Data Cleansing: an application of Data Science

Kishan K Ural, Dr. Ravikumar G K, Vidhya S G

UG Student, Dept. of CSE, BGSIT, BG Nagar, Karnataka, India

Professor& R&D Head, Dept. of CSE, BGSIT, BG Nagar, Karnataka, India

Assistant Professor, Dept. of ISE, BGSIT, BG Nagar, Karnataka, India

ABSTRACT: Data science is a multidisciplinary field that uses scientific methods or processes or algorithms and systems to extract knowledge and insights from structures and unstructured data. This project is aimed to develop an efficient data cleansing model using various machine learning algorithms.

Administratively incorrect and inconsistent data will lead to false conclusions and misdirected investments on both public and private scales. It's important to have access to reliable data to avoid incorrect important decisions.

I. INTRODUCTION

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing or deleting the dirty or coarse data.

Data cleansing can be done interactively with data wrangling tools or as batch processing through scripting.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage or by different data dictionary definitions of similar entities in different stores.

Data cleansing differs from data validation, in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

II. EXISTING SYSTEM

When a quality screen records an error, it can either stop the dataflow process, send the faulty data somewhere else than the target system or tag the data. The latter option is considered the best solution because the first option requires, that someone has to manually deal with the issue each time it occurs and the second implies that data are missing from the target system and it is often unclear what should happen to these data.

III. DISADVANTAGES OF EXISTING SYSTEM

- Project costs:

It is expensive to store and retrieve the data.

- Time:

It is not timely.

- Security:

It requires sharing information that is not secure.

IV. PROPOSED SYSTEM

This project is aimed at developing an efficient data cleansing model using various machine learning algorithms In the proposed system, data cleaning is done by software. It must be monitored and inaccurate data sets need to be reviewed. This is why building a model for data cleaning is necessary.

V. ADVANTAGES OF PROPOSED SYSTEM

- It removes major errors and inaccurate datasets that are inevitable during retrieving.
- More reliable and efficient data.
- The ability to offer the best classification in the past/present and future data.

VI. SYSTEM ARCHITECTURE

Data cleansing involves following procedures, namely:

- Data auditing:

The data is audited with the use of statistical and database methods to detect anomalies and contradictions.

- Workflow specification:

The detection and removal of anomalies are performed by a sequence of operations on the data known as the workflow.

- Workflow execution:

In this stage, the workflow is executed after its specification is complete and its correctness is verified.

- Post-processing and controlling:

After executing the cleansing workflow, the results are inspected to verify correctness. Data that could not be corrected during the execution of the workflow is manually corrected, if possible.

Good quality source data has to do with “Data Quality Culture” and must be initiated at the top of the organization. It is not just a matter of implementing strong validation checks on input screens, because almost no matter how strong these checks are, they can often still be circumvented by the users.

There is a four-step guide for organizations that wish to improve data quality.

- Parsing:

For the detection of syntax errors. A parser decides whether a string of data is acceptable within the allowed data specification.

- Data transformation:

It allows the mapping of the data from its given format into the format expected by the appropriate application.

- Data Elimination:

Duplicate detection requires an algorithm for determining whether data consists of duplicate representations of the same entity.

- Statistical methods:

By analyzing the data using the values of mean, standard deviation, range, or clustering algorithms, an expert can find values that are unexpected and erroneous.

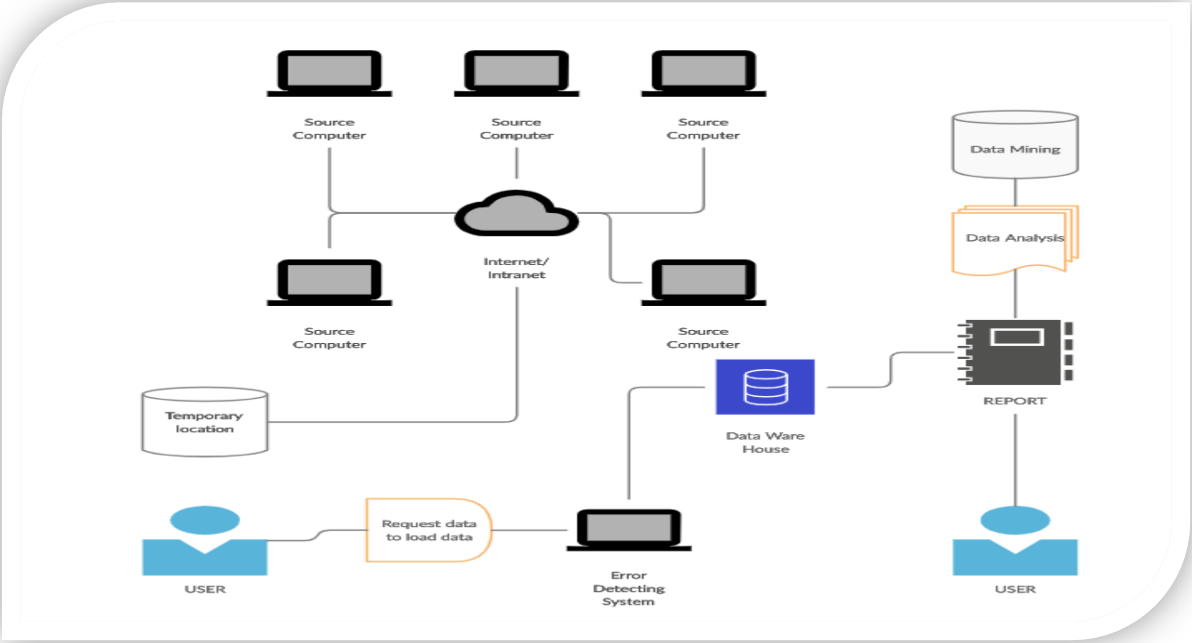

FIG: SYSTEM ARCHITECTURE

VII. METHODS

Data cleansing follows “Super-4 Mechanism” that involves the following procedures.

- Profiling.
- Data cleaning.
- Final datasets and reports.
- Automatization.

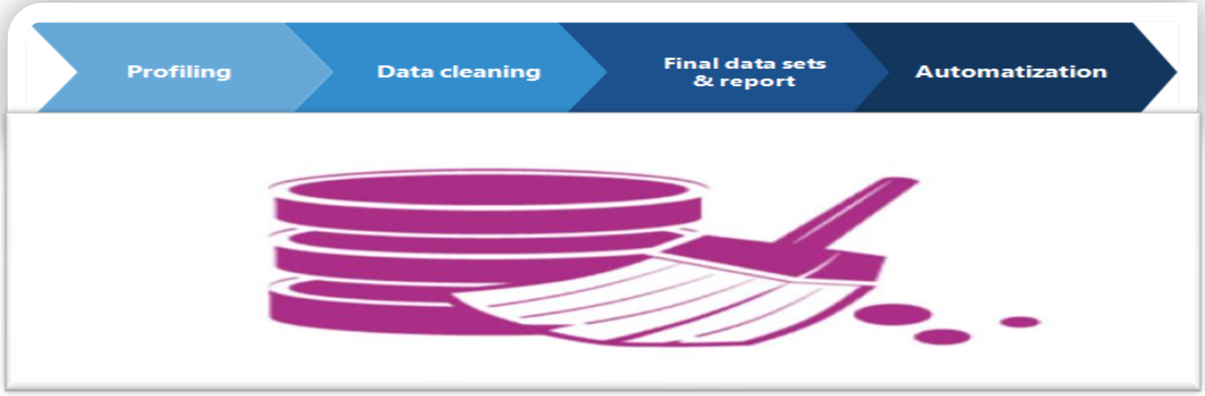

FIG: SUPER-4 MECHANISM

- Profiling: Detect faulty data.
- Data cleaning:Remove/modify the faulty data.
- Final data sets and report:Replacing the old data with modified data.
- Automatization:Model the automation of the above three steps.

VIII. APPLICATIONS

- Less time needed to data unification before each use.
- The correct interpretation of business data.
- Increased data reliability.
- Less time needed for the preparation of data for future analyses.
- Decreased marketing campaign costs as the result of the reduced number of duplicate shipments.

IX. CONCLUSION

It helps in removing major errors and inconsistencies that are fatal when multiple sources of data are getting pulled into one dataset. It will make everyone more efficient and they can quickly access the data whenever they need it.The final model will be user friendly

REFERENCES

1. Doing Data Science Straight Talk from Frontline by Cathy O'Neil, Rachel Schutt.
2. Python Data Science Hand Book By Jake Vanderlin.
3. https://www.researchgate.net
4. http://www.academia.edu
5. http://www.wikipedia.com