Big Data Quality Scoring for Structured Data Using MapReduce

Shalini Dhamodharan[1], Narmada Kokila[2], Chandrika Pathakamuri[3], Vinuthna Ghattamaneni[4], and Yalong Wu[5]

[all]Department of Computing Sciences , University of Houston-Clear Lake
[1]dhamodharans1298@uhcl.edu
[2]kokilan3198@uhcl.edu
[3]pathakamuric9332@uhcl.edu
[4]ghattamaneniv3051@uhcl.edu
[5]wuy@uhcl.edu

***Abstract*—In the era of big data, numerous applications are constructed with data as the foundational element. To ensure the reliability and trustworthiness of these application outcomes, they must be built upon high-quality data. Data quality serves as a metric, indicating how well the data meets business needs and requirements. Given the substantial surge in data volume originating from diverse sources, assessing data quality becomes a practical challenge. In this paper, we present an innovative Big Data Quality Score Model crafted to assess data quality in extensive datasets, utilizing the Hadoop MapReduce framework. While alternative sophisticated architectures exist, Hadoop capitalizes on the MapReduce programming model by facilitating seamless access to vast datasets stored in the Hadoop distributed file system. Importantly, our proposed model can provide data quality scores for large data sources, distinguishing it from other models that are tailored for either small datasets or resort to data sampling methods. We propose Accuracy, Completeness, Consistency, Timeliness, and Correlation as pivotal measures of data quality. Each data quality metric is assigned a rating on a scale from 0 to 100, and the overall final score is computed by scaling and normalizing the feature ratings. This research not only contributes a robust methodology for data quality assessment in the era of Big Data but also provides practical insights applicable to various domains, enhancing the integrity and usability of extensive datasets.**

***Index Terms*—Big Data, Assessment, Hadoop, MapReduce, Mapper, Reducer, Data Quality Score**

I. Introduction

The world has witnessed an unprecedented surge in the generation and accumulation of data. Various domains such as national security, healthcare, economics, and media, involve big data analytics [1]. Big data is characterized by high volume, velocity, and variety. Specific technologies including Hadoop MapReduce used for speedy computation and analytical methods such as Cluster analysis, Genetic algorithms, Natural Language Processing, Machine learning, Neural Networks, Predictive modeling, Regression Models, Social Network Analysis, Sentiment Analysis, Signal Processing, Data Visualization are required to transform big data into value [2]. We must ensure that, to extract knowledge from data and for practical data analysis, the data should be of good quality so as to obtain insights from it by analyzing through various transformations. The increase in data has brought new challenges in measuring data quality. The structure and complexity of data, along with the unpredictable volume of data, raise potential concerns regarding quantifying data quality.

How do we determine whether the information or data gathered is useful to extract insights? How can we measure and quantify the practical usage of the data present? It is by defining its quality. Data quality is defined as “fitness for use” [3]. It is the degree to which data accurately reflects reality. This can be affected by several factors, including the data collection methods, the tools used to process it, and how it is managed. The importance of high-quality data in decision-making processes cannot be overstated. For Instance, in the context of customer communication and marketing, a lack of customer information might lead to misdirected efforts, missed opportunities, wasted resources, and damage to the company’s reputation. Investing in data quality assurance processes becomes crucial to avoid these pitfalls and maximize the effectiveness of marketing campaigns. Poor data quality not only jeopardizes the accuracy of analyses but also leads to missed opportunities and misguided actions, incurring significant costs for businesses and organizations that are associated with production, marketing, and R&D. This research explores and addresses the multifaceted challenges of Big Data Quality Assessment that play a vital role in today’s business.

The notion of Big Data Quality Assessment has a plethora of approaches in the existing studies. However, these approaches are not open-sourced. Extending the literature to find a solution to quantify the big data quality assessment is the scope of study in this paper. Our primary objective of the study is to develop Hadoop MapReduce-based data quality scorecards to score the data. The data quality score measures how trustworthy and reliable the data source is. This includes identifying different data formats, devising data quality metrics, implementing MapReduce algorithms, and deriving a comprehensive scorecard for assessing data quality. We have explored and defined data quality scores based on data quality metrics such as Accuracy, Completeness, Correlation, Consistency, and, Timeliness. Additionally, we aim to open-source our methodology, contributing to the growing body of knowledge in the field.

This paper is structured as follows: Section II gives an overview of the existing data quality tools in the industry. Section III articulates the specific challenges addressed in this research project, delineating the methodologies employed. Section IV discusses the detailed implementation of the methodology for data quality assessment, leveraging the Hadoop Map-Reduce framework. Section V validates our methodology against other existing approaches. Finally, Section VI concludes the paper.

II. LITERATURE REVIEW

Data quality assessment and scoring techniques have played an integral role as a toolkit for Data Analysts and Engineers for quite some time. These methodologies are employed to verify the suitability of data for use and to ensure that it meets specific requirements. Notably, major players in the industry, such as IBM, Talend, Informatica, and, other proposed techniques like Great Expectations, have released scoring models for qualifying data quality. However, these models have often restricted their methodologies from users. The following parts of this section will provide a detailed explanation of how these entities score data quality, their approaches, and, methodologies.

IBM utilizes its Data Cloud Pak for Data, coupled with Watson's Knowledge Catalog, to establish a comprehensive platform for monitoring and analyzing data quality. Embedded within Watson's Knowledge Catalog, the data quality feature empowers users to evaluate data across various dimensions, including accuracy, completeness, consistency, timeliness, uniqueness, and validity. Supported data formats include csv, xlsx, text, and JSON files, accessible in both Watson Studio and Watson Knowledge Catalog[4]. IBM introduces the Entity confidence dimension, indicating the system's confidence in correct entity matches within the data. Data quality scores are calculated using the data quality rules that result in asset scores(dimension scores). A weighted average of the corresponding scores of its columns is taken. Despite the comprehensive evaluation of these core dimensions through data quality checks, the specific metrics for each scoring method are not disclosed by IBM. This lack of transparency underscores the importance of understanding how these metrics are measured[5]. It is worth noting that while IBM Watson Studio offers a free version with certain feature restrictions, obtaining the data quality score is not part of the free usage and may involve additional costs.

However, with the necessity of accurately interpreting data and conveying it to clients, and business leaders visually, companies tend to use some of these platforms like Talend. Talend is a data quality management solution that offers both on-premises and cloud-based solutions. It is designed to help organizations ensure the accuracy, completeness, and reliability of their data. It provides various options such as Talend Data Inventory to upload the datasets, understand the data, and make sense of it. Talend Data Preparation helps in cleansing and standardizing the data while Talend Studio and Talent Management Console transform and integrate the data[6]. Talend Data Inventory, within its suite of data quality features, offers a set of metrics to determine the overall data quality score for a dataset[7]. It also facilitates the uploading of our datasets, conducting the data quality assessments, and generating scores for the metrics. Users can define data quality rules for each metric, allowing for the customization and establishment of specific criteria governing metrics such as usage, validity, popularity, completeness, and discoverability.

Informatica is another cloud-based licensed data management solution that can help businesses drive better results[8]. The tool helps users cleanse, standardize, monitor, and profile data to improve customer data, generate more efficient operations, and deliver trusted data analytics. It offers a variety of data quality tools and services, including Cloud Data Quality, Cloud Data Governance and Catalog, and Data as a Service. Informatica's Cloud Data Quality solution [9] delivers trusted data throughout an enterprise using AI-powered automation to manage the entire data quality process, regardless of company size or data volume, ensuring reliable data for analytics and decision-making. Informatica's key features include fast data profiling and cleansing which involves profiling, cleansing, standardizing, and enriching data using a comprehensive set of prebuilt rules. With Informatica's Cloud Data Quality solution, businesses can leverage trusted data to drive better decision-making, enhance customer experiences, and optimize operations.

While IBM, Talend, and Informatica are cloud-based platforms that enable big-data monitoring and quality assessment. There exist other open-sourced packages that have been developed with Python, and used for data analysis. Great Expectations and Panderas libraries in Python have been the leading libraries used for data quality assessment. Great Expectations offers flexibility and control when creating your data quality tests. Great Expectations is an open-source library designed to bring structure, maintainability, and scalability to validating, documenting, and profiling data[10]. It provides a declarative language for expressing expectations about the structure, content, and statistical properties of datasets. It allows users to codify their assumptions about the data into "Expectation Suites" facilitating the automated validation of data as it evolves. Great Expectation has been used to quantify the metrics evaluation of the data quality dimensions[11]. The results for metrics defined for completeness, uniqueness, and consistency use the methods like expect_values_to_be_not_null(),expect_values_to_be_between(), and so on. Aiming for a data quality score that relies on Python as a programming language would extensively use great expectations to the fullest.

As these cloud-based architectures have their limitation we propose our methodology that can be modified based on the user requirements. The Big Data Quality Score Model proposed in this paper emphasizes the data quality dimensions of accuracy, consistency, completeness, timeliness, and correlation for the data in use. To achieve scalability and efficient processing in a cloud-based environment, the methodology uses MapReduce functions, that enable parallel processing of vast datasets across multiple nodes, aligning well with the requirements of large-scale data processing. For the model to be capable of evaluating any dataset quality, domain-specific or data-specific configuration files facilitate the requisites to quantify them. Correspondingly, the MapReduce functions could be defined to satisfy the requirements. The Big Data Quality Score Model focuses on quantifying the dimensions by implementing metrics that are tailored to the specific characteristics and demands of the dataset under consideration. The metric could vary for data quality dimensions like accuracy depending on how they are defined and examined. For instance, numerical and textual data may necessitate different sets of metrics for a comprehensive evaluation. By incorporating these techniques, the proposed approach not only provides a practical way of assessing data quality but also enhances the credibility and reliability of the chosen metrics. The following section gives a brief about the proposed methodology.

III. Methodology

In our paper, we have defined and established a Hadoop MapReduce-based framework to process and measure the quality of big data. Our model is novel and has the efficacy to handle huge volumes of structured data. Extensive research has been performed to identify key metrics that affect the data quality. The following metrics also known as data quality dimensions have been measured in our project:

Accuracy- Accuracy serves as the bedrock of data quality, which ensures the integrity and correctness of the information in a dataset. In [12], Wang et al define accuracy as “the extent to which data is correct, reliable, and certified”. To uphold the required standards of accuracy, thorough data validation techniques must be used. This research paper provides an approach for accuracy through data validation by employing MapReduce to check the correctness of data against a set of authoritative information that is predefined.

Completeness- Completeness, the significance of ’One Missing Piece’ in data cannot be underestimated, as even a single omission can give rise to a multitude of data integrity challenges. Such omissions can impede the decision-making and the efficacy of data analysis. In the research area of structured data, completeness is often related to null values. In general, a null value is a missing value i.e., a value that is present in the real world but absent from a data collection [12].

Consistency- Consistency stands as a paramount Imperative in the realm of maintaining data integrity. Consistency ensures uniformity in data types within respective columns is equally important. According to Silberschatz et al, data consistency in databases implies that data should be considered equivalent when data exists in various storage locations [13]. While it has many perspectives, it can also be interpreted as one being the consistency of the identical data values across tables [14], which has been outlined in this paper.

Correlation- Correlation is a measure of the linearity between two variables. The correlation coefficient defines the relationship between these two variables in a dataset. In terms of data quality, a dataset may require a substantial number of features for proper data analysis. A good data quality standard requires these features or attributes in a data set to be highly correlated with the target feature (i.e., the dependent features), and also the data quality standard requires the features to be not correlated among themselves. Therefore, correlation is one of the key factors in determining the data quality of a dataset. In this paper, we aim to demonstrate the correlation among all the numeric features in a Big Data Set.

Timeliness- Timeliness is a critical factor in the assessment of data quality, as it directly impacts the reliability and relevance of the information contained within a dataset [15]. A study performed by Barkhordari et al, Fisher & Kingma, in order to explain the role of data quality in the explosion of the Challenger spacecraft states that timeliness means the recorded value is up to date. Data that is up-to-date and aligned with current events and conditions is inherently more valuable for decision-making processes. Accurately classifying data records as “latest” or “old” enables the identification of data that can drive more informed and effective analyses and decisions.

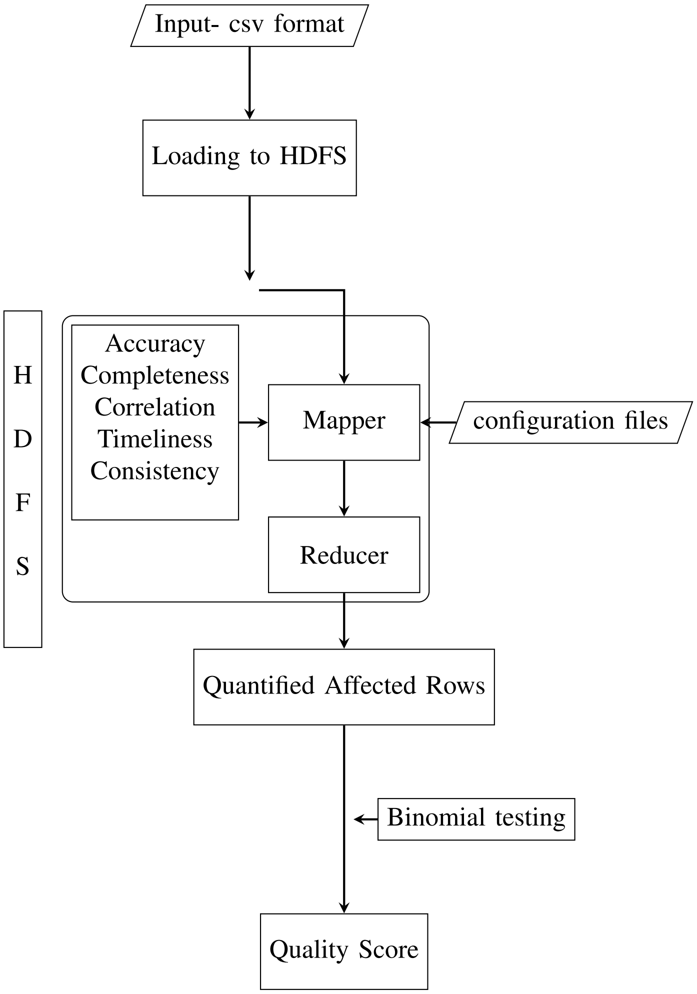

Fig. 1: Big Data Quality Score Framework

Fig. 1 demonstrates the Big Data quality assessment model framework. The data quality assessment framework is designed to harness the robust capabilities of the Hadoop MapReduce framework to conduct a scalable assessment of data quality. The framework entails loading the data into a Hadoop-Distributed File System(HDFS) and strategically partitioning it into manageable chunks to facilitate parallel processing as the first step as mentioned in Fig. 1. Hadoop is a crucial tool for quickly storing and handling huge amounts of data. Its power comes from using a distributed computing model, allowing fast data processing that can be expanded by adding more computing nodes. The data quality assessment requires a set of user-defined configuration files to establish data quality standards and requirements. The exponential increase in data sources has enabled a wide variety of domains, and the data requirement is unique for each domain. Thus, our model facilitates user-defined configurations to assess data quality. MapReduce is a programming technique basically used to handle and analyze large datasets. Map involves breaking the dataset into small chunks and processing them across multiple composting indeed while Reducer performs aggregate function to produce a final output. The Fig. 2 demonstrates the MapReduce architecture.

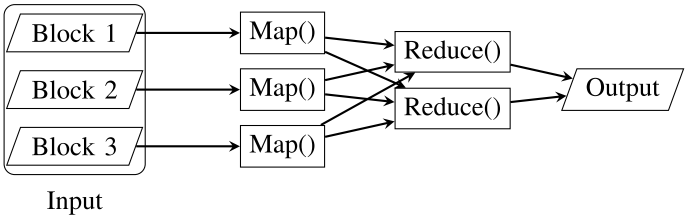

Fig. 2: MapReduce Architecture

Dedicated MapReduce functions have been devised for each critical data quality feature. The MapReduce function devised for each of the data quality features works as such Map identifies all the accurate or efficient records in the dataset and reduces the total number of inefficient rows as part of each of the data quality dimensions. As seen in Fig. 1, reading the HDFS system and configuration files, respective mapreduce blocks are computed. The output of the MapReduce function is the number of good-quality rows. By having a total number of accurate or efficient rows in a dataset, we could determine the percentage of ineffective rows in the dataset using the following formula presented in Equation 1:

$$percentage\_of\_ineffective\_rows = \frac{\#ineffective\_rows}{total\_rows} \tag{1}$$

By referring to the above formula each data quality feature such as Accuracy, Consistency, Correlation, and Timeliness could be quantified in terms of percentage. In the subsequent phase, the framework incorporates the use of statistical binomial testing to evaluate if the quantified quality metric aligned with predetermined thresholds. These thresholds can be either predefined within a 95 % confidence interval or customized by users based on specific requirements. Binomial testing enables us to normalize the scores obtained by evaluating whether the number of affected rows is within the threshold range. If the number of affected rows is accepted, then the feature score would not be affected which would be 100 % and the score would be adjusted accordingly using an adjustment factor. Finally, the Big Data Quality Score Model employs an unweighted scoring method. Each of the quality dimensions is given equal weight. This approach ensures a balanced assessment of overall data quality. It simplifies the scoring process significantly. This unweighted scoring method contributes to a fair evaluation of data quality across various dimensions.

IV. Experimentation and Results

While experimenting, a rigorous and systematic approach was employed to validate the proposed Big Data Quality Score Model. The primary objective was to assess the effectiveness and efficiency of the Big Data Quality Score model in evaluating data quality on a large scale. To achieve this, a carefully designed experiment was conducted to elucidate the framework's ability to handle substantial volumes of data and to analyze its performance in various real-world scenarios.

A. Data Description

The data utilized in this research was sourced from "The Center For Disease Control and Prevention", a reputable repository for credible health information. The specific dataset used for the research is titled 'COVID-19 Vaccinations in the United States, County'. Data represents all vaccine partners encompassing jurisdictional partner clinics, retail pharmacies, long-term care facilities, dialysis centers, Federal Emergency Management Agency, and Health Resources and Services Administration partner sites, and federal entity facilities [16]. It comprises 1.96 million samples and 80 fields, where each sample corresponds to vaccine equity data at the county level which were gathered within the time frame spanning from December 13, 2020, to May 12, 2023. In addition to this, a dataset titled 'COVID-19 Vaccinations in the United States, Jurisdiction', containing similar information, is used for consistency checks.

B. Hadoop Environment and Implementation

A well-configured work environment lays the foundation for efficient and reliable big data processing. In this research, we have installed a Hadoop distributed file system framework suitable for the operating system which is equipped with the latest versions of Hadoop ecosystem components, such as HDFS for distributed storage and MapReduce for parallel processing. Further details on the hardware setup are presented in Table I. Configuring the Hadoop environment requires adjusting different configuration files such as mapred_site.xml, hdfs_site.xml, and core-site.xml to customize the setup to your specific needs. For optimal performance, we can adjust the resource allocation settings in yarn-site.xml Additionally, setting up the SSH keys for secure communication between nodes is very crucial.

Dealing with the complexities of processing large datasets poses numerous challenges. Multi-threading is an ideal approach for tackling these challenges. It enables parallelism,

Component	Specification
Processor	11th Gen Intel(R) Core(TM) i7-1195G7 @ 2.90GHz (2.92 GHz)
Memory(RAM)	16.0 GB915.8 GB usable)
Storage	512 GB SSD
Operating System	Ubuntu 20.04 on Windows 11 Home 64-bit
Software Dependencies	JD, HDFS, Hadoop configuration files(core-site.xml),hdfs-site.xml, mapred-site.xml) , Hadoop daemons(Namenode and Datanode), Python3, SSH.

TABLE I: : Hardware Configuration

where in the context of big data, multi-threading allows data processing tasks to be broken down into smaller units, and these units can be processed concurrently. It reduces the time required to process large volumes of data and makes efficient use of CPU and memory. As a part of our implementation, we have integrated multi-threading. To achieve parallelization in executing multiple Hadoop streaming commands, each corresponding to a MapReduce job, we have employed a ThreadPoolExecutor. This method allows us to efficiently distribute the workload across multiple threads. Our implementation is designed to handle an issue or failure during execution by promptly generating error messages. This kind of error reporting mechanism ensures that any exceptions are appropriately documented and made visible, aiding in the debugging process. This helps organizations make better use of their big data assets, empowering people to make data-driven decisions and gain valuable insights.

After establishing the necessary techniques, the essential algorithms and code are formulated for each phase of the framework. Our paper primarily focuses on developing a model that assesses the quality of a dataset, specifically applying quality metrics to the “COVID-19 Vaccinations in the United States, County” dataset through the implementation of a data quality assessment framework. As a preliminary step, the “COVID-19 Vaccinations in the United States, County” data is loaded into the HDFS platform, and corresponding configuration files are placed in their designated folders to calculate the data quality.

C. Accuracy

In our context, accuracy corresponds to the grammar check for the string data types, and format check for the float data types etc., the predefined set comprises three sets of configuration files, encompassing State codes, County names, and Percentage fields on vaccine equity. Additionally, a configuration file will be employed to store the field names of the dataset. Initially, the mapper’s task is to verify whether each record within the string data field corresponds to the predefined configuration file, which is typically a grammar check, and for the Percentage fields, verify if the data contains a valid float value between 0 and 100. Next, the reducer handles the output from the mapper, typically in the form of key-value pairs where the key represents a field name, and the associated value is ‘1’ if the conditions are met. The reducer’s task is to generate aggregated values for each field based on the key-value pairs. Following that, the combiner computes an accuracy value for the entire dataset. For example, in the County field in the ‘COVID-19 Vaccinations in the United States’, it is validated again county configuration file. The configuration file contains all counties in the United States of America. Fig 3. illustrates the MapReduce framework for accuracy. The MapReduce validates the values in the county column against the configuration file and computes the total number of accurate rows which is 1912999. The score for accuracy is calculated using the following formula:

$$Accuracy\_Score = \frac{\#accurate\_records}{total\_records} \qquad (2)$$

- #accurate_records imply the number of values that are accurately classified within the standard values provided in the configuration files.
- total_records are the total number of values that are checked in the dataset.

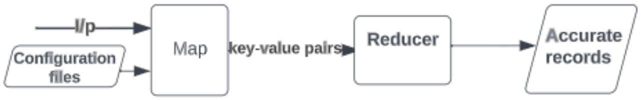

Fig. 3: Accuracy

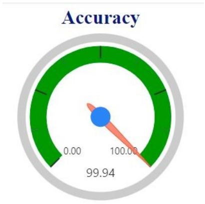

Fig. 4: Accuracy

Fig. 4 represents the final accuracy score.

D. Completeness

In this paper, we present a comprehensive analysis of completeness through a thorough examination of null values employing the MapReduce framework. In the initial step, the mapper employs an approach to identify the null values within each record. Subsequently, it emits key-value pairs wherein the key remains constant while the value assigned is 1 if the identified record is null and 0 if it is not. Subsequently, the reducer aggregates the counts associated with each field. Fig. 5 illustrates the MapReduce framework for completeness. For example, MapReduce has identified the total number of Null values in the Administered_Dose1_Pop_Pct to be 24085. The score for the Completeness is calculated using the following formula:

$$Completeness\_Score = \frac{\#nonempty\_records}{total\_records} \quad (3)$$

- #nonempty records are the complete records that have non-null values.
- total_records are the total number of values that are checked in the dataset.

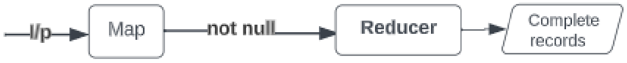

Fig. 5: Completeness

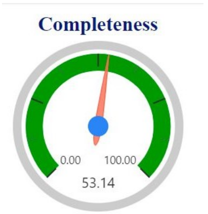

Fig. 6: Completeness

Fig. 6 represents the final completeness score.

E. Consistency

The entire process for consistency validation is implemented using the MapReduce framework, employing two mappers. Each mapper is assigned a distinct input dataset for comparison. Initially, these mappers create a composite primary key, formed by combining Date and Location, which serves as the unique identifier for the data within both datasets. Then produce the field name, primary key, and associated values as output. Subsequently, the role of the reducer is to conduct a comparative analysis of the outputs produced by the mappers, considering the field name and primary key it checks whether the value across both the datasets is the same, if not the reducer identifies the record as inconsistent. Finally, it compiles an aggregated count of inconsistent records. The Fig. 7 illustrates the MapReduce framework for consistency. The consistency between the datasets 'COVID-19 Vaccinations in the United States, County, and COVID-19 Vaccinations in the United States, Jurisdiction' is validated using the MapReduce function.

$$Consitency\_Score = \frac{\#consistent\_records}{total\_records} \quad (4)$$

- #consistent_records are the records that match primary keys.
- total_records are the total number of values that are checked in the dataset.

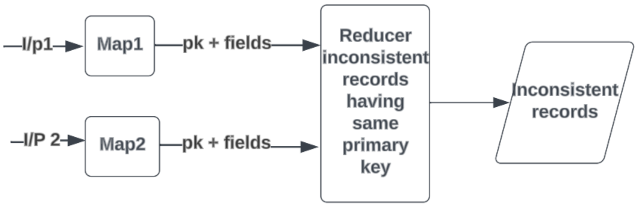

Fig. 7: Consistency

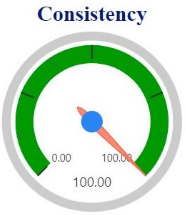

Fig. 8: Consistency

Fig. 8 represents the final consistency score.

F. Timeliness

The task involves a two-step process consisting of a Mapper and a Reducer component. In the Mapper task, Firstly, it entails extracting date records from the dataset. Subsequently, it computes the maximum and minimum dates within the dataset. To establish a range, the Mapper identifies unique dates in the dataset. This range is then employed to determine a threshold, which is set at 20 percent of the range. This range is dynamic and can be changed based on user requirements. In the classification phase, the Mapper categorizes dates falling below this threshold as "old", while any dates exceeding this threshold are designated as "latest". The Reducer task follows the Mapper's output and is responsible for aggregating the results. It calculates and reports the count of records classified as "latest" and "old" based on the Mapper's categorization. Fig. 9 Illustrates the MapReduce framework for Timeliness. The score for Timeliness is calculated using the following formula:

$$Timeliness\_Score = \frac{\#new\_rows}{total\_records} \quad (5)$$

- #new_rows are the records whose date exceeds the threshold set.
- total_records are the total number of values that are checked in the dataset.

Fig. 10 represents the final timeliness score.

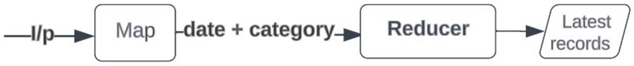

Fig. 9: Timeliness

Timeliness
0.00
100.00
100.00

Fig. 10: Timeliness

G. Correlation

In this paper, we aim to demonstrate the correlation efficiency among all the numeric features in a Big Data Set. We have implemented the Pearson Correlation method. The formula to calculate the Pearson correlation coefficient between two variables, X and Y, with n data points, is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

- x_i and y_i are individual data points for variables x and y.
- $\bar{x}$ and $\bar{y}$ are the means (averages) of variables x and y, respectively.

The correlation coefficient values range from -1 to +1. An absolute correlation value greater than 0.9 indicates a stronger correlation between the two variables. We have implemented the Hadoop Map-Reduce technique for calculating the correlation coefficient. The mapper function creates pairs of features to be measured for each row in the dataset, and the reducer performs the calculation using the formula. Fig. 11 Illustrates the MapReduce framework for Correlation. The score for Correlation is calculated using the following formula:

$$Correlation\_Score = \frac{\#independent\_records}{total\_rows} \tag{7}$$

- #independent_rows are the values that are not highly correlated.
- total_records are the total number of values that are checked in the dataset.

Fig. 12 represents the final accuracy score.

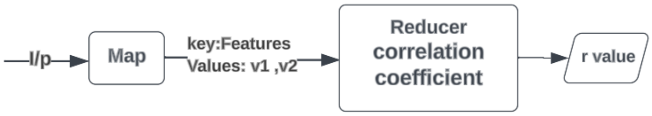

Fig. 11: Correlation

Correlation
0.00
100.00
52.00

Fig. 12: Correlation

The individual scores for data quality metrics, namely Accuracy, Completeness, Consistency, Correlation, and Timeliness, are determined using MapReduce algorithms, as previously discussed. The specific scores for each metric are as follows:
Accuracy: 99.94 %
Completeness: 52.5 %
Consistency: 100 %
Correlation: 57.50%
Timeliness: 100 %.

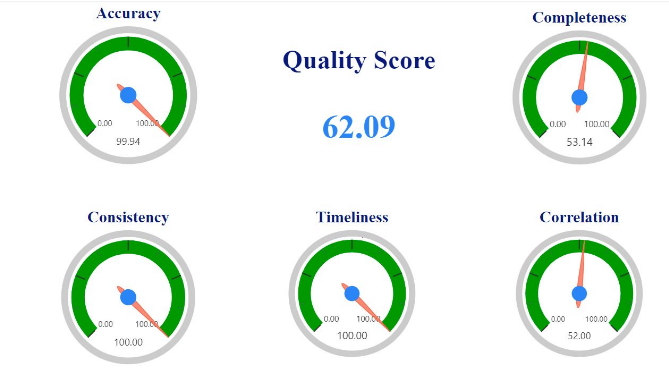

Fig. 13: Quality Scorecard

The overall dataset score is **62.5 %** which is computed by employing a binomial test to assess whether the total number of affected rows falls within an acceptable range of 95 %. If not, the metric scores are adjusted accordingly. Fig. 13 visually represents the contribution of each of the data quality metrics to the Big Data Quality Score for the dataset.

The final score for the dataset is derived by assigning equal weightage to each metric, resulting in a percentage for the “COVID-19 Vaccinations in the United States, County” dataset.

V. Validation

Initially, the dataset has been examined to validate the quality of the data using the Great Expectations package on the Python environment. Completeness and accuracy, the methods defined within the package supported the validity checks, resulting in good data quality that justified the scores obtained from our proposed model. Then our model output undergoes validation against the Talend output, specifically for the dataset titled “COVID-19 Vaccinations in the United States, County”. This validation process utilizes existing tools to ensure the accuracy and reliability of our model. The Talend data quality assessment model assigns a score of 2.38 out of 5 to “COVID-19 Vaccinations in the United States, County”, with corresponding metric values provided by Talend.Talend dashboard uses a radar chart to visualize the trust score. To compare our score with Talend and to understand the distribution of each dimension, a radar chart is plotted. Fig.14 represents the quality score distribution for “COVID-19 Vaccinations in the United States, County” dataset.

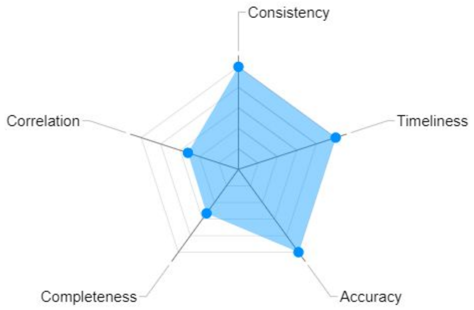

Fig. 14: Data Quality Score Distribution

It is noteworthy that the data quality score generated by the Talend tool closely aligns with the assessment from our data quality framework. The similarity in scores underscores the robustness and consistency of both evaluation methods, reinforcing the confidence in the accuracy of our model’s outputs.

VI. Conclusion

This paper has presented a comprehensive exploration of data quality scores in the context of big data, offering a novel approach through the utilization of the Hadoop MapReduce framework. The data Quality Score Model focuses on data quality dimensions including accuracy, completeness, consistency, correlation, and timeliness with each of the dimensions having metrics associated that can be tweaked depending on the domain we work with. The incorporation of multithreading and statistical binomial testing adds robustness to our model and ensures that data quality adheres to accepted standards. Remarkably, our results are comparable to those obtained through existing data quality scoring techniques. Through this paper, we aim to contribute to the evolving field of big data quality assessment by introducing a scalable, open, and versatile approach. Our proposed model stands as a valuable contribution to this goal, providing a framework that can be adapted and extended in the field of big data analytics.

References

[1] Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques, and technologies: A survey on Big Data. Information sciences, 275, 314-347.

[2] De Mauro, Andrea&Greco, Marco&Grimaldi, Michele. (2016). A formal definition of Big Data based on its essential features. Library Review. 65. 122-135. 10.1108/LR-06-2015-0061.

[3] Cai, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2, pp.1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002

[4] “Supported data sources,” www.ibm.com. https://www.ibm.com/docs/en/cloud-paks/cp-data/4.7.x?topic=data-supported-sources.

[5] “Data quality scores (Watson Knowledge Catalog),” www.ibm.com. https://www.ibm.com/docs/en/cloud-paks/cp-data/4.7.x?topic=results-data-quality-scores.

[6] “Welcome to Talend Help Center,” help.talend.com. https://help.talend.com/search/all?query=Talend+Trust+Score

[7] “Talend - A Cloud Data Integration Leader (modern ETL),” Talend Real-Time Open Source Data Integration Software. https://www.talend.com/

[8] “Enterprise Cloud Data Management — Informatica,” www.informatica.com. https://www.informatica.com/

[9] “Cloud Data Quality - Data Quality Management Tool — Informatica,” www.informatica.com. https://www.informatica.com/products/data-quality/cloud-data-quality-radar.html.

[10] “Welcome — Great Expectations,” docs.greatexpectations.io. https://docs.greatexpectations.io/docs/

[11] W. Elouataoui, I. El Alaoui, S. El Mendili, and Y. Gahi, “An Advanced Big Data Quality Framework Based on Weighted Metrics,” Big Data and Cognitive Computing, vol. 6, no. 4, p. 153, Dec. 2022, doi: https://doi.org/10.3390/bdcc6040153.

[12] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–52, Jul. 2009, doi: https://doi.org/10.1145/1541880.1541883.

[13] Cai, Li & Zhu, Yangyong. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal. 14. 10.5334/dsj-2015-002.

[14] Pipino, Leo & Lee, Yang & Wang, Richard. (2003). Data Quality Assessment. Communications of the ACM. 45. 10.1145/505248.506010.

[15] Wang J, Liu Y, Li P, Lin Z, Sindakis S, Aggarwal S. Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. J Knowl Econ. 2023 Feb 10:1–20. doi: 10.1007/s13132-022-01096-6. Epub ahead of print. PMCID: PMC9912223.

[16] “COVID-19 Vaccinations in the United States,County — Data — Centers for Disease Control and Prevention,” data.cdc.gov. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh