What is Data Scrubbing?

Scrubbing is also known as data cleaning. The data cleaning process detects and removes errors and anomalies and improves data quality. Data quality problems arise due to misspelling during data entry, missing values, or any other invalid data.

In basic terms, Data Scrubbing is the process of guaranteeing accurate and correct collection of information. This process is especially for companies that rely on electronic data during the operation of their business. During the process, several tools are used to check the stability and accuracy of documents.

By using data cleansing software, your system will be fed up with unnecessary material that reduces the system.

Reasons for ‘Dirty’ Data Dummy Values:

- Absence of data
- Multipurpose fields
- Cryptic data
- Contradicting data
- Inappropriate use of address lines
- Violation of business rules
- Reused primary keys
- Non-unique identifiers
- Data integration problems
- Why data cleaning or cleansing is required?
- Source Systems data is not clean; it contains certain errors and inconsistencies.
- Specialized tools are available which can be used for cleaning the data.
- Some leading data cleansing vendors include Validity (integrity), Harte-Hanks (Trillium), and First brick.

Data Scrubbing as a Process

1. The first step in data scrubbing as a process is discrepancy detection. The discrepancy can be caused by a number of factors, including human errors in data entry, intentional errors, and data delays. Discrepancies can also arise from consistent data representation and inconsistent use of code.

After detecting the discrepancy, we will use the knowledge we already have about the properties of the data to find the noise, extrinsic, and abnormal values that need to be investigated.

Data about unique rules, consistent rules, and null rules should also be examined.

- A unique rule states that each value of a given attribute must be different from all other values for that attribute.
- A consecutive rule states that there can be no missing value between the lowest and highest value for an attribute and all values must be unique.
- A null rule specifies the use of a blank, question mark, special character, or other string that represents null conditions and how such values should be handled.
- The null rule should specify how to record the null condition.

2. Once we find discrepancies, we typically need to define and apply the transformation to correct them. The two-stage process of anomaly detection and data transformation. Some changes may introduce more discrepancies.

The new method of data scrubbing emphasizes increasing inhumanity. In this tool, the change can be specified as an underline. The results are immediately shown on the record appearing on the screen. The user can choose to undo the change so that the change that introduces additional errors can be erased.

Steps in Data Cleansing/Scrubbing

1. Parsing: Parsing is a process in which individual data elements are located and identified in source systems and then these elements are separated into target files. For example, parsing of name into the First name, Middle name, and Last name or parsing the address into a street name, city, state, and country.

2. Correcting: This is the next step after parsing, in which individual data elements are fixed using data algorithms and secondary data sources. For example, in the address attribute replacing a vanity address and adding a zip code.

3. Standardizing: In standardization, process conversion routines are used to transform the data consistent format using both standard and custom business rules. For example, the addition of a prename, replacing a nickname, and using a preferred name.

4. Matching: The matching process involves eliminating duplication by searching for records with parsed, corrected, and standardized data using certain standard business rules. For example, identifying similar names and addresses.

5. Consolidating: Consolidation involves merging the records into one representation by analyzing and identifying the relationship between the recorded records.

6. Data Scrubbing must deal with many types of eventual errors:

- There may be many errors in the data such as missing data, or incorrect data on one source.
- When more than one source is involved there is a possibility of inconsistency and conflicting data. So Data Scrubbing must deal with all these types of errors.

7. Data Staging:

- Data staging is an interim step between data extraction and the remaining steps.
- Data is stored from asynchronous sources, using various processes such as native interfaces, flat files, FTP sessions.
- After a certain predefined interval, data is loaded into the warehouse after the transformation process.
- No end-user access is available to the staging file.
- For data staging, the operational data store may be used.

Importance of Data Scrubbing

- **More Storage Space**: Given that we are removing all those unnecessary entries, we are freeing up a considerable amount of storage space for all our other data.
- **Much more Accurate**: Through using this software program, our database is suitable for executing more accurate and accurate data. This will also help you to get more relevant information in less time.
- **Low marketing cost**: This is achieved by executing the method of extracting duplicate documents from the data source, resulting in reduced ad shipping costs.