СЕКЦІЯ 1. МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ

УДК 004.4
А. Кашосі, О. Кишкевич, Наталія Загородна
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

УПРАВЛІННЯ ЯКІСТЮ ДАНИХ В ПРОЦЕСІ ЕТЛ В УМОВАХ РЕСУРСНИХ ОБМЕЖЕНЬ

UDC 004.4
A. Kashosi, O. Kyshkevych, N. Zagorodna

DATA QUALITY MANAGEMENT IN ETL PROCESS UNDER RESOURCE CONSTRAINTS

It is impossible to picture today's firms' marketing success without Big Data. Data quality (DQ) evaluation in Big Data operations provides unique evaluation challenges. Several research [1, 2] have attempted to provide a more exact description of this term, which represents information assets with high volume, velocity, and variability that require particular technology and analytical approaches to convert into value.

A data integration phase is required inside the BI system for a variety of reasons, including heterogeneous formats, data formats that are difficult to comprehend or ambiguous, legacy systems that employ antiquated databases, and the structure of the data source that varies over time. All these data source factors make DQ unreliable. [3]

Dakrory et al. conducted research [4] to automate testing procedures that validate Data Quality Dimensions (DQD) such as completeness, consistency, uniqueness, validity, timeliness, and accuracy. The goal of their study was to track DQ in ETL to ensure the correctness of ETL methods and assess whether they need to be changed to solve data problems.

Moreover, the study in [4] leads to a more realistic representation and classification of DQ tests in ETL methods.

- Completeness: ensures that all relevant data is stored, which includes checking the number of records, duplicates, integrity constraints, and data boundaries.
- Consistency: ensures that all values are consistent across all datasets. This includes field mapping, integrity constraints, aggregation of metrics, and hierarchy-level integrity checks.
- Uniqueness: ensures that there are no duplicates in the stored data.
- Validity: Ensures that the data conforms to the syntax (format, type, range) of its description, and its implementation includes integrity constraint checking, field data type checking, and field length checking.
- Timeliness: ensures that all data is stored within the specified period. Data access and freshness are the tests that are applied.
- Correctness: ensures that all data accurately describes the «real object or event being described» Field-to-field comparison, data boundaries, and integrity constraints are all difficult to implement.

The implementation of DQ in an ETL process dealing with Big Data faces a problem related to the nature of the data to be validated, which presents characteristics such as «volume», «velocity», and «variety», to name a few, as Nagham and Laden argued in their research [5] that the number of Vs in Big Data varies by sector.

Volume is a particular challenge when developing tests to validate the quality of the dataset going through the ETL process. As a result, we have explored a sampling strategy that allows us to obtain a representative DQ result without having to process each data item.

Sampling has been successfully applied in areas such as network traffic measurement [7], where it has been shown to improve memory allocation and analysis performance.

To this end, we analyze the type of system that uses Big Data in this study and examine processes in batch rather than in stream. This is significant because it determines the type of sampling method to be used. We used stratified random sampling in this study.

Stratified random sampling divides the sampling units of the population into homogeneous groups called strata, and then draws a simple random sample within each stratum.

Strata are determined by information other than the attributes being assessed that is known or assumed to vary with the attributes of interest. Stratifying the population into homogeneous groups of sampling units minimizes sampling error; estimates obtained within strata are more accurate than simple random samples drawn from the same population. [6]

In this study, we consider the model of a data management system in the Python programming language. One of the many microservices manages the data integration process with various Customer Relationship Management (CRM) data sources, namely ZIP, TXT, CLS, XLS and XLSX files. The process of extracting, transforming, and loading is done sequentially, followed by the final transformation processes, i.e., sampling and performing the DQD tests and generating the DQ assessment reports.

Algorithm:

1. Divide the entire population into non-overlapping strata
2. Select a simple random sample from within each stratum

L = number of strata

Ni = number of sample units within stratum i

N = number of sample units in the population

The population mean (μ) is estimated with:

$$\hat{\mu} = \frac{1}{N}\left(N_1\hat{\mu}_1 + N_2\hat{\mu}_2 + \cdots + N_L\hat{\mu}_L\right) = \frac{1}{N}\sum_{i=1}^{L} N_i\hat{\mu}_i$$

where Ni represents the total number of sample units in stratum i, L represents the number of strata, and N represents the total number of sample units in the overall population.

Variance of the estimate is just the weighted average of estimated variances of the mean from a sequence of random samples selected from stratum i through L, although it appears to be a little more complicated:

$$\hat{\text{var}}(\hat{\mu}) = \frac{1}{N^2}\left[N_1^2\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right) + \cdots + N_L^2\left(\frac{N_L - n_L}{N_L}\right)\left(\frac{s_L^2}{n_L}\right)\right] = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

And si^2 is an estimate of the overall population variance from each stratum i through L.

Similarly, estimating the proportion of the population with a particular trait (p) using stratified random sampling involves combining estimates from multiple simple random samples, each generated within a stratum. The population proportion is estimated with the sample proportion:

$$\hat{p} = N_1\hat{p}_1 + N_2\hat{p}_2 + \cdots + N_L\hat{p}_L = \sum_{i=1}^{L} N_i\hat{p}_i$$

Variance of the estimate p^ is:

$$\hat{\text{var}}(\hat{p}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\,\hat{\text{var}}(\hat{p}_i) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}\right)$$

Standard error of p^ is the square root of var^(p^).

Using stratified random sampling necessitates deciding how to distribute a set amount of work among the many strata. There are several methods for assigning sampling effort. Data on variability within each stratum, as well as the relative cost of acquiring and testing a sample unit, aid in increasing overall sampling efficiency [6].

References

1. De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. B AIP Conference Proceedings. INTERNATIONAL CONFERENCE ON INTEGRATED INFORMATION (IC-ININFO 2014): Proceedings of the 4th International Conference on Integrated Information. AIP Publishing LLC. URL: https://doi.org/10.1063/1.4907823.
2. Udofia, E., Buduka, S., Akpabio, J., Egwu, S., Udofia, E., & Olagunju, D. (2020). Digital Transformation: After the Big Data, What Next? B Day 1 Tue, August 11, 2020. SPE Nigeria Annual International Conference and Exhibition. SPE. URL: https://doi.org/10.2118/203614-ms.
3. Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. B. (2019). Data quality in ETL process: A preliminary study. B Procedia Computer Science (Вип. 159, с. 676–687). Elsevier BV. URL: https://doi.org/10.1016/j.procs.2019.09.223.
4. B., S., M., T., & A., A. (2015). Automated ETL Testing on the Data Quality of a Data Warehouse. B International Journal of Computer Applications (Вип. 131, Issue 16, с. 9–16). Foundation of Computer Science. URL: https://doi.org/10.5120/ijca2015907590.
5. Saeed, N., & Husamaldin, L. (2021). Big Data Characteristics (V's) in Industry. B Iraqi Journal of Industrial Research (Вип. 8, Issue 1, с. 1–9). Corporation of Research and Industrial Development. URL: https://doi.org/10.53523/ijoirvol8i1id52.
6. Stratified Random Sampling. (n.d.). cals.arizona.edu. Retrieved December 1, 2022. URL: https://cals.arizona.edu/.
7. Cormode, G., & Duffield, N. (2014). Sampling for big data. B Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '14: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. URL: https://doi.org/10.1145/2623330.2630811.