Purity: a New Dimension for Measuring Data Centralization Quality

Lander Bonilla
High Performance Architecture, Tecnalia, Basque Research & Technology Alliance (BRTA)
Spain
lander.bonilla@tecnalia.com

Maria José López Osa
High Performance Architecture, Tecnalia, Basque Research & Technology Alliance (BRTA)
Spain
mariajose.lopez@tecnalia.com

Josu Diaz-de-Arcaya
High Performance Architecture, Tecnalia, Basque Research & Technology Alliance (BRTA)
Spain
josu.diazdearcaya@tecnalia.com

Ana I. Torre-Bastida
High Performance Architecture, Tecnalia, Basque Research & Technology Alliance (BRTA)
Spain
isabel.torre@tecnalia.com

Aitor Almeida
DeustoTech, DeustoTech, University of Deusto
Spain
aitor.almeida@deusto.es

Abstract

Data has become an asset for companies, originating from various sources, such as IoT paradigms. It is crucial to safeguard its life cycle using suitable, scalable, and effective technologies, like those enabled by cloud computing models. However, in order to extract value from this data, complementary processes of collection, refinement, cleaning, or modeling, among many others, are required. Furthermore, organizations greatly vary in their methodologies and approaches to handling data, which further emphasizes the need for standardized techniques. In this regard, data management methodologies promote the adoption of the various dimensions of data quality in order to ensure the reliability of data across different systems and processes. The main contribution of this manuscript is the proposal of a new data quality dimension, coined purity, to measure the importance of the data in a processing pipeline topology. As a result, organizations can better guarantee the quality of their datasets in order to raise the success of data-driven endeavors within organizations. The proposed methodology is validated in an urban mobility use case.

CCS Concepts

• **Computer systems organization → Cloud computing**; • **Information systems → Data access methods**.

Keywords

Data Quality, DataOps, Centrality, Computing Continuum, Big Data

ICCBDC 2024, August 15–17, 2024, Oxford, United Kingdom

ACM ISBN 979-8-4007-1725-3/24/08
https://doi.org/10.1145/3694860.3694862

ACM Reference Format:
Lander Bonilla, Maria José López Osa, Josu Diaz-de-Arcaya, Ana I. Torre-Bastida, and Aitor Almeida. 2024. Purity: a New Dimension for Measuring Data Centralization Quality. In *2024 8th International Conference on Cloud and Big Data Computing (ICCBDC 2024), August 15–17, 2024, Oxford, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3694860.3694862

1 Introduction

In the era of digital transformation, the adoption of cloud computing has offered organizations scalable and flexible infrastructure to store, process, and analyze massive volumes of data [32]. Cloud paradigms provide unparalleled opportunities for flexibility, cost-effectiveness, and security [7, 35]. However, amidst the advantages, a critical challenge emerges: the quality of the data residing in these cloud environments.

The term data quality is defined as the suitability of the data for the use case, emphasising its relative and dynamic nature, the context, and the requirements that depend on and may change over time [29]. In the context of urban mobility data, as cloud services continue to evolve, incorporating technologies such as machine learning and artificial intelligence, the role of high-quality data becomes even more important. The success of predictive models and intelligent algorithms hinges on the quality of the training data they receive. Poor data quality not only undermines the performance of these advanced technologies but also introduces biases and inaccuracies in the model performance that can affect the final decisions of the models [10].

Unfortunately, real-life data is often dirty[1], which negatively impacts the accuracy of the insights that can be obtained from that data [22]. There are numerous challenges associated with data management specially in this era, which are only aggravated by the growing size of data generated per year [28]. These challenges range from the security, privacy, and infrastructure [24]; to dependency from other organizations, lack of communication between teams, and increased responsibilities of the data pipeline owner [27]. In

[1]Data that is out of date, incorrect, incomplete, not integrated, or duplicated, among other things.

the following section, we elaborate on the problem we address in this manuscript.

1.1 Problem Statement

The success of data-driven decision-making and analytics in the cloud continuum is heavily reliant on the quality of the underlying data in the field of urban logistics. Recognizing the difficulty in mandating fixed quality standards for data providers, the decision has been made to shift the quality calculation to a central node in the cloud [11].

Despite the potential benefits, the absence of standardized and stringent data quality measures in cloud paradigms raises serious concerns. Inaccuracies, inconsistencies, and incompleteness within datasets can have far-reaching implications, impacting the reliability of analytical outcomes [21, 40]. Compromising the integrity of decision-making processes and diminishing the overall performance of cloud-based applications [14].

Data quality issues in cloud environments can manifest at various stages of the data life cycle, from the point of data ingestion to extraction and analysis. Challenges such as data duplication, format inconsistencies, and outdated information can proliferate, leading to a cascade of downstream effects that deteriorate the trustworthiness of insights derived from cloud-hosted data [1, 31]. The inefficiency in data can lead to long and redundant access times, also increasing operating costs associated with bandwidth and storage.

1.2 Contribution

This paper aims to delve into the importance of data quality in the computing continuum in an urban mobility use case. Emphasizing the establishment of robust data quality standards, protocols, and governance frameworks within the cloud ecosystem for organizations. Addressing these challenges is essential not only for ensuring the accuracy and reliability of data analytics but also for fortifying the foundation upon which critical business decisions are made.

To this end, the main contribution of this manuscript is the proposal of a new data quality dimension, coined 'Purity', which evaluates how pure different datasets are in comparison to others in the network. The purity of a dataset is measured by its significance in interacting with other datasets within the same network, as well as by how important its values are for the rest of the datasets. Furthermore, this dimension will be useful for predicting the quality of a dataset, which depends on the origin of its data. This dimension of quality will be validated in an urban mobility scenario.

The rest of the paper is organized as follows. Section 2 provides a brief introduction of the background and the related work is presented in Section 3. In Section 4, the workflow for the proper measurement of data quality is explained. Section 5 provides an overview of the new quality dimension purity, while Section 6 goes into further detail about the mathematical estimation for it. Section 7 portrays the validation scenario. Finally, the conclusions and future works are drawn in Section 8.

2 Background

In this section, we have delved into aspects of data life cycle and related paradigms that must be known to understand the contribution presented in this article on data quality.

2.1 Data Centralization on Cloud

Data centralization brings together all data into one place so it can be more effectively managed and accessed[13]. As businesses rely on a larger number of data sources than ever before, the importance of having a centralized approach to store and manage it has never been greater. The storage architectures that allow such centralization are several, the most well-known in recent times being Data Warehouse[39] and Data Lake[26]. The cloud computing model has allowed a democratization of these architectures. Unlike traditional data storage solutions, cloud-based options offer great scalability and flexibility at a reasonable cost. Besides, centralizing data increases collaboration among teams and ensures that everyone has access to validated, complete datasets, modelled in a unified manner. Data silos make it impossible to gain a clear, unified view of business data [30]. In the work of data engineers, significant effort is invested in reconciling disparate data sets, and there are many inefficiencies in not establishing ways to standardize, unify, and reproduce these data governance phases. In this context, a large part of the challenges is always associated with activities related to data quality[4], from the methodology in which it is measured, to the dimensions or indicators that reflect the reality of the data and the transformations that have suffered. It is important to highlight that in this work we are framed in a cloud computing model, and in the specific problem of centralized storage architectures, which aim to solve and manage the data quality activity from the central repository that represent the unified storage of all data sources. The problem of quality measurement in federated or fully distributed environments is beyond the scope of this work, and is also a topic of special interest today, for example in the mobility domain[16].

2.2 Existing Data Quality Dimensions

Data quality management is an extremely important process for organizations to tackle the problem of poor data quality, establishing a set of best practices for improving this quality and enhancing the value of the data and hence the outcomes obtained from them. Data quality dimensions[25] describe the characteristics by which the quality of the data is measured. Since the quality of the data is different for every organization, the metrics to measure differ from another. Different articles and studies define several of them, but the six most used dimensions are the following: [6]:

- Consistency: data is uniform, accurate, and coherent across diverse datasets within the company.
- Accuracy: the data represents the "reality" that the organization wants to analyze.
- Completeness: all the data the company needs is available and findable.
- Timeliness: the organization should have the data at the specific moment they are needed.
- Uniqueness: one specific data appears once in the storage system.
- Validity: the format and type of data are as expected.

3 Related Work

The usual definition of data quality is the idea to which the data meets the expectations of data consumers based on their intended use of the data[5]. However, a significant challenge arises because

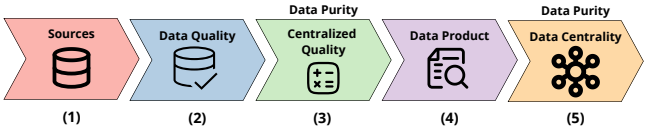

Figure 1: Data Purity Methodology Flow

the intended use of data is often not fixed and can vary depending on the specific use case. In decentralized environments oriented toward exploratory analysis, this complicates the development of appropriate methodologies and technologies to measure the quality of the data from its origin to its consumption.

We present an analysis of the main related works in the literature regarding existing tools and frameworks for quality measurement in decentralized storage environments. Reviewing the current literature, we have found the following studies that we consider interesting in our work. In [2], M. Altendeitering et. al. proposed a software reference architecture for data quality tools that guides organizations in creating state-of-the-art solutions. The work [17] presented a new big data quality framework in which four new dimensions are considered integrity, accessibility, ease of manipulation, and security. Whereas, in the paper [20], the researchers tried to generalize the quality assessment operations by providing a new ISO-based declarative data quality assessment framework (BIGQA), that supports data quality assessment in different domains and contexts. In [38], I. Taleb et. al. introduced a new concept called Big Data Quality Profile, where the intention is to define quality outline, requirements, attributes, dimensions, scores, and rules related to quality dimensions. Finally, the work [41] provides a novel technique for detecting Big Data quality anomalies relating to six quality dimensions: accuracy, consistency, completeness, conformity, uniqueness, and readability.

In conclusion, after analyzing the previous studies we are not aware that there is any work, either from a conceptual point of view, or from a tool implementation point of view, that considers quality measures oriented to the issues or degradation in the data or its value that imposes its centralization in a common repository.

4 Methodology for measuring the purity and the quality of the data

There are a series of stages in the flow for the proper measurement of the data quality, these are shown in Figure 1. First, in **(1)**, the different data sources need to be selected. These data sources are comprised of the various databases stored in the cloud system, and the data quality of the dataset is calculated **(2)**. Once the sources to work with have been selected, before joining the data, the quality that the future dataset is calculated through mathematical formulation **(3)**. This allows notifying the user of the expected outcome before performing the data union, economizing data processing in case the result is not as anticipated. These objectives are formalized in further detail in Section 6.

Once the join is accepted by the user, it is time to create the data product by merging the sources and adding metadata to the resulting dataset **(4)**. After the merge, the metadata that will depict the dataset is formed by provider identifiers, the type of join that has been performed, the quality of the sources, its own quality, and the importance it holds within the cloud network **(5)**. Using these data, the resultant dataset is introduced into the dataset network in order to determine its centrality in relation to the other datasets in the network.

The last two identifiers are the ones that will have the greatest impact on the development of this paper. The quality dimensions are responsible for assigning intrinsic value to the data, because the higher the quality, the more reliable its use will be for future work. In addition, centrality indicators will determine the value of the dataset within the network, quantifying its importance in relation to other data in the cloud. This enhances decision-making when working with data and provides greater efficiency to search processes. The following scenarios are responsible for calculating the six data quality dimensions (Completeness, Accuracy, Consistency, Validity, Uniqueness, and Timeliness) and for performing centrality calculations using the purity dimension (Degree Centrality, Betweenness Centrality, and Closeness Centrality). All these quality indicators are detailed in Section 5.

5 Data Purity Dimensions

This section describes the main contribution of this manuscript to data quality with the new data purity dimension. The Purity dimension is responsible for measuring how central the various datasets are. It becomes critical for success, complex, and interdependent within the company's network, which is why data centralization offers organizations benefits such as improved efficiency, quality, and decision-making.

Data plays a crucial role in informed decision-making, impacting the quality of business decisions based on data quality and availability. Efficient data management enhances internal processes, leading to increased operational efficiency and reduced costs. Organizations effectively leveraging their data tend to be more competitive, adapting quickly to market changes and understanding customer needs. Additionally, service quality can be improved by collecting and analyzing data related to customer feedback, production efficiency, and other relevant factors [33]. All these advantages can be measured with the data purity dimension; that is why this section will explain the different Key Performance Indicators (KPIs) designed to measure centrality of each dataset in the network that is created with all the data sources of the company.

5.1 Degree Centrality

As explained in [18, 19], degree centrality is the simplest and most straightforward measure. It quantifies the number of direct connections a node has in a network. Multiple ties to the same node are counted only once.

In simpler terms, the degree centrality of a node is the number of connections it has. The higher the number of connections a node has, the greater its degree centrality, and consequently, it is considered more central in the network [23]. In the context discussed in this research, this indicator is used to analyze how many connections each dataset has to understand their importance in the cloud environment. Thus, datasets with higher degree centrality are deemed more important, indicating that their data purity has a greater value.

5.2 Betweenness Centrality

In [3, 19], unlike degree centrality, betweenness centrality focuses on the importance of a node as an intermediary in the shortest paths between other nodes in the network. In other words, it evaluates how often a node acts as a bridge or intermediary in communication routes between other nodes.

Betweenness centrality is commonly used in situations where one aims to identify nodes that play a crucial role in the transfer of information or resources across a network [3]. In the context discussed in this paper, this metric will determine the importance of each node in the dataset network. In the event that a node disappears, it will assess how much communication and efficiency might deteriorate, particularly if a domain wants to access data held by another domain.

5.3 Closeness Centrality

In [19], closeness centrality measures how quickly a node can reach all other nodes in the network. It is based on the average shortest path length from a node to all other nodes. The importance of the node resides in proximity to the other nodes of the network.

Closeness centrality is calculated by considering the geodesic distance (the number of links in the shortest path) between a particular node and all other nodes in the network [34]. The shorter the paths connecting a node to other nodes, the higher its closeness centrality. In practical terms, this implies that a node with high closeness centrality can communicate or interact with other nodes in the network more efficiently than a node with low closeness centrality.

Closeness centrality is especially relevant for the purity indicator in situations where the speed and efficiency of communication or information transmission are crucial. In the context discussed in this research, this metric will determine the importance of each node in the dataset graph by adding value to nodes that are closer to a greater number of nodes. This approach will provide insights into how different domains act upon their various datasets.

5.4 Centralized Quality

These indicators measure the quality of the dataset that is about to be formed before it is created, allowing notification to the user about the expected level of quality of the dataset. For this calculation, the six dimensions (accuracy, completeness, timeliness, validity, uniqueness, and consistency) of the base datasets used for the merger will be considered.

Each quality dimension will have a formula capable of predicting that dimension in the target dataset, this calculation is explained in Section 6. By computing this in advance, the system obtains the following attributes:

- Problem identification: centralized quality metrics can help identify specific issues within the dataset. This allows addressing these problems before using the data in analyzes or models.
- Informed decision making: with established quality metrics, professionals can make more informed decisions about how to approach the creation of new datasets with the available sources in the system.
- Efficiency in the collection and preprocessing: by establishing quality metrics from the beginning, you can optimize data collection and preprocessing processes. This can save time and resources by avoiding the need to correct data quality issues later in the workflow.
- Ease of collaboration: having objective and quantifiable metrics facilitates communication and collaboration between teams working with the data. Everyone shares a common understanding of data quality, reducing misunderstandings.
- Resource savings: identifying and addressing data quality issues from the beginning can save resources in the long term. Avoiding quality problems before they impact decision-making processes and analysis can reduce costs and time.

6 Mathematical Computing for Purity Dimension

This section offers an overview of the new quality dimension that is explained in this paper, coined purity. The focus will be on exploring the various mathematical formulations that make up the different indicators of the dimension.

6.1 Centrality Formulas

This subsection contains the mathematical formulation of the centrality of a single node of the network. Let V be a set of objects connected together by links, where $|V| > 0$ and given n nodes, each node i ∈ V. In said set, three centrality indicators can be calculated: degree, betweenness and closenesss. Degree centrality can be calculated with the following formula 1. Let N_i be a set composed of neighbors of node i.

$$DegreeCentrality(i) = |N_i| \tag{1}$$

In the case of calculating the betweenness centrality indicator for the node i two new variables need to be instantiated. Let $\sigma_{uv} = \sigma_{vu}$ be the number of shortest paths between nodes u and v, where σ_{uu}=1 and u ≠ i ≠ v ∈ V. And let $\sigma_{uv}(i)$ be the number of shortest paths between nodes u and v that pass through node i [9, 36].

$$BetweennessCentrality(i) = \sum_{u \neq i \neq v \in V} \frac{\sigma_{uv}(i)}{\sigma_{uv}} \tag{2}$$

For determining closeness centrality, let d_{ij} be the geodesic distance between nodes i and j, the minimum length of any path between vertices i and j. By definition, $d_{jj} = 0$ for every j ∈ V and $d_{ij} = d_{ji}$ for i, j ∈ V [9, 15, 36].

$$ClosenessCentrality(i) = \frac{1}{\sum_{j \in V} d_{ij}} \tag{3}$$

High centrality scores indicate that a node can reach to other nodes on more optimized paths. Improving the efficiency of communication in the transmission of data. These values will grant how pure each dataset is in the network, specifying the importance of the dataset for the company its selves.

6.2 Centralized Quality Formulas

To perform centralized quality analysis, an indicator has been created for each of the dimensions analyzed in the previous quality

section (Accuracy, Completeness, Consistency, Timeliness, Uniqueness, and Validity).

Once analyzed the six dimensions of the data quality, a series of similarities can be observed among them. One of these similarities is found in the dimensions of timeliness and uniqueness, which validate groups of values in the datasets, they measure the quality of rows instead of unique values. The other similarity could be found in the accuracy, completeness, consistency, and validity dimensions, that they quantify the quality of individual values of the datasets. With the division of these groups, a formula has been created for each one to predict the corresponding dimension with the least possible error. These formulas are explained in Sections 6.2.1 and 6.2.2.

6.2.1 Timeliness and Uniqueness. Equation 4 calculates the overall quality of a dataset in the dimensions of timeliness and uniqueness by averaging the quality of each individual dataset. However, it gives more weight to datasets that have more rows in the intersection of the datasets, because in the dimensions that operate with rows, the more rows a dataset has in the join, the greater its weight will be in terms of quality regarding the outcome.

Let's n be the total number of datasets that are going to be merged, being Q_i the quality of each dataset and row_i the number of rows of the dataset itself. Let row_{total} be the number of rows the final dataset will have after the join. Let $row_{\cap}$ be the number of rows in the intersection of the datasets. This is formalized as follows.

$$PredQuality = \frac{1}{\text{row}_{total}} \cdot \sum_{i=1}^{n} Q_i \left(\text{row}_i - \frac{row_{\cap}}{n}\right) \quad (4)$$

6.2.2 Accuracy, Completeness, Consistency, and Validity. The following formula calculates the overall quality of a dataset in the dimensions of accuracy, completeness, consistency, and validity by calculating the number of erroneous values in a column of the dataset after the join. After predicting the quality of each column, the average of all values should be calculated. It is important to consider not counting repeated columns.

This formula is based on the principle of inclusion-exclusion, thus allowing the calculation of how many additional errors will be added to each of the columns in the new dataset. Through this, a good approximation to the real value in the corresponding quality dimension of the dataset can be achieved.

The first part of the equation, represented by equation 5, calculates the number of erroneous values in column i before forming the join, which is stored in the variable $errors_i$. Then, it iterates over all of the other dataset's j and calculates the number of erroneous values in column i that were introduced by joining dataset.

Once calculated the error for each column in $i \in col_{total}$, for getting the final centralized quality value, in the formula 6 is calculated the mean of all the columns.

$$\epsilon_i = errors_i + \sum_{\emptyset \neq J \subseteq \{1,..,n\} \setminus \{i\}} (-1)^{|J|+1} \left| \left(\bigcap_{j \in J} row_j \right) \setminus row_i \right| \quad (5)$$

$$PredQuality = \frac{1}{col_{total}} \sum_{i=1}^{col_{total}} \left(1 - \frac{\epsilon_i}{row_{total}}\right) \quad (6)$$

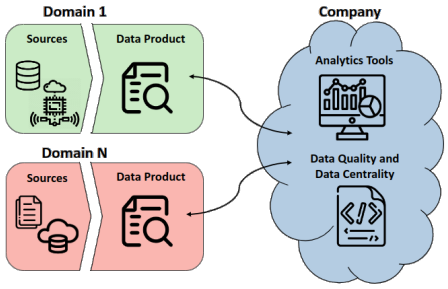

Figure 2: Usage example of the purity dimension

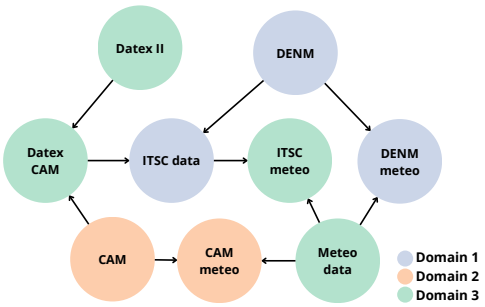

Figure 3: Mobility use case dataset graph

The formula in the image calculates the default quality of a product. Default quality is a measure of a product's quality before any testing or inspection is conducted. It is calculated as the sum of the expected errors in each feature of the product.

7 Validation Scenario

In this section, we advocate for the use of the new purity dimension in a use case for the validation of the methodology presented in Section 4 and the formulas explained in Section 6. Section 7.1 presents the use case scenario, followed by Section 7.2 which shows the use case's outcomes.

7.1 Mobility Use Case

As a matter of illustration, we present a mobility use case in the field of ITS [2]. Let's imagine a company that needs to optimize its real-time traffic control system using datasets from various sources. All these data sources are not stored within the company but are decentralized between the different providers of mobility sensors. Among these data we have different mobility standards, such as CAM [3] [8], DENM [4] [37] or Datex II [12], and we also have meteorological data from the cities.

The objective is to measure the quality of all data before using it in different artificial intelligence models. However, the challenge faced is that each data distributor assesses data quality using different standards. Hence, we cannot impose our quality standards

[2]Intelligent Transportation System
[3]Cooperative Awareness Messages
[4]Decentralized Environmental Notification Messages

	Dataset	D	B	C
1	CAM data	0.25	0	0
2	DENM data	0.25	0	0
3	DATEX data	0.12	0	0
4	Meteo data	0.38	0	0
5	CAM meteo	0.25	0	0.25
6	DENM meteo	0.25	0	0.25
7	CAM Datex	0.38	0.07	0.25
8	ITSC data	0.38	0.07	0.33
9	ITSC meteo	0.25	0	0.38

D: Degree, B: Betweenness, C: Closeness

Table 1: Results of centrality indicators

		Timeliness F. 4		Completeness F. 5 and 6	
Amount of Data	Duplicate Rows	Real Quality	Predicted Quality	Real Quality	Predicted Quality
500	0	0.5999	0.5999	0.5547	0.5547
500	90	0.5849	0.5554	0.555	0.553
500	180	0.5979	0.5870	0.5602	0.5533
500	300	0.5784	0.5694	0.5604	0.5521
1500	0	0.617	0.617	0.557	0.557
1500	90	0.6158	0.5915	0.5584	0.5536
1500	180	0.624	0.6139	0.5622	0.5544
1500	300	0.6367	0.6144	0.5636	0.5536
3000	0	0.6086	0.6086	0.5544	0.5544
3000	90	0.6103	0.6022	0.555	0.5525
3000	180	0.6286	0.6069	0.5592	0.5529
3000	300	0.6364	0.6154	0.5735	0.5598

Table 2: Result comparison of the centralized quality formulas

on them. Figure 2 illustrates a common example of data decentralization in the cloud. Each domain is formed by a distributor in the network, and each one will have the corresponding datasets. In Figure 3 it is shown how all the datasets of the different domains are related. As can be seen in the image, each dataset forms a node within the network, and the connection between the nodes is responsible for representing the relationships between the different datasets.

In this scenario, the purity dimension gains strength when wanting to work with data from other domains. It will allow us to calculate the relevance of each dataset within the network, considering the significance for the company. This enables us to understand the potential impact in case of changes to any dataset. Also, the implementation of the dimension will save cloud resources through centralized quality prediction. Because the quality of the resulted dataset could be uncertain in the six data quality dimensions. In these types of cases, the newly introduced quality dimension in the document becomes suitable.

7.2 Results

This section presents the main results of the process to validate the new purity dimension in the mobility use case. To verify the feasibility of purity dimension as a new beneficial dimension of data quality in decentralized scenarios, it is relevant to check the different indicators presented with the data provided in the use case.

As explained in the use case, we have four datasets from different domains. These domains initiate requests to integrate their datasets with those accessible on the network. In Table 1, it can be seen the results of centrality (degree, betweenness, and closeness) of the different datasets created with the main four datasets.

As can be observed, there are three different cases. The first one concerns datasets from which data originates, which have a zero in fields **B** and **C** since they are located at the edge of the graph. On the other hand, we can see datasets that are in the middle of the network; these have values in both **B** and **C**, indicating that they depend on another dataset for their creation, and in turn, another dataset depends on them. Finally, we have datasets that have no dependent dataset connected to them. These have a zero in **B** since they are not in the middle of the network, but they have a non-zero value in **C**.

As previously stated, these values represent the node's relevance in the graph, the most important datasets are the ones with a zero in **B** and **C**, because they are the origin of the data, and the one with high values in **B** and **C**, because they have more datasets depending from them. If changes are made to that data, it would affect the other datasets in the network will greatly be affected.

The implementation of data purity in the use case is also essential cause the big amount off data that could be stored in each domains. To avoid processing data with potential poor outcomes in various quality dimensions, the formulas presented in Section 6.2 have been tested on the merge of the four datasets of the use case. To achieve this, tests were performed with varied quantities of data in each dataset and with different row duplication, allowing us to see how these factors might affect the formulas performance. We have tested the formula with varying numbers of rows in the datasets to observe how this might impact the results, especially when dealing with large amounts of data. Additionally, we measured the effect based on the quantity of duplicated rows in each dataset. This consideration arises because, as formulas evaluate the error for each column by computing values across rows, repeated values can influence the final prediction outcome. However, as demonstrated in the validation scenario of this manuscript, even when duplicating a substantial number of rows, the result is not significantly affected. These results can be observed in Table 2, where both the actual and predicted values of the two measured dimensions (timeliness and completeness) for the test are displayed. The Mean Squared Error (MSE) of formula 4 is $1,26\%$, while in Formula 5 and 6 is $0,45\%$. Considering the low error in these predictions, it is reasonable to assume that they perform well.

8 Conclusions and Future Work

In this research, we provide a new dimension for the data quality known as Purity, which measures the relevance of the datasets in the network and helps to mitigate unnecessary processing usage in the cloud. The addition of purity as a component of data quality offers a more comprehensive view of data quality in complex and decentralized systems. This technique can be especially useful in instances where hierarchy and dependency among datasets are essential components of the decision-making. On one hand, the new dimension predicts the quality of the future data before it is merged with the six dimensions of data quality, so the data manager

can decide whether to compute the merger or not. On the other hand, three indicators of data centrality have been implemented to determine the importance of each of the datasets, taking into account how close it is to the rest of the data, how many nodes are connected to it, and how central a node is in the communication between others. Purity dimension has been analyzed through the mobility use case presented in the Section 7. In it, the proper functioning of the formulas presented in the manuscript has been demonstrated for different mobility domains.

Given our findings on cloud quality measurement, it is clear that a more effective approach is to assess quality from the outset at the edge. Therefore, our future research will focus on advancing edge quality measurement and addressing the rise of decentralized computing and reliance on edge devices. We aim to contribute insights by exploring innovative methodologies for quality assessment and enhancing the performance of distributed systems. Investigating quality metrics and monitoring mechanisms at the edge is crucial, providing potential for optimizing service and application delivery in our evolving technological landscape.

Acknowledgments

The work described in this article is partially funded by the Datamite project, co-funded by the HORIZON Innovation Actions program of the European Union (HORIZON-CL4-2022-DATA-01) under Grant Agreement No. 101092989.

References

[1] Omar Almutiry, Gary Wills, Abdulelah Alwabel, Richard Crowder, and Robert WaIters. 2013. Toward a framework for data quality in cloud-based health information system. In *International Conference on Information Society (i-Society 2013)*. IEEE, 153–157.

[2] Marcel Altendeitering, ISST Fraunhofer, and Tobias Moritz Guggenberger. 2024. Data Quality Tools: Towards a Software Reference Architecture. (2024).

[3] Marc Barthelemy. 2004. Betweenness centrality in large complex networks. *The European physical journal B* 38, 2 (2004), 163–168.

[4] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–52.

[5] Carlo Batini, Anisa Rula, Monica Scannapieco, and Gianluigi Viscusi. 2015. From data quality to big data quality. *Journal of Database Management (JDM)* 26, 1 (2015), 60–82.

[6] Sovit Bhandari, Navin Ranjan, Yeong-Chan Kim, Jong-Do Park, Kwang-Il Hwang, Woo-Hyuk Kim, Youn-Sik Hong, and Hoon Kim. 2021. An Automatic Data Completeness Check Framework for Open Government Data. *Applied Sciences* 11, 19 (2021), 9270.

[7] Abdur Rahim Biswas and Raffaele Giaffreda. 2014. IoT and cloud convergence: Opportunities and challenges. In *2014 IEEE World Forum on Internet of Things (WF-IoT)*. IEEE, 375–376.

[8] Annette Bohm, Magnus Jonsson, and Elisabeth Uhlemann. 2011. Adaptive cooperative awareness messaging for enhanced overtaking assistance on rural roads. In *2011 IEEE Vehicular Technology Conference (VTC Fall)*. IEEE, 1–5.

[9] Ulrik Brandes. 2001. A faster algorithm for betweenness centrality. *Journal of mathematical sociology* 25, 2 (2001), 163–177.

[10] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2022. The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529* (2022), 5–11.

[11] Corinna Cichy and Stefan Rass. 2019. An overview of data quality frameworks. *IEEE Access* 7 (2019), 24634–24648.

[12] Datex II. 2024. Welcome to Datex II. https://datex2.eu/. Accessed 2024-01-26.

[13] Primavera De Filippi and Smari McCarthy. 2012. Cloud computing: Centralization and data sovereignty. *European Journal of Law and Technology* 3, 2 (2012).

[14] Tharam Dillon, Chen Wu, and Elizabeth Chang. 2010. Cloud computing: issues and challenges. In *2010 24th IEEE international conference on advanced information networking and applications*. Ieee, 27–33.

[15] R Eballe and I Cabahug. 2021. Closeness centrality of some graph families. *International Journal of Contemporary Mathematical Sciences* 16, 4 (2021), 127–134.

[16] Johann Eder and Vladimir A Shekhovtsov. 2021. Data quality for federated medical data lakes. *International Journal of Web Information Systems* 17, 5 (2021), 407–426.

[17] Widad Elouataoui, Imane El Alaoui, Saida El Mendili, and Youssef Gahi. 2022. An Advanced Big Data Quality Framework Based on Weighted Metrics. *Big Data and Cognitive Computing* 6, 4 (2022), 153.

[18] Martin G Everett and Stephen P Borgatti. 1999. The centrality of groups and classes. *The Journal of mathematical sociology* 23, 3 (1999), 181–201.

[19] Martin G Everett and Stephen P Borgatti. 2005. Extending centrality. *Models and methods in social network analysis* 35, 1 (2005), 57–76.

[20] Hadi Fadlallah, Rima Kilany, Houssein Dhayne, Rami El Haddad, Rafiqul Haque, Yehia Taher, and Ali Jaber. 2023. Bigqa: Declarative big data quality assessment. *ACM Journal of Data and Information Quality* 15, 3 (2023), 1–30.

[21] Wenfei Fan. 2015. Data quality: From theory to practice. *Acm Sigmod Record* 44, 3 (2015), 7–18.

[22] Wenfei Fan and Floris Geerts. 2022. *Foundations of data quality management*. Springer Nature.

[23] Linton C Freeman et al. 2002. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge* 1 (2002), 238–263.

[24] Ammar Gharaibeh, Mohammad A Salahuddin, Sayed Jahed Hussini, Abdallah Khreishah, Issa Khalil, Mohsen Guizani, and Ala Al-Fuqaha. 2017. Smart cities: A survey on data management, security, and enabling technologies. *IEEE Communications Surveys & Tutorials* 19, 4 (2017), 2456–2501.

[25] ISO/IEC. 2022. ISO/IEC 25012. https://iso25000.com/index.php/en/iso-25000-standards/iso-25012. Accessed on January 23, 2024.

[26] Natalia Miloslavskaya and Alexander Tolstoy. 2016. Big data, fast data and data lake concepts. *Procedia Computer Science* 88 (2016), 300–305.

[27] Aiswarya Raj Munappy, Jan Bosch, and Helena Homström Olsson. 2020. Data pipeline management in practice: Challenges and opportunities. In *Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings 21*. Springer, 168–184.

[28] Joe Myers. 2021. This is how much data we're using on our phones. https://www.weforum.org/agenda/2021/08/how-the-pandemic-sparked-a-data-boom/. Last accessed 23 January 2024.

[29] Anastasija Nikiforova. 2020. Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment. *Baltic Journal of Modern Computing* 8, 3 (2020).

[30] Jayesh Patel. 2019. Bridging data silos using big data integration. *International Journal of Database Management Systems* 11, 3 (2019), 01–06.

[31] Leo L Pipino, Yang W Lee, and Richard Y Wang. 2002. Data quality assessment. *Commun. ACM* 45, 4 (2002), 211–218.

[32] Ling Qian, Zhiguo Luo, Yujian Du, and Leitao Guo. 2009. Cloud computing: An overview. In *Cloud Computing: First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. Proceedings 1*. Springer, 626–631.

[33] Thomas C Redman. 1995. Improve data quality for competitive advantage. *MIT Sloan Management Review* 36, 2 (1995), 99.

[34] Yannick Rochat. 2009. *Closeness centrality extended to unconnected graphs: The harmonic centrality index*. Technical Report.

[35] Ahmed Shawish and Maria Salama. 2013. Cloud computing: paradigms and technologies. In *Inter-cooperative collective intelligence: Techniques and applications*. Springer, 39–67.

[36] Kshitij Shukla, Sai Charan Regunta, Sai Harsh Tondomker, and Kishore Kothapalli. 2020. Efficient parallel algorithms for betweenness-and closeness-centrality in dynamic graphs. In *Proceedings of the 34th ACM International Conference on Supercomputing*. 1–12.

[37] Dayong Song, Yanheng Liu, Jian Wang, Weiwen Deng, and Heekuck Oh. 2017. Performance Modeling and Analysis of Decentralized Environmental Notification Message in Vehicular Networks. *Adhoc & Sensor Wireless Networks* 39 (2017).

[38] Ikbal Taleb, Mohamed Adel Serhani, Chafik Bouhaddioui, and Rachida Dssouli. 2021. Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data* 8, 1 (2021), 1–41.

[39] Alejandro Vaisman and Esteban Zimányi. 2014. Data warehouse systems. *Data-Centric Systems and Applications* (2014).

[40] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.

[41] Elouataoui Widad, Elmendili Saida, and Youssef Gahi. 2023. Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis. *IEEE Access* (2023).