. Formalization implies structure and refining of the acquired knowledge. The result is a mental model, called the conceptual schema, that corresponds to the modelled part of the world. In developing that mental model of reality an analyst must determine what knowledge to represent, how to express and organize it and what constraints to introduce in order to keep it a consistent model of the real world.

Because of the cognitive nature of requirements modelling, the formalisms employed are known as Conceptual Models. In [5], see (Figure 1), a CM recommended to formalize a data process quality framework based mainly on [6,7], [8], [9], [10], [11] and [12,13] among others.

In this paper the main objectives, are defined as follows: Specify a CM that should be used to formalize data-process quality mechanism in the era of big data, as part of an online analytical and predictive mining environment (OLAM)?

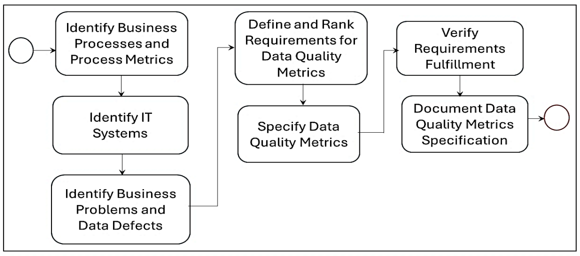

Figure 1: DQ Process and Metrics

How beneficial such a data quality mechanism for the case of Oil and Gas industry? Here we consider the case of Oil & Gas industry sector in Libya in the domain of Mellitah Oil & Gas (MOG).

II. CONSIDERATIONS

The main aim of conceptual modelling is to capture the knowledge about the UoD and represent it in such a way as to enable the developer to reason about this knowledge, communicate this understanding of the UoD to end users for purposes of validation and specify the allowable structures of and transitions on the information base, [14]. In [14] is suggested that quality business processes (BP)/facts cannot be defined directly, but their result is determined by the dimensions in the sense of Golfarelli et al [15], involved, when answering an OLAM query.

At the same time Golfarelli et al. proposed a CM named dimensional fact model (DFM) for a Datawarehouse repository development, that consists of a set of Facts/ Business Processes in the sense of Otto et al. The elements are facts, dimensions, and hierarchies. In (Figure 2) we defined the fact Sale in terms of its dimensions (time, location, product) measurable in quantities sold and total returns.

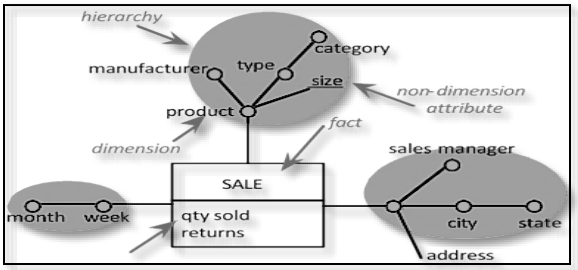

Figure 2. Multi-dimensional fact sale.

So, the question here is can we develop an integrated data quality conceptual framework for a Big data repository that integrates the Business process analysis needs as identified by Golfarelli et al, while incorporating or preserving the Business-Oriented Data Quality Model, proposed by Otto et al [5]?

III. Data Quality Modelling Considerations

The DQ meta-model is based on a (GQM) [6] to start managing quality from the stakeholders needs, based on the the ISO/IEC 25012 standard. It builds on the notion of a Fact. Our conceptual base proposes a set of predefined dimensions, attributes, metrics etc. Stakeholder can define the quality attributes and the corresponding metrics for its evaluation.

At the same time using the main concepts of the initial proposed quality metamodel we define the metamodel of the DFM model as below:

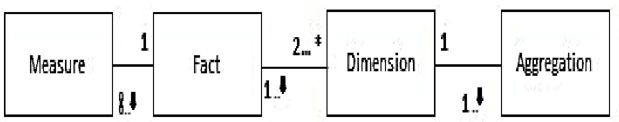

Figure 3. The (DFM) Metamodel

In this sense a Fact represents a business process i.e. sale a dimension determines the factors adopted in defining facts, measures are quantitative models to determine business process performance while aggregation represent the hierarchical organization of a data domain behind a dimension.

IV. Data Quality Metamodel

Meta-data modelling of a particular model is concerned with the product dimension and captures the static aspects of a technique. It tries to assist the process of verifying the good quality of a specific conceptual schema. The attempt to formalise both the process and the product of a modelling technique is a great one to undertake, but only then there exists a complete and forma specification of a conceptual model.

There are two main categories of metamodels. The first one deals with the steps of a modelling language and is referred to as meta-activity modelling. In this, by applying metamodeling to the way a technique works, the dynamic aspects of this technique are formally captured, i.e. the way that the technique must be applied in order to build a specification. Its use leads to an understanding of the procedure of the requirements modelling technique and this knowledge can be used when the technique is applied to a specific situation.

The second category of metamodels deals with the products of the formalism, the actual models, and is referred to as meta-data modelling.

Meta-data modelling of a particular model is concerned with the product dimension and captures the static aspects of a technique. It tries to assist the process of verifying the good quality of a specific conceptual schema. The attempt to formalise both the process and the product of a modelling technique is a great one to undertake, but only then there exists a complete and formal specification of a conceptual model

The main blocks are facts/business processes each among *N* dimensions or entity types (E_1, E_2,…,En) as a set of associations. Each fact instance is a truth proposition in the real world of business processes about one or more business process instances. A Data Quality Metric is a quantitative measurement and goodness estimation of a given dimension.

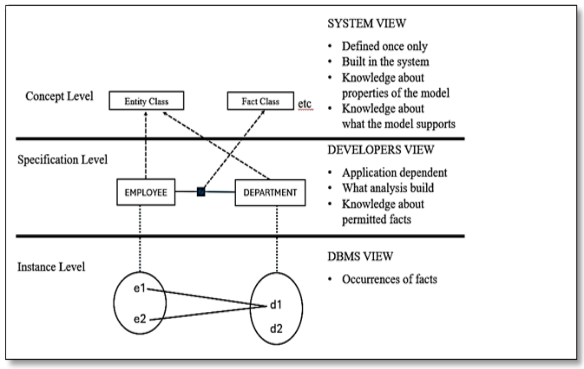

Figure 4. The Levels of a Fact

Each fact can have a number of Performance and Quality Measures, compromising of at least two Dimensions one of which is usually the Time dimension. Dimensions may be common between dependent business process/facts in terms of performance evaluation.

A DQ metric must be specified as a class-subclass hierarchy for each dimension of interest. The reason for this modelling choice is that the cardinality of a dimension |D| is significantly less than of a business process/fact F. Moreover, a fact F is defined as the product of the involved dimensions {D_1… D_N}, noted as follows:

$$F = D_1 \times D_2 \times \ldots \times D_N \quad (1)$$

Therefore, applying the DQ metric at the dimension level ensures less computational effort, especially as a number of dimensions (i.e., location, customer) may be shared between dependent business processes i.e. (sales and shipment) in different nodes.

The DQ metamodel created for this paper is presented at a high level as a conceptual data model in Figure 4: This model presents the fundamental set of classes and associations that should be defined and implemented to properly support the essentials of DQ querying.

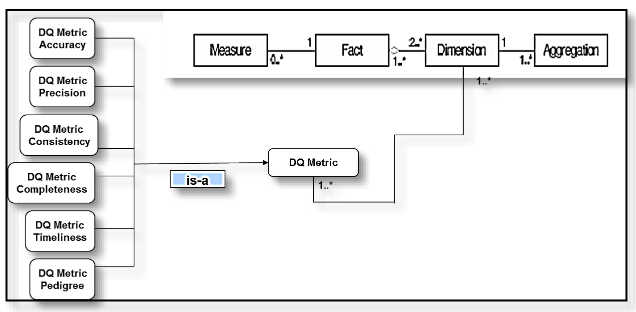

Figure5. Data Quality Metamodel

V. Big Data Quality Matrix Representation

A. Big Data Quality Metrics

In this regard, BDQ metrics can be defined given a set of data stored and located in different Nodes referred to as 'Si', each Node (Site) refers to a data wrapper (or the site) where data is maintained and the 'I' signifies the number of Nodes (sites) where database servers are located. Thus, the Big Data process platform (BDP) is a set of 'Si' of 'ith' Nodes.

$$|S| \rightarrow N \text{ nodes in BDP} \quad (1)$$

$$S = \{S1, S2, \cdots, SN\} \quad (2)$$

The expected Fact 'F' are calculated at the level of number of fetched records (rows) 'm' and attributes (columns) 'n' as follows:

$Fe = \{F11, \cdots, Fmn\}$ the expected dataset:

Fm refers to expected records.

Fn refers to expected attributes (columns) (3)

Meanwhile, after calculating the expected records and attributes by the above equation during performing BDQ, some records and columns (attributes) are received from the processed dataset(s):

$Fr = \{FR11, \cdots, FRmn\}$ the received dataset:

Frm refers to received records.

Frn refers to received attributes (columns) (4)

Based on that, the total received dataset during the BDQP can be dimensionally computed by the following formula:

$| Fmne | = Fm * Fn$ (5)

$| Fmnr | = Frm * Frn$ (6)

Referring to previous section which introduced and defined a mechanism that BDQ be processed in four steps to approach to calculate BDQ dimensions.

VI. IFS Foundations of Data Quality Metrics

This section presents an Intuitionistic Fuzzy Set (IFS) based, definitions of BDQ metrics based on [17],[18] [19], [20] [21].

Each element of an Intuitionistic fuzzy set [24], [25], [26] has degrees of membership or truth (μ) and non-membership or falsity (ν), which don't sum up to 1.0 thus leaving a degree of hesitation margin (π).As extension to the classical definition of a fuzzy set is given by

$$\tilde{A} = \{< x, \mu_{A'}(x) > | x \in X\}$$

where: $\mu A(x)$ [0, 1] is the membership function of the fuzzy set , an Intuitionistic fuzzy set A is given by

$$A = \{< x, \mu_A(x), v_A(x) > | x \in X\}$$

such that $0 \leq \mu_A(x) + v_A(x) \leq 1$ and $\mu_A(x)\ v_A(x) \in [0, 1]$ and $\mu A(x), vA(x)$ denote a degree of membership and a degree of non-membership of $x \in A$, respectively.

For each IFS in X, we will call $\pi A(x) = 1 - \mu A(x) - vA(x)$ an Intuitionistic fuzzy index (or a hesitation margin) of x A which expresses a lack of knowledge of whether x belongs to A or not. For each $x \in A$, $0 \leq \pi_A(x) \leq 1$.

With reference to our big data platform (BDP) and Fact 'F' we define the following IFS data quality parameters:

$$\mu_F = |Fr| / |Fe| \quad (6)$$

$\mu_F \leq 1$ is the number of fact objects that are common in the expected and received data sets. In the case where $Fr = Fe$ then $\mu_F = 1$ since no missing data or Null Values in Fr.

$$\pi_F = |NullValues| / |Fe| \quad (7)$$

$\pi_F \leq 1$ is the number of Null Values in the received data set. In the case where zero null values exist then $\pi_F = 0$ since no Null Values in Fr

$$v_F = 1 - \mu_F - \pi_F = 1 - (|Fr| + |NullValues| / |Fe|) \quad (8)$$

$v_F \leq 1$ is the number of Missing Values in the received data set, excluding the null values. In the case where $(|Fr| - |NullValues|)/|Fe| = 1$ no missing values exist in Fr thus $v_F = 0$.

VII. Completeness Metric

Generally, data completeness is one of most important dimensions of data quality. If data without missing information is considered complete, see [17].

DQ completeness (*Compl*) of Big Data at the fact level F can be modelled as follows:

$$Compl(F) = 1 - VF, \quad Compl(F) = Compl(F) * 100 \quad (9)$$

DQ completeness per Fact attribute level can be modelled as follows:

$$Compl(Fri) = 1 - VFri \quad (10)$$

$$VFri = 1 - \mu Fri - \pi_{Fri} = 1 - (|Fri| + |NullValues| / |Fei|)$$

$$Compl(Fri) = Compl(Fri) * 100$$

– DQ completeness of Big Data at over a time span can be modelled as follows:

- Using Fact level F

$$Compl(F) = \frac{Aggr\ (Compl(F))}{d\ |Hr - H|} \quad (11)$$

Where, 'Hr' represents the time dimension and its hierarchical levels for a given domain and 'd' duration of time. $d\ |Hr - H|$ for the month level, is the total days for which data received.

- Using the Fact attribute level

$$Compl(Fri) = \frac{Aggr\ (Compl(Fri))}{d\ |Hr - H|} \quad (12)$$

VIII. Data Quality Model Accuracy

According to different concepts discussed in many research fields, the data accuracy dimension [22] is defined differently. A number of different definitions also exist in [12],[22],[6],[7].

We utilize the definition of accuracy as in in [23], and list of metrics related to accuracy. DQ Accuracy of Big Data at the record level can be modelled as follows:

- Accuracy(Fi) = Fri - Fei
- IF Fri $\subseteq$ Fei, snd $\mu_{Fi} \leq \mu_{Ei}$ and $v_{Fi} \leq v_{Ei}$ then Accur = 1 otherwise 0. Hence:

$$\text{Accur(F)} = \frac{\sum_{j=1}^{n} \text{Accur } j}{n} \quad (13)$$

- DQ Accuracy (*Accur*) of Big Data at the attribute level can be modelled as follows:

$$\text{Accur (Fri)} = \frac{\sum_{i=1}^{m} \text{Accur } i}{m} \quad (14)$$

- DQ Accuracy (*Accur*) of Big Data at the overtime level can be modelled as follows:

$$\text{Accur}_{(F_{Hr})} = \frac{\sum_{i=1}^{|Hr-1|} \text{Accur}}{d|Hr-1|} \quad (15)$$

IX. DATA QUALITY MODEL TIMELINES

Timelines means how the information is available right when it's needed by a business, with reference to the data quality dimension of time [23]. If your information isn't ready exactly when you need it, it doesn't fulfil that dimension. As an example, a business requires financial information every quarter about its revenues, if the required data is ready at the time that supposed to be then it's timely.

– DQ timeliness of Big Data at the fact level can be defined as follows:

IF Fri $\subseteq$ Fei, snd $\mu_{Fi} \leq \mu_{Ei}$ and $v_{Fi} \leq v_{Ei}$ then the result of the calculated timeliness of Big Data at fact level assuming time expected (tie) = time received-(tir), then Timeliness rec = 1, otherwise = 0.

– DQ timeliness of Big Data and overtime at record level of a hierarchy manner can be defined as follows:

Timeliness

$$\text{FHr} = \frac{\sum_{i=1}^{|F|} \text{Timeliness(F)} \quad < \mu\text{Fi} , v\text{Fi} >}{|F|\text{Hr}}$$

(16)

X. DQ MODEL UNIQUENESS

A distinctive "Unique" information means that there is only one occurrence of data that appears in its data mart . Data duplication is a repeated fact about something stored in a database, i.e., "John Smith" and "J. Smith" might be address to the same name of same person [23].

– DQ Uniqueness of Big Data at the Fact level of a hierarchy sequence can be modelled as follows:

IF Fri $\subseteq$ Fei, and $\mu_{Fi} \leq \mu_{Ei}$ and $v_{Fi} \leq v_{Ei}$ then the uniqueness is measured at record level and no match found then set it to 1 otherwise set it to 0.

The following equation measures the uniqueness at record level in a hierarchy sequence.

$$\text{Uniqueness}_{(F_{Hr})} = \frac{\sum_{i=1}^{|Hr-1|} \text{Uniqueness (F)} < \mu\text{Fi} , v\text{Fi} >}{d|Hr-1|}, \quad (17)$$

XI. DATA QUALITY MODEL VALIDITY

Based on [16], the dimension validity is used to measure the degree to which business rules or definitions are precisely encoded. By other means data must be defined appropriately and representative of the business metrics it describes.

Data validity is defined as the accumulation of three DQ metrics completeness, accuracy and timelines [24].

– DQ Validity of Big Data at the fact level can be modelled as follows:

Based on the validation criteria which depending on a set of criterion elements defined to be used to validate the data of the selected records or attributes. However if creation result of all C={C1,C2, …., Ck} is TRUE then Validity Val = 1, otherwise = 0 which means data is not valid. DQ Validity overtime at the record level can be modelled as follows:

$$\text{Validity}_{(F_{Hr})} = \frac{\sum_{i=1}^{M(Hr)} \text{Validity (F)} < \mu\text{Fi} , v\text{Fi} >}{\text{MHr}}$$

(18)

XII. DATA QUALITY MODEL CONSISTENCY

The consistency dimension has different definitions introduced by numerous research groups. Here are some of consistency dimension definitions as follows [22] :

– "Consistency captures the violation of semantic rules defined over tuples or tables [6], or records in a file.". Consistency rules can either defined on attribute-value level or on tuple-level. In alignment to the accuracy and completeness metric calculated, consistency can be calculated on table - or database-level as the arithmetic mean of the tuple-level consistency.

– However, if no difference found in both datasets either sent or received, then the Consistency Val at *i* iteration is set to True (1), otherwise Consistency Val is set to False (0). At end the Consistency of a dataset can be defined as Validity.

XIII. DATA QUALITY MODEL USABILITY

The term data usability defines data can be used to confirm and inform business decisions. To measure Usability three data quality metrics altogether such as Completeness, Accuracy and Timelines by applying the following equation.

– DQ Usability of Big Data at the record level can be modelled as follows:

$$\text{Usability}_{(F)} = \frac{\text{Compl(F)} + \text{Accu(F)} + \text{Timelines(F)} < \mu\text{F} , v\text{F} >}{3}$$

(19)

The big data quality operators as proposed above are defined for the *Procurement (purchasing, warehouse) and Maintenance* of "Mellitah Oil & Gas" main business processes. The goodness of the operators can be verified by comparing the value of quality operators against the predicted values via regression analysis.

XIV. THE MELLITAH OIL AND GAS GROUP CASE STUDY

The initial phase of this research was to develop a general BDQ framework that provides complete and clear definition of its components. The next goal of this is to identify, collect, analyze, and evaluate the quality metrics for business process master data; that are of high value ones and have business impacts for the Mellitah Oil & Gas group. In MOG case, there are two critical datasets of big size comprise two essential types of MOG data related to *Procurement (purchasing, warehouse) and Maintenance*

.

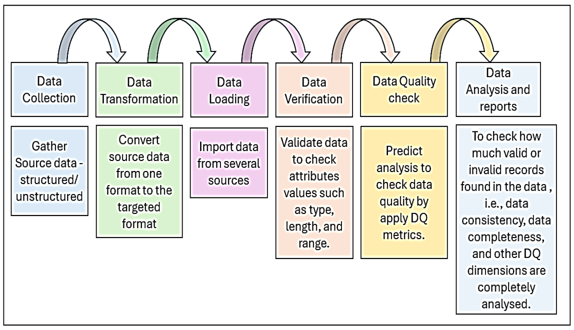

Figure 6: Data Validation Process

The main target is to adopt such a quality framework a set of rules and policies defined to be applied during the evaluation process of DQP before/after of the pre-processing stage.

XV. The Mellitah Oil Case study and Predictive Data Quality Modelling

Here we defined how to calculate each data quality parameter with the help of two well-known methods Regression and Deep learning. Regression analysis is used to derive the set of weights for the Deep Network Architecture.

Figure 7 presents a neural network, that allow us to predict the completeness metric values for the present Sales/Purchasing data with only one hidden layer with three neurons. i.e. 3-3-1.

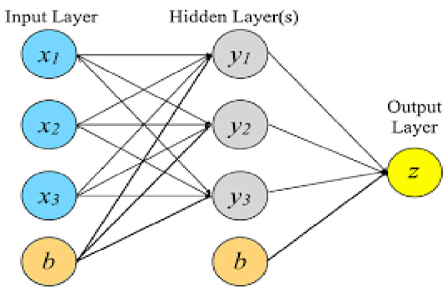

Figure 7: Neural network 3-3-1 for IFS Completeness

XVI. Data Quality Metric Trial

We present the IFS completeness data quality metric. In this regard the following steps took place:

- data validation process and checklist are applied (see Figure 6).
- Secondly, datasets values imported, to validate and check the quality level of data values of the selected datasets.
- Thirdly, predictive mining and visualization of the overall data quality (Figure 7), at the node level; and completeness of Purchasing and Materials Accounts data set separately, (Figures 8), using Power BI are presented.

The correctness of the models can be checked by calculating the value of parameters with the help of the proposed models in this case study compared with predicted values by ANN/regression analysis..

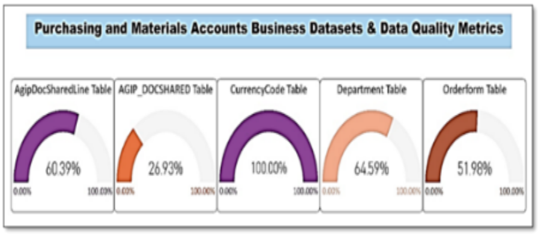

Figure 8: Purchasing and Materials Accounts, node level DQ

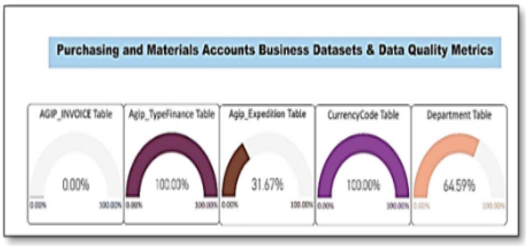

Figure 9: Purchasing and Materials Accounts DQ completeness

XVII. Conclusion

Big data quality can be achieved through conceptual modelling at the levels of metamodels and models. This paper puts forward a big data quality meta-model for evaluating the quality of conceptual designs before implementation. The proposed big data quality meta-model is generic, and can be mapped to different types of conceptual models (ER models, UML diagrams etc.)

Big data quality parameters/metrics defined that can be useful for measuring and predicting data quality as part of predictive dashboard utility empowered by ANN's and Regression analysis techniques, that can deal with crisp and imprecise data. Work reported in [27] is only dealing with crisp data as the authors proposed framework and operators can not cope with imprecise data

The definition of the data quality metrics with reference to the time dimension on a big data repositories and evolving ecosystems due to changes in the data domain, paves the way to the next question, which is to what extent close world assumption (CWA) is valid? Is Open World Assumption OWA valid? What are the implications for Big Data Quality Metrics and their interpretation under OWA?

REFERENCES

[1] I, Taleb M. Serhani, C. Bouhaddioui R. Dssouli “Big data quality framework: a holistic approach to continuous quality management”. Journal of Big Data 2021;8. doi:10.1186/s40537-021-00468-0.

[2] P. Chountas, I. Petrounias “Modelling and Representation of Uncertain Temporal Information” Requir. Eng. 5(3): 144-156, (2000)

[3] B. Otto, “How to design the master data architecture: Findings from a case study at Bosch.” International Journal of Information Management 32.4 (2012): 337-346.

[4] B. Otto, K. M. Hüner, and H. Österle. “Identification of Business Oriented Data Quality Metrics.” ICIQ. 2009.

[5] B. Carlo, and M. Scannapieca. “Data quality: concepts, methodologies and techniques”. Springer, 2006.

[6] B., Carlo et al. “Methodologies for data quality assessment and improvement.” ACM Computing Surveys (CSUR) 41.3 (2009): 16-26.

[7] L. P. English, “Improving data warehouse and business information quality”: methods for reducing costs and increasing profits. Vol. 1. New York: Wiley, 1999

[8] W. L, Yang et al. “AIMQ: a methodology for information quality assessment.” Information & management 40.2 (2002): 133-146.

[9] D. Loshin, “Master data management”. Morgan Kaufmann, 2010.

[10] D.Loshin, “The practitioner's guide to data quality improvement”. Access Online via Elsevier, 2010.

[11] R. Y Wang, D. M. Strong, and L.M. Guarascio. “Beyond accuracy: What data quality means to data consumers.” J. of Management Information Systems 12.4 (1996): 5-33.

[12] R. Y Wang, P. Martin. Reddy, and B. Kon Henry. “Toward quality data: An attribute-based approach.” Decision Support Systems 13.3 (1995): 349-372.

[13] B., Caldiera, V. R. G., H. D.. Rombach, The Goal Question Metric Approach. Encyclopedia of Software Engineering - 2 Volume Set, pp 528-532, John Wiley & Sons, Inc. (1994). Available at http://www.cs.umd.edu/users/basili/papers.html

[14] M. Golfarelli, M. Dario, R. ,Stefano: “The Dimensional Fact Model: A Conceptual Model for Data Warehouses”. Int. J. Cooperative Inf. Syst. 1998, 7(2-3): 215-247

[15] L. Chun X. Gao, T. Chuanqi “Big Data Validation Case Study” Big Data Service April 2017: 281-6 doi:10.1109/BigDataService.2017.44.

[16] L. Ehrlinger, W. Wöß “A Survey of Data Quality Measurement and Monitoring Tools”. Front. Big Data 2022;5. doi:10.3389/fdata.2022.850611.

[17] B. Saha, D. Srivastava Data quality: The other face of Big Data. 2014 IEEE 30th International Conference on Data Engineering 2014:1294-7. doi:10.1109/ICDE.2014.6816764.

[18] I. Caballero, M. Piattini CALDEA: a data quality model based on maturity levels. Third International Conference on Quality Software, 2003:380-7. doi:10.1109/QSIC.2003.1319125.

[19] X. Chunli, J. Gao, T. Chuanqi “Big Data Validation Case Study”. Big Data Service 2017: 281-286

[20] T. Haegemans, Monique Snoeck, Wilfried Lemahieu: “Towards a Precise Definition of Data Accuracy and a Justification for its Measure”. ICIQ 2016: pp 144-156

[21] C.W. Fisher, Eitel J. M. Lauría, C. Matheus: “An Accuracy Metric: Percentages, Randomness, and Probabilities”. ACM J. Data Inf. Qual. 1(3): 16:1-16:21, 2009

[22] L. Pipino, Y. Lee, and R. Y. Wang. “Data quality assessment.” Communications of the ACM 45.4 2002, 211-218

[23] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, A. Mustapha “Data quality: A survey of data quality dimensions”. 2012 International Conference on Information Retrieval & Knowledge Management 2012, 300-4.

[24] K. Atanassov, “ Intuitionistic fuzzy sets”. *Fuzzy Sets and Systems*, 20(1), 1986, 87–96.

[25] K. Atanassov, “Remark on the intuitionistic fuzzy sets”. *Fuzzy Sets and Systems*, 1992, 51(1), 117–118.

[26] K. Atanassov, *“Intuitionistic Fuzzy Sets: Theory and Applications”*, 1999, Springer, Heidelberg,

[27] Desai, Khushali Yashodhar, "Big Data Quality Modeling And Validation" (2018). *Master's Theses*. 4898. DOI: https://doi.org/10.31979/etd.c68w-98uf