International Journal of Applied Research 2015; (7): 388-390

ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2015; 1(7): 388-390
www.allresearchjournal.com
Received: 15-04-2015
Accepted: 17-05-2015

Monika
Student, M.Tech Dept. of Computer Science & Engg., Manav Institute of Technology & Management, Jevra, Hisar(Haryana)

Madhurima
Guide, Assistant Professor, Dept. of Computer Science & Engg., Manav Institute of Technology & Management, Jevra, Hisar (Haryana)

Dr. Vijay Bhardwaj
Co-Guide, Associate Professor & HOD, Dept. of Computer Science & Engg., Manav Institute of Technology & Management, Jevra, Hisar (Haryana)

Correspondence:
Monika
Student, M.Tech Dept. of Computer Science &Engg., Manav Institute of Technology & Management, Jevra, Hisar(Haryana)

A vague assisted association analysis approach to repair Bigdata impurities

Monika, Madhurima, Vijay Bhardwaj

Abstract
Big data consist of huge amount of data. There are many of challenges when we accessing Big data. One of such major challenge is identified in terms of dataset impurities. These impurities are visible sometimes in terms of missing or incomplete information. But sometimes this kind of impurities is hidden in terms of non-valuable attribute or unnecessary information. In this paper, a two layered work is defined to improve the dataset integrity. In first phase, the analysis over the dataset is performed and later on the impurities are removed. Once the problems are identified, the particular attribute or the tuple are removed from the dataset. To verify the dataset integrity, the association rules are generated.

Keywords: Big data, Vague Association Approach, Hadoop, Mapreduce.

Introduction
Big data is the collection of data sets large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications. Big data are about turning unstructured, invaluable, imperfect, complex data into usable information [1].

Big data is most advanced form of data representation and storage with relative query management in the relative environment. Big data is able to process on large amount on unstructured data patterns. These kinds of patterns are identified respective to the data management, visualization, analysis, prediction and the data modeling capabilities. The information structure is here derived with storage specification, handling and the relative information derivation. The basic structure of big data is shown in figure

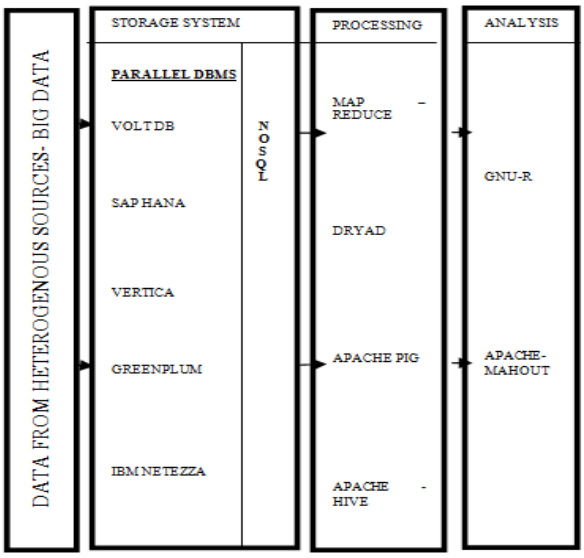

Fig 1: Architecture of Bigdata

The big data processing also includes the architectural specification in the distributed architecture with the specification of a centralized view defined with the level constraints. These constraints are relative to the environment, application and the processing activity so that the more relative decision can be taken. This kind of information processing is done under architecture rule generation and derivation.

2. Related Work

Lot of researchers provided different work model provides the solution for big data integrity. Some of the contributions of earlier researchers have been discussed in this section. EdmonBegoli[1] present three system design principles that can inform organizations on effective analytic and data collection processes, system organization, and data dissemination practices.. KapilBakshi[2] focus on the unstructured data like imagery, sensors, telemetry, video, documents, log files, and email data files. AvitaKatal, Mohammad Wazid& R.H. Gourdar[3] describe about the Big Data and its importance in the current and upcoming market because today everywhere people like to work on the system and use large database. Antonia Azzini[4] exploits two different procedures. The first one is aimed at computing the mismatch among the data sources to be integrated. The second uses mismatch values to extend data to be processed with a traditional map reduce algorithm. ZibinZheng[5], present how to store, manage, and create values from the service-oriented big data become an important research problem. Anita A. Parmar, Udai Pratap Rao, Dhiren R. Patel [6], propose a blocking based approach for sensitive classification rule hiding. First we find the supporting transactions of sensitive rules. Then we replace known values with unknown values in those transactions to hide a given sensitive classification rule. Finally the sanitized dataset is generated from which sensitive classification rules are no longer mined.Dr. PayalPahwa[7] classify the methods in two categories domain dependent and domain independent. This paper presents a survey and review of data cleansing methods, classification of existing methods and comparison between them.Aditya B. Patel[8] have done prototype implementation of Hadoop cluster, HDFS storage and Map Reduce framework for processing large datasets by considering prototype of big data application scenarios. Seref SAGIROGLU [9] presents an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern on it. Timothy E. Ohanekwu[10] proposes a technique that eliminates the need to rely on match threshold by defining smart tokens that are used for identifying duplicates. This approach also eliminates the need to use the entire long string records with multiple passes, for duplicate identification.

3. Research Methodology

The presented work is about to generate the associated rules for RFID data. RFID dataset is generally collected from multiple sources because of this dataset is having number of related impurities. This work is about to remove these dataset impurities and generate the effective rules over the database. The work is divided in three main stages. In first stage, high level analysis is performed over the dataset to generate the effective results. This analysis includes the vertical pruning under the support and confidence value analysis. Once the vertical pruning is done, in second stage, the dataset partitioning is done to perform the parallel processing over the dataset. Once the analysis is done, in next stage, the generation of rules will be done. Here the vague improve approach is defined for filtration of the dataset. As we Defined the complete process of Bigdata Cleaning and classification is divided in the following Steps. The basic step wise work procedure is shown in figure

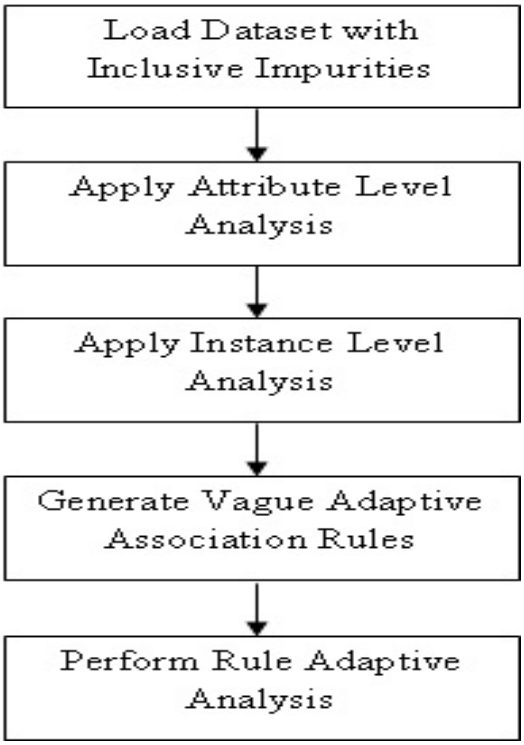

Fig 2: Work Model

Firstly load the dataset. The dataset can be collected from various locations so that the data with inclusive deficiencies will be obtained. This kind of dataset will have number of integrated impurities. This work is here defined to identify these big data associated impurities problem at the early stage so that the improvement to the data processing will be done. The work model is to identify the attributes that are not having the significant number of entries and having the lesser contribution in instance or record representation. This kind of analysis is performed to identify the missing values under specific attribute.

This attribute level analysis is also performed at the association level so that the effective feature adaptive analysis will be obtained. Now identify the instances that are not having the associated value. To improve the significance ofthe dataset, it is required to remove such values from the dataset. If the associated is low complete data values will be removed from the dataset.

4. Result

The presented work is here defined to improve the distributed Bigdata system by removing the integrated impurities. The improvement is here obtained in terms of effectiveness analysis and the ruleset generation over the dataset. The time effective analysis obtained from the work on different size dataset is given in this section

Time Based Analysis

As the work is derived from external web source, the analysis of the work is here performed to improve the data

integrity and efficiency. To improve the efficiency the partition adaptive approach is applied. The work is applied on different size datasets. The analysis results are here shown in table.

Table 1: Time Adaptive Analysis

Number of Records	Time Taken (Seconds)
1000	2.1
2000	3.4
3000	4.9
4000	5.4
5000	6.8

Here table is showing the analytical results obtained on different size data processing. The results are taken in terms of time taken by the work

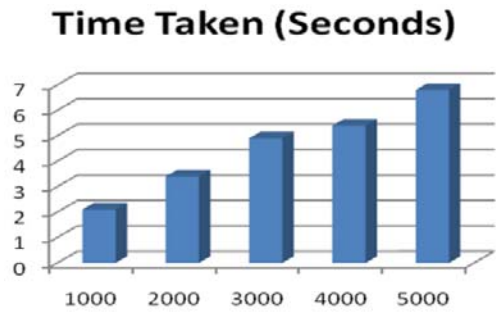

Fig 3: Time Based Analysis

Here x axis shows the size of dataset and y axis shows the time taken. The figure shows that as the number of records increases, the time taken by the work also increased.

Dataset Pruning

Here table is showing the analytical results obtained on different size data processing. The results are taken in terms of time taken by the work

Table 2: Pruned Result

Number of Records	Pruned Records
1000	78
2000	189
3000	356
4000	625
5000	896

Here table is showing the analytical results obtained on different size data processing. The results are taken in terms of time taken by the work

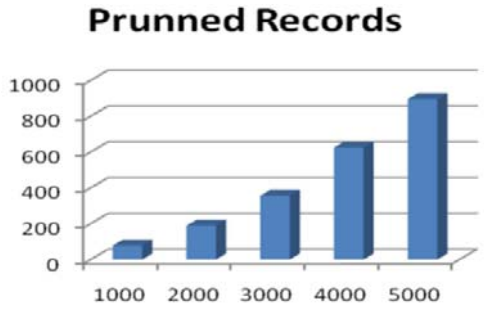

Fig 4: Pruning Analysis

Here figure is showing the pruning analysis on different size dataset. Here x axis shows the size of dataset and y axis shows the number of pruned records. The figure shows that as the number of records increases, the chances of pruning increased very fast.

5. Conclusion

In this paper, vague adaptive approach is used to remove the big data impurities & provide the significant information from the dataset. The filtration is applied to remove the associated dataset impurities. The high level impurities are identified in terms of missing values and such attributes and tuples are removed. Later on value specific association analysis is applied to identify the integrated impurities. The association rules are here generated to obtain the aspect driven values. The results show that the work is able to derive significant rules from the filtered dataset.

6. References

1. EdmonBegoli, James Horey.Design Principles for Effective Knowledge Discovery from Big Data, Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture, 2012.
2. KapilBakshi. Considerations for Big Data: Architecture and Approach, IEEE, 2012.
3. AvitaKatal, MohammdWazid, GoudarRH.Big Data: Issue, Challenges, Tools and Good Practices, IEEE, 2013.
4. Antonia Azzini. Consistent Process Mining Over Big Data Triple Stores, IEEE International Congress on Big Data 978-0-7695-5006-0/13 © 2013 IEEE, 2013.
5. ZibinZheng, "Service-generated Big Data and Big Data-as-a-Service: An Overview", IEEE International Congress on Big Data 978-0-7695-5006-0/13 © 2013 IEEE, 2013.
6. Anita AParmar,Udai Pratap Rao, Dhiren R Pate.Blocking based approach for classification Rule hiding to Preserve the Privacy in Database, International Symposium on Computer Science and Society, 978-0-7695-4443-4/11© 2011 IEEE,2011.
7. Dr. PayalPahwa. Domain Dependent and Independent Data Cleansing Techniques, IJCST ISSN: 2229 – 4333.
8. Aditya PatelB.Addressing Big Data Problem Using Hadoop and Map Reduc, 2012 Nirma University International Conference on Engineering.
9. Seref SAGIROGLU, Big Data: A Review, 978-1-4673-6404-1/13 ©2013 IEEE.
10. Timothy OhanekwuE. A Token-Based Data Cleaning Technique for Data Warehouse Systems, 2010, 1-7.