IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11769

Comparative study of popular Data Quality tools

Mahesh S P[1], Dr. G S Mamatha[2]

Undergraduate Student, Department of Information Science and Engineering,

RV College of Engineering, Bengaluru, India1[1]

Professor and Associate Dean (PG Studies), Department of Information Science and Engineering,

RV College of Engineering, Bengaluru, India[2]

Abstract: Organizations now place a high priority on data quality since it might be the key differentiator. There are already several software applications on the market that solve problems with data quality. This study examines the evaluation of data quality solutions created using open-source development practices or made available as free trial versions. The evaluation of Talend Open Studio, Data Cleaner, Informatica Data Quality, Oracle Enterprise Data Quality, and Ataccama's DQ Analyser is specifically covered in this article. Based on performance characteristics and the kinds of data quality issues they addressed, the tools were assessed. In the report, applications of practical tools in data quality programmes that support data governance and master data management activities are also briefly discussed.

Keywords: Data quality, Data Quality Tools, Data Cleansing, Data Integration

I. INTRODUCTION

Data quality in simple words shows us how trustworthy a specific piece of data is and whether or not it would be suitable for a user to utilize in decision-making. The degree of this quality is frequently measured. Data completeness, accuracy, timeliness (e.g., is it up to date?), consistency, validity, and uniqueness are few of the criteria used to determine the status of data. The fundamental or core dimensions of data quality are six. Analysts use these measures to assess the data's feasibility and usefulness to those who require it.

• Accuracy-The information must be accurate and reflect things that actually happen in the world. The measure of accuracy, which is derived by how closely the numbers agree with the verified right information sources, should be confirmed by analysts using verifiable sources.

• Completeness-The data's capacity to properly give all the required values is measured by its completeness.

• Consistency-The uniformity of the data across applications, networks, and when it originates from several sources is referred to as data consistency. The same datasets should be the same and not conflict if they are stored in separate places, which is another definition of consistency. Keep in mind that reliable data can be incorrect.

• Timeliness-Information that is known as timely is the information that can be accessed easily whenever it is needed. This aspect also includes maintaining the data's accuracy; data should be updated in real-time to make sure it is always available and accessible.

• Uniqueness-The absence of duplicate data or redundant information across all datasets is referred to as uniqueness. There should be no duplicate or repeated records in the dataset. Data deduplication and cleansing are used by analysts to assist overcome a low uniqueness score.

• Validity-Data must be gathered in accordance with the business policies and guidelines established by the firm. All dataset values should fall within the specified range, and the data should also follow the necessary, accepted formats. The logic for mapping and transformation can be incorrect throughout the data migration procedure. Data corruption can be caused by problems like run-time errors, network outages, or broken transactions. Such mistakes may result in the retention of invalid data. These could lead to many of the data quality problems, including:

• Improper or Missing records
• Incomplete values
• Invalid values
• Duplicate entries
• Poorly structured values
• Lost relationships within records or systems.

The evaluation portion of this paper includes a number of criteria, including determining the tool's suitability for the-variety of data sources available, performing data extraction and user interface, application framework, and the system's ability to generate reports into exception files or tables. The four main elements of any quality programme created to support organizational data governance and data management projects are data profiling, data integration, data cleansing, and monitoring.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022
DOI: 10.17148/IJARCCE.2022.11769

II. DATA QUALITY TOOLS

Data quality tools that are evaluated and compared in this paper are:

A. Talend Open Studio for data quality

An open-source data integration tool called Talend Open Studio has enhanced capabilities that increase the effectiveness of data integration task design and verified flexibility to guarantee the best execution. In its Magic Quadrant of Data Quality Tools, Gartner promoted Talend from "visionary" in 2016 to "leader" in 2017. One of the most mentioned DQ applications is Talend Open Studio. Both solutions (Open Studio and Enterprise) provide strong support for Big Data analysis using Spark or Hadoop as well as a range of data profiling and cleaning functionalities. We assessed TOS for Data Quality version 6.5.1, and found that it can compete in terms of the data profiling capabilities, business rule administration, and user interface (UI) with a number of commercial DQ products (which charge a fee). DQ monitoring features, on the other hand, are only supported by the Enterprise edition and are not available in the free version.

B. Data Cleaner

Kasper Srensen from Copenhagen developed the open-source data quality tool known as Data Cleaner [9]. Users of this programme can profile, validate, and compare data using these methods. Four aspects of data quality are addressed by the characteristics of the data cleaner. The profiler provides reports on the state of the data and enables users to browse through the content of the data. Statistics like the frequency of null values in records, the greatest and lowest results, and more intricate studies like looking for trends in the data are included in the reports. Users can create validation rules using the validator. These rules for validation can be as straightforward as not accepting null values in specific columns or specifying an input value range, or they can be considerably more complicated, such as comparing input data to a dictionary or a pre-set set of values. Usually, data comparison and profiling will serve as the foundation for the creation of validation criteria. The comparator is employed to determine whether two data sources are comparable. With the use of this function, DBAs can examine the contents of systems which are not constantly synced and spot any data inconsistencies. To ascertain whether the records are identical, it compares tables and records. To ascertain whether the records are identical, it compares tables and records. The monitor is a development of the input validation functionality that gives the DBA the ability to set validation rules and provide alerts when unreliable data enters the system. The user can get the validation reports via the internet thanks to this web-based functionality.

C. Informatica Data Quality

In order to provide enterprise information quality capability in a variety of scenarios, you can combine Informatica Data Quality with Informatica Power Centre. Data Quality Workbench, Data Quality Server are the core components of the system. Data Quality Workbench plans are used to develop, validate, and deploy data quality processes. Workbench helps in quick data exploration and validation of data quality methods by allowing you to test and execute plans flexibly. Server for Data Quality. Used to make exchange of plans and files possible and to run plans in a networked setting. Data Quality Server interacts with Workbench via TCP/IP and provides networking through service domains. An engine and repository for data quality are installed with both Workbench and Server. Users cannot build or change plans using the Server, but they can independently run a program to any Data Quality engines using runtime instructions or through PowerCenter. When running data quality plans through a Data Quality engine, users have the option to apply parameter files, which alter plan operations, to runtime commands. A Data Quality Integration connector for PowerCenter is also available from Informatica. With the help of this plug-in, PowerCenter users can add proper data quality plan directions to a PowerCenter and execute the plan from a PowerCenter session to the Data Quality engine.

D. Oracle Enterprise Data Quality

The business tool Oracle Enterprise Data Quality (EDQ) is a Java Web Based application that stores data using a relational database management system (RDBMS), a Structured Query Language (SQL), and a Java Web Start graphical user interface. In addition to traditional data profiling features, EDQ provides certain DQ monitoring as well as data purification (parsing, standardization, match and merge, address verification). The average user interface (GUI) was criticized for having a rigid data source connection to databases and files. Oracle EDQ requires a snapshot of the real data connection to be created before performing any profiling or DQ measurement tasks, in contrast to other DQ tools in which a connection can be directly accessed and reused. This method avoids an automatic update of the information source.

E. Ataccama's DQ Analyzer

As an Adastra Corp. spin-off, Ataccama Corp. is a multinational corporation that was established in 2007. The creation of tools to enable solutions for Data Quality Management, Master Data Management, and Data Governance is Atacama's area of expertise. The DQ Analyzer is a profiling tool with a user-friendly interface that can handle massive amounts of data processing, statistical frequency distribution analysis for column values, task analysis, primary key and foreign key analysis, dependency analysis, or business rule documentation. For a deeper knowledge of data structures, mask analysis, quantiles, and grouping analysis methods are all options. In order to establish the logic and rules that will be applied to the input data in order to get the desired output, a plan file may be built up in the DQ Analyzer. Steps can carry out a wide range of tasks, including data transformation, filtering, categorization, and reading. A succession of Steps can be visually

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11769

grouped together by placing comments around them to visually group the Steps together or to add notes or other information to the canvas.

Figure 2.1: Data quality dimensions

III. EVALUATION METHODOLOGY

The criteria used for the evaluation of these tools are:

Data Source Connectivity: This criterion assesses how well the tool works with the variety of data sources that are available. It also considers the technology used to make these connections.

Data Storage Management: Assesses the tool's control over data extraction and storing the extracting data. Additionally, it will assess whether the programmer could alter the criteria used to select this data and the capability to integrate multiple data sources records. Interface: This criterion assesses the type and usefulness of the tool's user interface, considering the capability of creating objects by dragging and dropping them, as well as graphical engagement with the data quality procedure. When evaluating the user interface, the following specifications are taken into consideration [2].

1.**System status visibility:** Users should always be aware of their location and the current situation.

2.**System and real-world alignment:** As closely as feasible, the system should reflect the user's actual environment. Use terms, concepts, etc. that the user is familiar with. Procedures should be in logical order.

3.**Freedom and control:** Never "trap" a user. Support functions for exit, undo, and redo that are clearly marked. Avoid locking them into a protracted, non-interruptible linear series of actions.

4.**Standardization and uniformity:** Consistently use verb tenses and objects. Observe platform customs.

5.**Not recall but recognition:** Give the user access to visual cue cards, actions, and alternatives to help with input and navigation. You shouldn't count on them to remember commands.

6.**Efficiency and adaptability in use:** Expert users' interactions can be sped up by accelerators, which are hidden from novice users. Whenever feasible, let users personalize frequently performed actions.

7.**Attractive and simple design:** Information that is only seldom needed should not be seen. Each unit of data gets less clear when a lot of information is displayed on the screen.

8.**Additional documentation and online assistance:** A very well system can be operated without documentation and assistance, but it is possible that additional information is still required.

9.**Error handling:** A good error code is preferable to a design that foresees mistakes. Assist users in identifying, diagnosing, and fixing issues. When feasible, suggest a fix rather than just pointing out the issue.

10.**Development Platform:** This refers to the programming language used to create the application and if it is able to accept additional features or add-ons created in any other language.

11.**Report Creation:** Determines whether the tool's ability to produce reports relevant to the data quality procedure is accessible or whether specialized software is required.

12.**Data Quality dimensions**: Evaluates the tools based on their capabilities in metrics to measure accuracy, completeness, consistency, timeliness and other data quality dimensions.

13.**Rule-based checks**: Evaluates the tools based on their capabilities creating business rules, executing those rules against datasets.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11769

IV. COMPARISON

Talend Open Studio and Informatica data quality provide several downloadable components that allow these tools to connect to a wide range of data sources. Data Cleaner is tested successfully using various data sources, as it uses Java Database Connectivity it's list of connections can be extended to almost any data source. Oracle EDQ also works with Java DataBase Connectivity where it adds new connections by every new released version. Atacama's data quality analyser supports only popular data source connections. Most of the tools tested have a user-friendly interface with drag and drop component manipulation. All the tools are also capable of producing outputs in the form of tables or datasets and most of the tools support creation of rules. The details of comparisons based on important criteria is shown in table IV.1

Criteria name	Talend Open Studio for data quality	Data Cleaner	Informatica Data Quality	Oracle Enterprise Data Quality	Ataccama's DQ Analyzer
Data Source Connectivity	Talend Open Studio supports connect to all wide range of data sources.	It supports wide range of data sources, as it uses JDBC, this can be increased to almost any data source	It supports access to extract data from and deliver data to almost any system, application, or database	It supports connections to data sources that uses standards such as JDBC and ODBC.	It supports connections to most popular data sources by downloading profiles.
Data Storage Management	Uses Extract Load and Transform process	Uses Extract Transform and Load Method.	This tool uses repository which contains two database schemas: the Results and the config schema.	Uses Extract Transform and Load Method.	Uses Extract Load and Transform process
Interface	Almost complete graphical interface is available. Rated 8 out of 10	A complete graphical interface is available. Rated 10 out of 10	Almost complete graphical interface is available. Rated 8 out of 10	Moderate graphical interface is available. Rated 6 out of 10	A complete graphical interface is available. Rated 10 out of 10
Error handling	Uses dedicated customized components such as Asset and AssertCharter	Uses validation tool known as BODS	Uses continuous rules validation technique	Inspect logs in warehouse builder and detecting operators causing error	Automatically spot potential issues in your data flow using tool known as DQV
Development Platform	uses languages like Java, Perl and SQL that allows users to create components	This tool uses languages that include Java, Perl	This tool uses languages that include Java, SQL	It is a Java Web app which uses a Java Servlet Engine, a Java web gui,sql	This tool uses java to develop its data quality tool

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022
DOI: 10.17148/IJARCCE.2022.11769

Report Creation	It produces output in the form of tables or datasets and pie or bar graphics.	It produces output in the form of tables or datasets and pie or bar graphics.	It produces output in the form of various visualization charts like trend, bar chart	The system can produce output in the form of tables or datasets	It produces output in the form of pie or bar graphics.
Data Quality dimensions	It checks all dimensions of data quality	It checks all dimensions except uniqueness	It checks all dimensions except accuracy, timeliness	It checks all dimensions of data quality	It checks all dimensions except completeness
Rule-based checks	It supports creation and application of business rules	It supports creation and application of business rules	It supports creation and application of business and general rules	It supports creation and validation of business rules	It does not support creation of rules

Table IV.1: Comparison of various data quality tools

VI. CONCLUSION

Companies have been looking for effective, affordable solutions to these ever-growing issues since it was realised how important information quality is to the success of a business. These reasons have increased interest in open-source data quality technologies as businesses view them as a solution to their information quality management requirements.
As with choosing any tool, there should be an initial clear plan that specifically outlines its goal, range, and anticipated outcomes. Open-source tools may be quite valuable in this situation. Even while none of the methods we used for our study is a perfect answer, they all have some benefits and could offer a fast glimpse into the fundamental causes of data quality issues.

ACKNOWLEDGMENT

The authors of this paper express their gratitude to **Mr. Prasanth Prakash** and his colleagues of Akamai Technologies for their time and guidance.

REFERENCES

[1]. Val Pushkarev, Henry Neumann, Cihan Varol, John R. Talburt. (2019). An Overview of Open Source Data Quality Tools. International Conference on Information & Knowledge Engineering, IKE 2010, July 12-15, 2019
[2]. Lisa ehrlinger, Elisa rusz,. Wolfram wob, (2019). A Survey of Data Quality Measurement and Monitoring Tools. Frontiers in big data, published doi: 10.3389/fdata.2022.850611
[3]. Alexandre Santuchida, CunhaaFernando, CunhaPeixotobDiego, MartinezPratac, (2016) Data Governance in Large Organizations with Linked Data, 2016 IEEE 8th International Conference on Cyber security and cloud computing
[4]. Mohammad Khodizadeh Nahari; Nasser Ghadiri; Zahra Jafarifard; Ahmad Baraani Dastjerdi; Joerg R. Sack, (2016) Data Governance Model To Enhance Data Quality In Financial Institutions, 2016 International Conference on Communication, Information & Computing Technology (ICCICT)
[5]. J.Sreemathy; S.Priyadharshini; K. Radha; K. Sangeerna; G. Nivetha, Data Quality Using TALEND, 2019 IEEE 5th International Conference on Advanced Computing & Communication Systems (ICACCS)
[6]. Quan Li; Lan Lan; Nianyin Zeng; Lei You; Jin Yin; Xiaobo Zhou; Qun Meng, A Framework for Big Data Governance to Advance RHINs: A Case Study of China, 2019 IEEE 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)
[7]. Alexandre Santuchida, CunhaaFernando, CunhaPeixotobDiego, (2020) MartinezPratac, Robust data reconciliation in chemical reactors, 2020, IEEE International Conference on Services Computing
[8]. L. Seligman, A. Roenthal, Data Governance Model To Enhance Data Quality In Financial Institutions, 2016 International Conference on Communication, Information & Computing Technology (ICCICT)

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11769

[9]. Natasha Micic, Daniel Neagu, Felician Campean, Esmaeil Habib Zadeh, Towards a Data Quality Framework for Heterogeneous Data, 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

[10]. Dan Sui, Olha Sukhoboka & Bernt Sigve Aadnøy, Improvement of Wired Drill Pipe Data Quality via Data Validation and Reconciliation, 2017, International Journal of Automation and Computing

[11]. V. Goasdoué, A. Nugier, D. Duquennoy, and B, Laboisse, "An evaluation framework for data quality tools", Proceedings of the International Conference for Information Quality (ICIQ'07).Alan Gates. 2011. Programming Pig (1st. ed.). O'Reilly Media, Inc.

[12]. S. Madnick, R.Y. Wang, Y. Lee, Y., and H. Zhu, "Overview and Framework for data and Information Quality Research", ACM Journal of Data Quality, Vol. 1, No. 1 Article 2, June 2009.John Russell. 2014. Getting Started with Impala (1st. ed.). O'Reilly Media, Inc.

[13]. F. Dravis, "Information Quality: The Quest for Justification", Business Intelligence Journal, October 26, 2009.