International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Data Cleaning in Cloud Platform

V Ramya, Jayasimha S R

ABSTRACT: ***Data is very valuable and it is generated in large volumes. The Use of high-quality data for making quality decisions has become a huge task which helps people to make better decisions, analysis, predictions. We are surrounded by data with errors, Data cleaning is a delayed, complicated task and considered costly. Data polishing is important since it is necessary to remove errors from the data before transferring to the data warehouse since poor quality data is eliminated to get the desired results. The Error-free data will produce precise and accurate results when queried. Hence consistent and proper data is required for the decision making. The characteristics of data polishing is data repairing and data association. Identifying the homogeneous object and linking it to the most associated object is defined as Association. The process of making the database reliable by repairing and finding the faults is defined as repairing. In the case of big data applications, we do not use all the existing data, we use only subsets of appropriate data. Association is the process of converting extensive amounts of raw data to subsets of appropriate data that are useful. Once we get the appropriate data, the available data is analyzed and it leads to knowledge [14]. Multiple approaches are used to associate the given data and to achieve meaningful and useful knowledge to fix or repair [12]. Maintaining polished quality of data is referred to as data polishing. Usually the objectives of data polishing are not properly defined. This paper will discuss the goals of data cleaning and different approaches for data cleaning platforms.***

KEYWORDS: ***polishing, Clustering, Association, Deduplication, Repairing***

I.INTRODUCTION

Data is considered a very beneficial resource. The availability of High-grade data helps the organization to make better decisions and help in better prediction and analyses. low-grade data is not considered appropriate to accomplish its intended purpose [2]. Maintaining polished quality of data is referred to as data cleaning. Data cleaning includes fixing or removal of damaged or inaccurate data which is mandatory for Big Data as the errors may result in poor conclusions [3]. The fundamental of data polishing is repairing the data and associating. Association identifies a similar object and links it to the associated objects, data repairing makes the database reliable by finding and fixing the faults [13]. A limited amount of data is used in big data applications. In majority cases, we use only subsets of appropriate data. Association is the process of converting extensive amounts of raw data to subsets of appropriate data that are useful. Once we get the appropriate data, the available data is analyzed and it leads to knowledge [12].

Manuscript received on April 02, 2020.
Revised Manuscript received on April 20, 2020.
Manuscript published on May 30, 2020.

***** Correspondence Author

V Ramya*, PG Student Department of Master of Computer Applications RV College of Engineering®, Bangalore

Jayasimha S R, Assistant Professor Department of Master of Computer Applications RV College of Engineering®, Bangalore

Maintaining polished quality of data is referred to as data polishing. Usually the objectives of data polishing are not properly defined. Data should be accurate before using it. Hence Datasets from multiple origins of data need to be polished before combining [4]. Data generated for one purpose and used for another purpose is not considered as good quality. Data generated from Social media arrives with ambiguous and corrupt information, this leads to a concern when extensive heterogeneous data are combined from different sources. faults in the database will result in reporting errors. Hence these faults lead to inappropriate decisions by the organization [5]. Faults in data come from multiple sources, example, when a User enrollment records are ingested many important fields like Address, Email ID or the phone number may be missing or wrong which is mandatory for delivering the product for the customer and the sales agent, do not attempt to figure out the appropriate or the missing data and ingest some random default value [13]. So, the information about the Customer is incorrect. Massive number of faults are encountered in the warehouse when data from multiple sources are combined with it, then it results in quality issues after merging [1].

II.LITERATURE REVIEW

2.0. Generic data cleaning tasks

2.0.1 Record Matching

The objective of matching records is to discover if the records within one or multiple relations constitute the actual real-world entity, called "matching". The record matching function should be solved when we dedupe records in specific relations. Record matching is defined in numerous ways [2]. The objective of matching records is to match each record from the collection of records with records in other tables. It is usually done when new sets of entities are ingested to the desired relations and assure that the insertion task does not result in duplicate copies of entries in the resulting target or destination relation.

2.0.2 Schema Matching

Schema matching is done before record matching. Schema matching is the task of arranging attributes across multiple schemas [4].

2.0.3 Task Deduplication

Target of deduplication is consolidating the records that speak to a similar certifiable entity. duplication can be inexactly alluded to as a fluffy or surmised variation of the socially selected unmistakable activity. It has a contribution as a table and a lot of segments; the yield is the segment of this table where every individual gathering speaks to a lot of records that are practically equivalent to the predefined columns [2]. The target of deduplication is to consolidate records in a table with the end goal that every blend of the gathering of records speaks to a similar substance. This task is performed when the database is being populated or cleaned at the underlying time.

2.2 Generic Data cleaning operators

2.2.1 Similarity Join

The target of record matching is to coordinate the pair of records over numerous relations. The work incorporates numerous forecasts and one significant forecast is to quantify closeness between the records [3]. The objective of similarity join is to distinguish comparative records across relations, where it is estimated by closeness function [20]. The similarity join is characterized as social join and the state of the join is characterized by the closeness work likeness join can be communicated in organized inquiry language by characterizing the join predicates utilizing the UDFs [10]. Set-Similarity Join administrator is an essential crude and it is utilized for similitude joins on different string similitude capacities.

2.2.2 Clustering

The bunching is utilized in information duplication and different information cleaning undertakings. Clustering is characterized as the activity of isolating a lot of things and consolidating them into comparative gatherings dependent on similarity [5]. For instance, a rundown of clubs might be grouped dependent on comparable cooking styles, or dependent on their expense, or the mix of cost and food.

2.2.3 Parsing

Parsing is characterized as fragmenting an information string into its trait values [6]. The sectioned information records are contrasted and are embedded with the objective table whenever required.

2.3 Data Standardization

Information normalization is done before information cleaning or cleaning errands and is considered as a basic activity, for example, record coordinating or information deduplication. Amending the quality qualities and normalizing the arrangement can bring about exactness in different information cleaning undertakings, for example, deduplication and record coordinating [20].

2.4 Data profiling

Cleaning the information is a ceaseless and dull procedure. The nature of information is a significant factor and it is important to check the quality before starting the information cleaning or cleaning process, and accordingly evaluates its success [9]. The way toward checking the nature of information is characterized as information profiling, it includes gathering different totaled information measurements. A casual target of information quality is to guarantee that the qualities coordinate with the desires.

III.FUNDAMENTALS

3.0 Implementation of Data Cleaning Operations and tasks

3.0.1 Generic data cleaning tasks

Record matching, schema matching and Task Deduplication

Schema matching aligning the schemas across various relations [11]. As a model, Given a connection R(R1,R2,...Rm) acquires a mapping of R to a connection R(K1, K2,...Km)[9]. Here we need to populate the tuples from R(City,Pincode ,Nation,Mobilenumber) into connection K(Name,CityAddress,Country,Phone). Here it includes two difficulties (1) first is to decide the pair of characteristics in K and R alluding to same ideas for instance (Phone, Mobilenumber) (Company, Name) (City, CityAddress) (Nation, Country) (2) The subsequent test is to utilize the trait correspondences and form them to get a capacity to change over tuples of R into tuples of for Example. City Address in K is acquired from R by connecting street, city, and pin code also the portable number is obtained by legitimately taking the telephone from K [20]. The first subtask above is typically alluded to as blueprint coordinating while the subsequent assignment is alluded to as pattern mapping. The Attribute name, Attribute worth, and Attribute names are considered for figuring the similitude or relatedness score. Then, sets with a high score are held.

Record matching is done post blueprint coordinating, which guarantees that the characteristics in the two relations have been adjusted appropriately [20]. In the Enterprise Data Warehousing Scenario: when a new clump of client records is being brought into a good sales database. For this situation, it is essential to check if a similar client is spoken to in the current just as the approaching sets, and we should just hold one record in the conclusive outcome. Because of illustrative contrasts and blunders, records in the two groups could be extraordinary and may not coordinate precisely on their key traits (e.g., name and address or the Customer ID) [10].

Table 2.1 sets of customer data

ID.NO	FIRST NAME	CITY	PIN
001	Abc textiles	Bangalore	560093
002	Cable services	Delhi	567902
111	Abc textiles ltd	Pune	467783
112	Cable TV services	Delhi	567902

Table 2.1 represents customer data with same entries with different FIRST NAME resulting in redundancy, hence using record matching, it should be eliminated and only a unique set of records should be retained in the master table.

Data Deduplication is the process where data blocks are analyzed to identify duplicate blocks where it stores only one copy of it and deletes the rest.

As of now utilization of cloud storage is expanding and to conquer expanding information issues, Data deduplication procedures are utilized [20]. Also the Cloud storage administration is given by outsider cloud suppliers along these lines security of information is required.

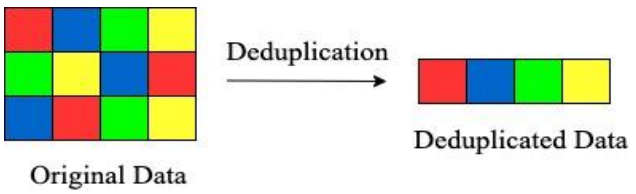

Fig. 3.1 deduplicated data

Fig. 3.1 shows the original data with duplicate blocks and using the deduplication it stores only one copy of each and deletes the other repeating blocks.

Identifying and eliminating duplicate blocks of similar data present within the set is referred to as data deduplication [20]. It is a similar approach to compression of data, which identifies the duplicate and repeating blocks in one particular record. Deduplication finds repeating blocks of similar data among files and records from multiple directories, locations, and different types. The way the deduplication works is that the data is chopped into chunks or slices, a slice is one or multiple contiguous blocks of information and a series of slices will be compared against all the previous slices of data [10]. The cryptographic hashing algorithm like SHA-1 or SHA-2 and SHA-256 creates a hash and chunks or slices of data are run through this algorithm. For example, "The hungry lion jumps over the zebra" into an SHA-1 hash calculator and considers the following 2FD4E1C67A2D28FCED849EE1BB76E7391B93EB12. If the hashes of these chunks or slices of data match, they are considered as identical and a minute change results in the hash to change. A SHA-1 hash is 160 bits. If we are creating a 160-bit hash for a 10 MB chunk of data [19], we can save 10 MB every time we backup that similar chunk of data. Hence dedupe is referred to as a memory saver.

3.0.2 Generic Data cleaning operators

Graph-based clustering is a method used to identify groups of similar cells or samples. There are no prior assumptions about the clusters in the data, meaning the number, size, density, and shape of clusters does not need to be known or assumed before clustering. Graph-based clustering is useful for identifying clustering in complex data sets [10].

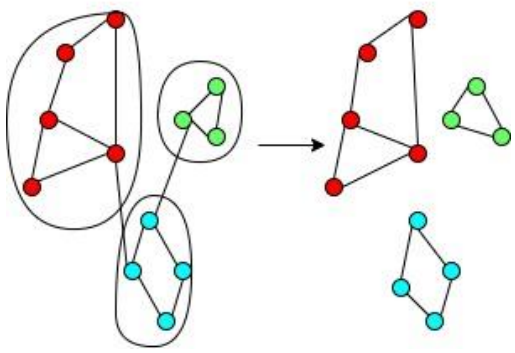

Fig. 3.2 Graph-based clustering

Fig. 3.2 shows the clustering of similar cells or samples using the graph-based clustering.

Minimum spanning tree

- Obtains MST for the input
- Removing the k-1 heaviest edges
- Resulting in k clusters

Spanning Tree is a connected subgraph with no cycles that includes all vertices in the graph as shown in the following example.

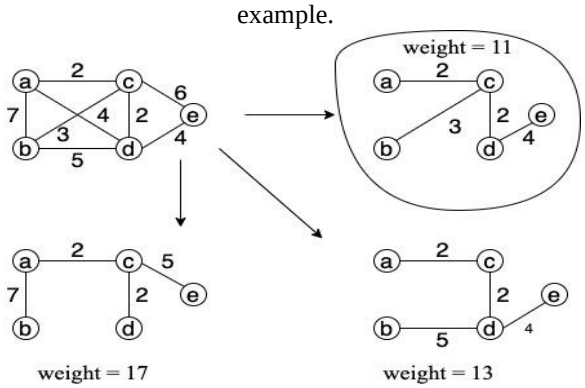

Fig. 3.3 Minimum spanning tree

Fig. 3.3 shows the calculation of the weights for each graph and the graph with minimum weight is chosen and the ones with highest weights are ignored.

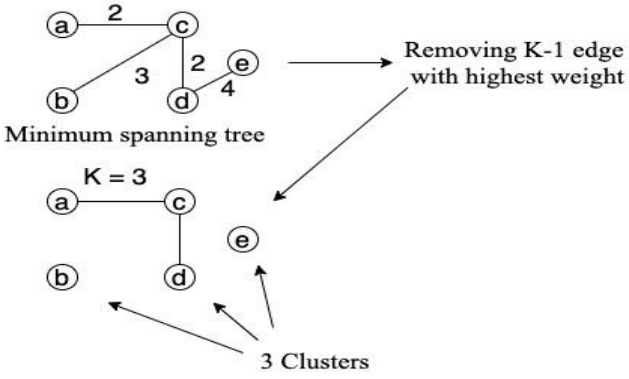

Fig. 3.4 k Spanning tree (3 clusters)

Fig. 3.4 shows the graph with the minimum spanning tree and removing the k-1 edge hence resulting in the k clusters for the graph.

4.0 cleaning and transforming data using queries

SQL can help expedite this task. Different functions commonly used to clean, transform, and remove duplicate data are by using String and other functions like SUBSTRING(), LTRIM(), RTRIM(), LEFT(), RIGHT(), CHARINDEX(), LEN(), ISNUMERIC(), ISDATE().SQL Server provides some general-purpose SQL string functions for extracting and overriding strings.

The use of the WHERE and HAVING clauses in a SELECT statement control the subset of the output from the source tables.

- Compare search conditions using comparison operators
- Range search using between, and, and or clause
- List search using in operator and or clause.

IV. CONCLUSION

The Process of Fixing unfit, inappropriate, or imprecise data is referred to as data cleaning. Cleaning of data is considered very important in Big Data as inappropriate data results in faulty analysis, predictions and may lead to undesired conclusions, the data cleaning framework will upgrade the quality of data, In the above paper, we discussed different features of data cleaning technology, with goals and efficient techniques to implement productive solutions. Data cleaning solution addresses several critical high-level tasks including matching the records, duplication, and data parsing. The objective of matching the records is to successfully and precisely match the pair of records among different relations to evaluate if they are equivalent semantically. The task can be personalized to use few similarity functions or filters. The record matching task is used to check for duplicates in the ingested records and to avoid the ingestion of duplicates. The objective of data deduplication or data duplication is to combine the records concerning where the individual group of records represents a similar real-world entity. Extracting the attribute values from the input record or string before pushing them into the destination relation is the objective of parsing. For example, customers or products. Many times, input data or records are extracted from an unknown source and so the formats vary. The SSJ operator accurately matches the pair of records across relations and is used to precisely implement SSJ across relations by using a variety of similarity functions. Hence, a developer can easily build on top of the SSJ to implement an effective and precise solution for matching the records task. The clustering operation is used to group multiple records across a relation and allows the developers to develop an effective and precise solution for the duplication task. In summary, we implemented an overview of the multiple problems with approaches, and different techniques and solutions that are being developed for data polishing. We expect the data polishing will continue to develop in the future in research and commercial domains.

REFERENCES

1. G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, “Improving data quality: Consistency and accuracy,” in Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB ’07, pp. 315–326, 2005.
2. Hong Liu, Ashwin Kumar TK, Johnson P Thomas, Xiaofei Hou,“cleaning framework for bigdata,” IEEE Second International Conference on Big Data Computing Service and Applications 2016.
3. C. Mayfield, J. Neville, and S. Prabhakar, “ERACER: A database approach for statistical inference and data cleaning,” in Proceedings of the 2010 International Conference on Management of Data, SIGMOD ’10, pp. 75–86, 2010.
4. W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7(1):81–99, 2003.
5. Peter Christen. Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, 2012a. DOI: 10.1007/978-3-642-31164-2.
6. Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. Knowl. and Data Eng., 24(9):1537–1555, 2012b. DOI: 10.1109/TKDE.2011.127. 55.Mikhail
7. Bilenko, Beena Kamath, and Raymond J. Mooney. Adaptive blocking: Learning to scale up record linkage and clustering. In Proc. 2006 IEEE Int. Conf. on Data Mining, pages 87–96, 2006. DOI: 10.1109/ICDM.2006.13. 55.
8. Surajit Chaudhuri, Venkatesh Ganti, and Raghav Kaushik. A primitive operator for simi- larity joins in data cleaning. In Proc. 22nd Int. Conf. on Data Engineering, 2006b. DOI: 10.1109/ICDE.2006.9. 28.
9. Jiawei Han and Micheline Kamber. Data mining: concepts and techniques. Morgan Kaufmann, 2006. 34.
10. Venkatesh Ganti, Anish Das Sharma Data cleaning practical perpective. Morgan & claypool publisher
11. Philip A. Bernstein, JayantMadhavan, Erhard Rahm: Generic Schema Matching, Ten Years Later. PVLDB 4(11): pp. 695-701, 2011.
12. Surajit Chaudhuri, Anish Das Sarma, Venkatesh Ganti, and Raghav Kaushik. Leveraging ag- gregate constraints for deduplication. In Proc. ACM SIGMOD Int. Conf. on Management of Data, pages 437–448, 2007. DOI: 10.1145/1247480.1247530.
13. Lukasz Ciszak: Application of Clustering and Association Methods in Data Cleaning, Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 97 – 103.
14. Morteza Alipour Langouri, Zheng Zheng: Contextual data cleaning, 2018 IEEE 34th International Conference on Data Engineering Workshops.
15. X. Chu, I. F. Ilyas, and P. Papotti, “Holistic data cleaning: Putting violations into context,” in *ICDE*, 2013, pp. 458–469.
16. E. Rahm and H. H. Do. Data cleaning: Problems and current approaches
17. Elgamal, N. Mosa, and N. Amasha. Application of framework for data cleaning to handle noisy data in data warehouse.
18. C. Mayfield, J. Neville, and S. Prabhakar, “ERACER: A database approach for statistical inference and data cleaning,” in Proceedings of the 2010 International Conference on Management of Data, SIGMOD ’10, pp. 75–86, 2010.
19. MENG Jian , DONG Yi-sheng , WANG Yong-li. “A Rule- based Interactive Data Cleaning Technique”. Computer Technology and Development,2005,15(4):141-144.
20. Vinayak Borkar, Kaustubh Deshmukh, and Sunita Sarawagi. Automatic segmentation of text into structured records. In Proc. ACM SIGMOD Int. Conf. on Management of Data, pages 175–186, May 2001. DOI: 10.1145/376284.375682.

AUTHORS PROFILE

Ramya, BSC (CS), MCA VI semester from RV College of Engineering®, Bangalore. She is currently working as a data engineer intern in Bidgely Inc. Her areas of Interest include data analytics and cloud computing.

Prof. Jayasimha S R, working as an assistant Professor in the department of MCA at RV College of Engineering®, Bangalore. He served in the institution from 2013 to till the date. Currently he submitted his PhD Thesis to the VTU. His area of interest is cloud computing. He published more than 20 papers in national, international conferences and international Scopus indexed journals.