INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

DESIGN AND IMPLEMENTATION OF A KNOWLEDGE RETERIVAL SCHEME FROM BIG DATA FOR AVIATION INDUSTRY

[1]Vijay laxmi , [2]Archana Sandhu

[1]P.M. College of Engineering, DCRURST, Murthal
Kami, Sonipat-131001, India
malikvijaysonu@gmail.com
[2]P.M. College of Engineering, DCRURST, Murthal
Kami, Sonipat-131001, India
asandhu43@gmail.com

Abstract:- *In this paper, we describe a middle layer architecture that defined to perform the query analysis and the assessment of the big data. The work implemented on dynamic generated data section instead of whole database. The concept of Big Data actually concerns with a bulk of data presented in large volume with complex architecture and with increasing dataset. The data in such system can be taken from multiple sources and sometimes from autonomous sources. With the development of new centralized system, cloud environments, the use of Big Data is directly available to the end users so the criticality there exists in terms of fast retrieval of data from the system. To perform the analytical information retrieval from such data system there is the requirement of data driven model. In this present work, one of such middle layer model is been presented to derive the valuable and predictive information from big data. The presented model will store the all aspects of dataset in the form of meta data so that the user query will be performed on the selective dataset instead of whole dataset. As the user query will be performed, the analysis of the query will be performed on this meta data information and identify the most relative attribute set and the data section that can answer the query in effective way. Now the query will be performed on this partial dataset instead of whole dataset. . The presented work is divided in 4 stages. In first stage, the big data assessment will be done, in second stage, the relation between the meta data and user query will be established. In third stage, the dynamic sectioning of data will be performed and finally the analytical information will be retrieved from the dataset based on user query. The presented work will be implemented on aviation dataset. The work will use the java as the front end and will use the oracle or mysql as the backend.*

General Terms: - *Algorithms, Design, Performance.*
Keywords: - *Big Data, Data Analytics, Data Warehouse, Data Stream Management, Data Mining.*

1. INTRODUCTION

As the amount of data around us is increasing at an enormous rate. Organization extract rules, information, knowledge from these data to increase innovation, retain customer and increase operational efficiency. So, there was the need to extract information from these large growing data sets. A proper algorithm needs to be defined to increase the computation. And so that the result are driven in a less time. Data Mining is the technology to extract the knowledge from the data. It is used to explore and analyze the same. The data to be mined varies from a small data set to a large data set i.e. big data. Data Mining has also been termed as data dredging, data archaeology, information discovery or information harvesting depending upon the area where it is being used. The data Mining environment produces a large volume of the data. The information retrieved in the data Mining step is transformed into the structure that is easily understood by its user. Data Mining involves various methods such as genetic algorithm, support vector machines, decision tree, neural network and cluster analysis, to disclose the hidden patterns inside the large data set. Data mining process include understanding the business requirements and needs. While understanding the business requirement both data and business requirements are understood. Then, using this business requirement it identifies data source and data format in this the data is prepared and modelled for evaluation, and then using these data source and data format it build data model. This data model is used to build data structure. Then, the mining operation is performed on this data structure. Data mining field comprises of four main disciplines: Statistics: defines tools for measuring significance in the data. Machine learning: provide algorithm to induce knowledge from the data. Artificial intelligence: involve knowledge for encoding and search techniques. Data management and databases: provides an efficient way of accessing and maintaining data. And before getting into the algorithm for data mining to find the association rule, we discuss about what the big Data is. Why it is useful. And will study about some of the work done to handle big data. Then will discuss about Apriority and our work.

1.1 BIG_DATA

Big data 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, so data mining tools were used. Big Data are about turning unstructured, invaluable, imperfect, complex data into usable information. Data have hidden information in them and to extract this new information; interrelationship among the data has to be achieved. Information may be retrieved from a hidden or a complex data set. Browsing through a large data set would be difficult and time consuming, we have to follow certain protocols, a proper algorithm and method is needed to classify the data, find a suitable pattern among them. The standard data analysis method such as exploratory, clustering, factorial, analysis need to be extended to get the information and extract new knowledge treasure.

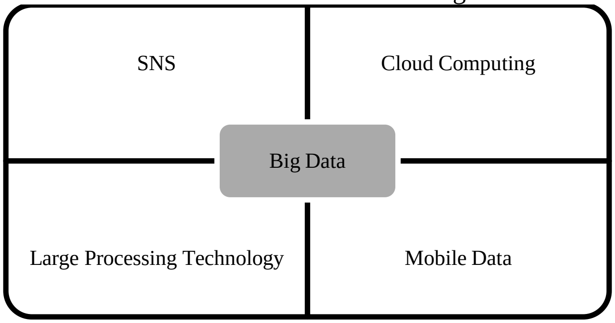


Fig. 1: Big Data modes

The benefit of analyzing the pattern and association in the data is to set the trend in the market, to understand customers, analyze demands and predict future possibilities in every aspect. It helped organizations to increase innovation, retain customers, and increase in operational efficiency. Big Data can be measured using the following that is categorized by 4 V's: variety, volume, velocity and value. Big Data architecture typically consists of three segments: storage system, handling and analysis.

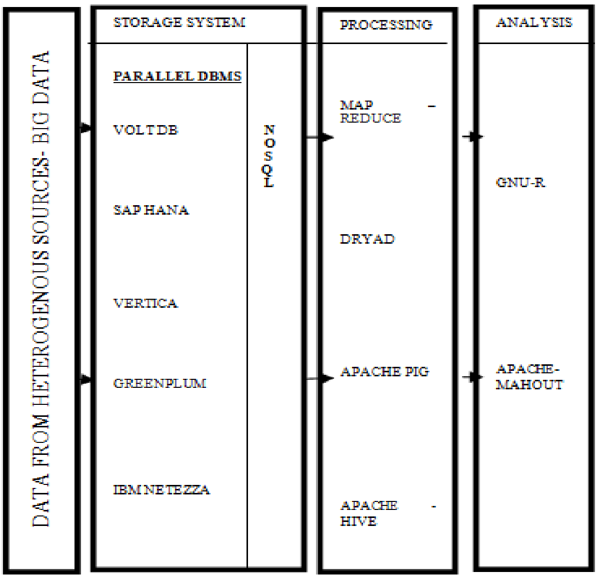


Fig. 2 Architecture of Big Data

Big Data typically differ from data warehouse in architecture; it follows a distributed approach whereas a data warehouse follows a centralized one.

One of the architecture laid describes about adding new 6 rules were in the original 12 rules defined in the OLAP system defined the methods of data mining required for the analysis of data and defined SDA (standard data analysis) that helped analysis of data that is in aggregated form and these were much well timed in comparison with the decision taken in traditional methods.

The Data Mining termed Knowledge discovery, in work done in "Design Principles for Effective Knowledge Discovery from Big Data", its architecture was laid describing extracting knowledge from large data. Data was analyzed using software Hive and Hadoop. For the analysis of data with different format cloud structure was laid. Then there arrived the prime need to analyze the unstructured data, so the Hadoop framework was being laid for the analysis of the unstructured large data sets. Many algorithms were defined earlier in the analysis of large data set. Will go through the different work done to handle Big Data. In the beginning different Decision Tree Learning was used earlier to analyze the big data. In work done by Hall. Et al. there is defined an approach for forming learning the rules of the large set of training data. The approach is to have a single decision system generated from a large and independent n subset of data. Whereas Patil et al, uses a hybrid approach combining both genetic algorithm and decision tree to create an optimized decision tree thus improving efficiency and performance of computation.

1.2 ARCHITECTURE FOR DATA MINING

To best apply these advanced techniques, they must be fully integrated with a Big data as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic Big data can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates architecture for advanced analysis in a large Big data.

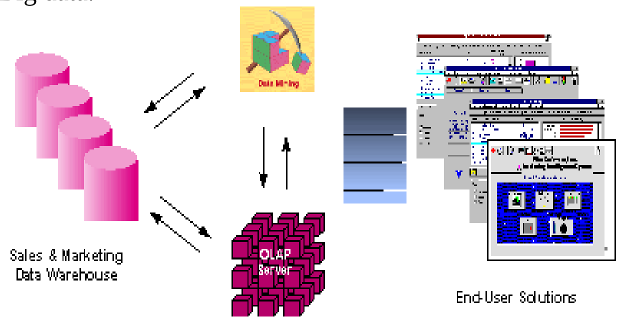


Figure 1.3: Integrated Data Mining Architecture

The ideal starting point is a Big data containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the Big data. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the Big data and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the Big data enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions. This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

A) ASSOCIATION MINING

Association mining was introduced by Agrawal et al.[1], it has emerged as a prominent research area. The association mining problem also referred to as the *market basket* problem can be formally defined as follows. Let $I = \{i_1, i_2, . . . , i_n\}$ be a set of items as $S = \{s_1, s_2, . . ., s_m\}$ be a set of transactions, where each transaction $s_i \in S$ is a set of items that is $s_i \subseteq I$. An *association rule* denoted by $X \Rightarrow Y$, where $X, Y \subset I$ and $X \cap Y = \Phi$, describes the existence of a relationship between the two itemsets X and Y. Several measures have been introduced to define the *strength* of the relationship between itemsets X and Y such as support, confidence, and interest. The definitions of these measures, from a probabilistic model are given below.

$Support(X \Rightarrow Y) = P(X,Y)$, or the percentage of transactions in the database that contain both X and Y.

$Confidence(X \Rightarrow Y) = P(X,Y)/P(X)$, or the percentage of transactions containing Y in transactions those contain X.

$Interest(X \Rightarrow Y) = P(X,Y)/P(X)P(Y)$ represents a test of statistical independence.

1.3 DATA MINING PROCESS

The data mining process works as follows in this architecture. First, the user defines the parameters for data mining in the graphical user interface. The data mining services on the client perform some pre processing prior to calling the data mining services on the middle tier. The first task on the middle tier is authentication and authorization of the users. Then the data mining services queue and execute the tasks of several clients and send back the results. These are used in the post-processing of the client, which computes the final outcome and presents it to the user. A client may start several data mining tasks in one session. Each of them includes a number of calls to the middle tier. Data mining services use the data access services on the middle tier in order to read from different types of data sources. This three-tier approach has several advantages compared to the two-tier architecture.

First, the data mining services can control the number of connections to the warehouse as well as the number of statements currently executed by the database system. The middle tier can control the number and kind of data mining tasks that are processed in parallel. This enables the system to influence the usage of system resources for data mining purposes, especially bandwidth and CPU cycles.

Second, the system can service users according to their priority and membership in user groups. This includes restricted access to data mining tables as well as user specific response behavior.

Third, a wide range of optimization strategies can be realized. The tasks of the data mining services can be distributed over the client and the middle tier. The middle tier can exploit parallelism by parallel processing on the middle tier hardware and parallel connections to the database layer. Additionally, the data mining services can reuse the outcome of data mining sessions and pre compute common intermediate results.

1.4 TYPES OF DATA MINING

Different types of Data Mining are given as

1. Predictive Data Mining
2. Descriptive Data Mining

Predictive data mining involves creation of model system based on and described by a given set of data. Descriptive data mining on the other hand produces new and unique information inferred from the available set of data.

A) RAW DATA:

Raw data is a term for data collected on source which has not been subjected to processing or any other manipulation. (Primary data), it is also known as primary data. It is a relative term (see data). Raw data can be input to a computer program or used in manual analysis procedures such as gathering statistics from a survey. It can refer to the binary data on electronic storage devices such as hard disk drives (also referred to as low-level data).

B) NORMALIZATION OF RAW DATA

Some data-mining methods, typically those that are based on distance computation between points in an n-dimensional space, may need normalized data for best results. Here are three simple and effective normalization techniques:

Suppose that the data for a feature v are in a range between 150 and 250. Then, the previous method of normalization will give all normalized data between .15 and .25; but it will accumulate the values on a small subinterval of the entire range. To obtain better distribution of values on a whole, normalized interval, e.g., [0, 1], we can use the min-max formula.

2. PROBLEM DEFINITION

Big Data actually concerns with a bulk of data presented in large volume with complex architecture and with increasing dataset. The data in such system can be taken from multiple sources and sometimes from autonomous sources. With the development of new centralized system, cloud environments, the use of Big Data is extending in all fields of business and technology. The Big Data is directly available to the end users so the criticality there exists in terms of fast retrieval of data from the system. There are number of such challenges for Big data processing.

To perform the analytical information retrieval from such data system there is the requirement of data driven model. In this present work, one of such middle layer model is been presented to derive the valuable and predictive information from big data. The presented model will store the all aspects of dataset in the form of meta data so that the user query will be performed on the selective dataset instead of whole dataset. This meta data will be generated based on the data mining assessment over the dataset and store the concrete information such as value count in each field, associatively between two or more attributes. As the user query will be performed, the analysis of the query will be performed on this meta data information and identify the most relative attribute set and the data section that can answer the query in effective way. Now the query will be performed on this partial dataset instead of whole dataset. The presented work is divided in 4 stages. In first stage, the big data assessment will be done, in second stage, the relation between the meta data and user query will be established. In third stage, the dynamic sectioning of data will be performed and finally the analytical information will be retrieved from the dataset based on user query. The presented work will be implemented on aviation dataset. The work will use the java as the front end and will use the oracle or mysql as the backend.

2.1 IMPROVEMENT OVER EXISTING WORK

The presented work is defined to identify the analytical answers on the user query performed on the aviation dataset. The presented work will be improved in following direction. In this work, middle layer architecture will be defined to perform the query analysis and the assessment of the big data. In the base paper, no such architecture is defined. The existing work based on the dynamic correlation analysis on data values whereas in this proposed work most of the mining assessment is defined statically so that efficient process will be done and the query assessment will done dynamically.

In this proposed work, the work will be implemented on dynamic generated data section instead of whole database whereas no such approach is defined in existing base paper work.

2.2 SIGNIFICANCE OF WORK

Big Data is an active research area is likely to see increased research activity in the near future as warehouses and data marts proliferate. And academic research into data warehousing technologies will likely focus on automating aspects of the warehouse. The presented big data will be taken from the aviation industry so that it is easy to answer the question of aviation system query. The proposed work is beneficial in terms of

The work will be able to answer the user query. As the work is based on one time meta data generation so that the dynamic analysis time will be reduced so that the efficiency of the system will be improved.

The managerial decisions are easy to take. It is quite easy to perform data mining and other database queries on this filtered data. The accuracy and efficiency both will be enhanced.

2.3 OBJECTIVES

In this proposed work following objectives are required to achieve:

- The first objective will be collect the aviation dataset from the secondary sources
- The main objective of work is to define middle layer architecture to perform the hybrid processing on big data.
- The objective of work is to perform the data mining assessment to define the meta data relative to big data.
- The objective of work is to perform decision based on query analysis and meta data information.
- The main objective of work is retrieving the results from the big data accurately and efficiently.

2.4 PROPOSED WORK

As we Defined the complete process of Big data Cleaning and classification is divided in the following Steps:

GENERATE DATASET:

The first step is to define a valid dataset with large amount of data. In this present work we are going to define an RFID database.

Every subset of a frequent item set is also frequent. Algorithms make use of this property in the following way - we need not find the count of an item set, if all its subsets are not frequent. So, we can first find the counts of some short item sets in one pass of the database. Then consider longer and longer item sets in subsequent passes. When we consider a long item set, we can make sure that all its subsets are frequent. This can be done because we already have the counts of all those subsets in previous passes.

Let us divide the tuples of the database into partitions, not necessarily of equal size. Then an item set can be frequent only if it is frequent in at least one partition. This property enables us to apply divide and conquer type algorithms. We can divide the database into partitions and find the frequent item sets in each partition. An item set can be frequent only if it is frequent in at least one of these partitions. To see that this is true, consider k partitions of sizes $n_1,n_2,...,n_k$.

Let minimum support be s.

The database cleaning is required to get the reliable results from the Big data. All the analytical and high level decisions are based on such kind of centralized database. As the dataset of warehouse is very large it required an optimized approach to perform the same. In this proposed work to get the accuracy and the efficiency we have combined three approaches to get the desired work of removing the impurities of database. The first dimension will identify the incomplete dataset and remove it. For this work we are using the association based mining along with fuzzy rule set. This work is here presented in the form of an example. To remove the duplicate data a clustered approach will be used. First of all the data will divide to the clusters. To perform the clustering the K means algorithms is defined. As the clustering will be performed all the data values will be defined in some clustered according to the defined values. If there is some value left that cannot be placed in any cluster, it will represent the case of inaccurate data. It means the data itself is incorrect. Such kind of data will be pruned from the dataset. The clustering process will eliminate the second kind of impurity called invalid dataset.

2.5 ASSOCIATION MINING

Association mining was introduced by Agrawal et al., it has emerged as a prominent research area. The association mining problem also referred to as the *market basket* problem can be formally defined as follows. Let $I = \{i_1,i_2, . . . , i_n\}$ be a set of items as $S = \{s_1, s_2, . . ., s_m\}$ be a set of transactions, where each transaction $s_i \in S$ is a set of items that is $s_i \subseteq I$. An *association rule* denoted by $X \Rightarrow Y$, where $X,Y \subset I$ and $X \cap Y = \Phi$, describes the existence of a relationship between the two item sets X and Y. Several measures have been introduced to define the *strength* of the relationship between item sets X and Y such as support, confidence, and interest. The definitions of these measures, from a probabilistic model are given below.

$Support(X \Rightarrow Y) = P(X,Y)$ or the percentage of transactions in the database that contain both X and Y.

$Confidence(X \Rightarrow Y) = P(X,Y)/P(X)$, or the percentage of transactions containing Y in transactions those contain X.

$Interest(X \Rightarrow Y) = P(X,Y)/P(X)P(Y)$ representsa test of statistical independence.

2.6 FUZZY LOGIC

Classically, a set is defined by its members. An object may be either a member or a non-member: the characteristic of the crisp set. The connected logical proposition may also be true or false. This concept of crisp set may be extended to fuzzy with the introduction of the idea of partial truth. Any object may be a member of a set 'to some degree'. Fuzzy set theory offers a precise mathematical form to describe such fuzzy terms in the form of fuzzy sets of a linguistic variable. To represent the shades of meaning of such linguistic terms, the concept of grades of membership or the concept of possibility values of membership has been introduced. M(x) represents the membership of some object in the set X. membership of an object will vary from full membership to non-membership:

1. for no membership
2. for full membership
3. for partial membership

Any fuzzy term may be described by a continuous mathematical function or discretely by a set of pairs of values {numeric values of linguistic variable and corresponding grades of membership}.

2.7 ALGORITHM

Usefulness of a rule can be measured with a minimum support threshold.This parameter lets to measure how many events have such itemsets that match both sides of the implication in the association rule. Rules for events whose itemsets do not match boths sides sufficiently often (defined by a threshold value) can be excluded. Database D consists of events T_1, T_2,… T_m, that is $D = \{T_1, T_2,…, T_m\}$. Let there be an itemset X that is a subregion of event T_k, that is $X \subseteq T_k$. The support can be defined as

$$\mathbf{sup(X)} = \frac{|\{T_k \in D \mid X \subseteq T_k\}|}{|D|}$$

This relation compares number of events containing itemset X to number of all events in database.

2.8 CODE MODULES

The description of different code modules associated with presented work is listed below:

Code File	Description
GUI.java	It contains the basic GUI to accept the database file and integrate all mining work. It includes the basic cleaning and horizontal and vertical pruning at earlier stage.
Main Test All Association Rules.java	This is the main file that will identify the association rules and perform the rule pruning. Currently this file has divided the dataset in partitions, horizontally and vertically. The file also defined the context attribute specification and perform the database analysis
DB.java	This is the library file to work on database such as connection establish, data retrieval etc.
AssociationRule.java	This file contains the basic working on association rule on single item or record. It includes the support and confidence based association rule generation.
AssociationRuleCollection.java	This file will combine all the association rules generated on all the dataset records
Bit.java	This file contains the bit wise operations on dataset records or values
ContextApriori.java	This file defines the context attribute or the main key attribute of the dataset based on which the effective rules will be generated.
ItemApriori.java	This file defines the single item under the apriority rule specification
ItemsetApriori.java	This file contains the apriority rule specification for complete dataset
Itemset.java	This file defines the dataset operations such as read operation, dataset retrieval
ItemsetCollection.java	Defines the complete dataset operations
Rule.java	Define the rule specification

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

2.9 RESEARCH DESIGN

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. The presented 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. To specify these rules, the context attribute specification based analysis is performed along with a priori algorithm. The basic flow of presented work is shown here under

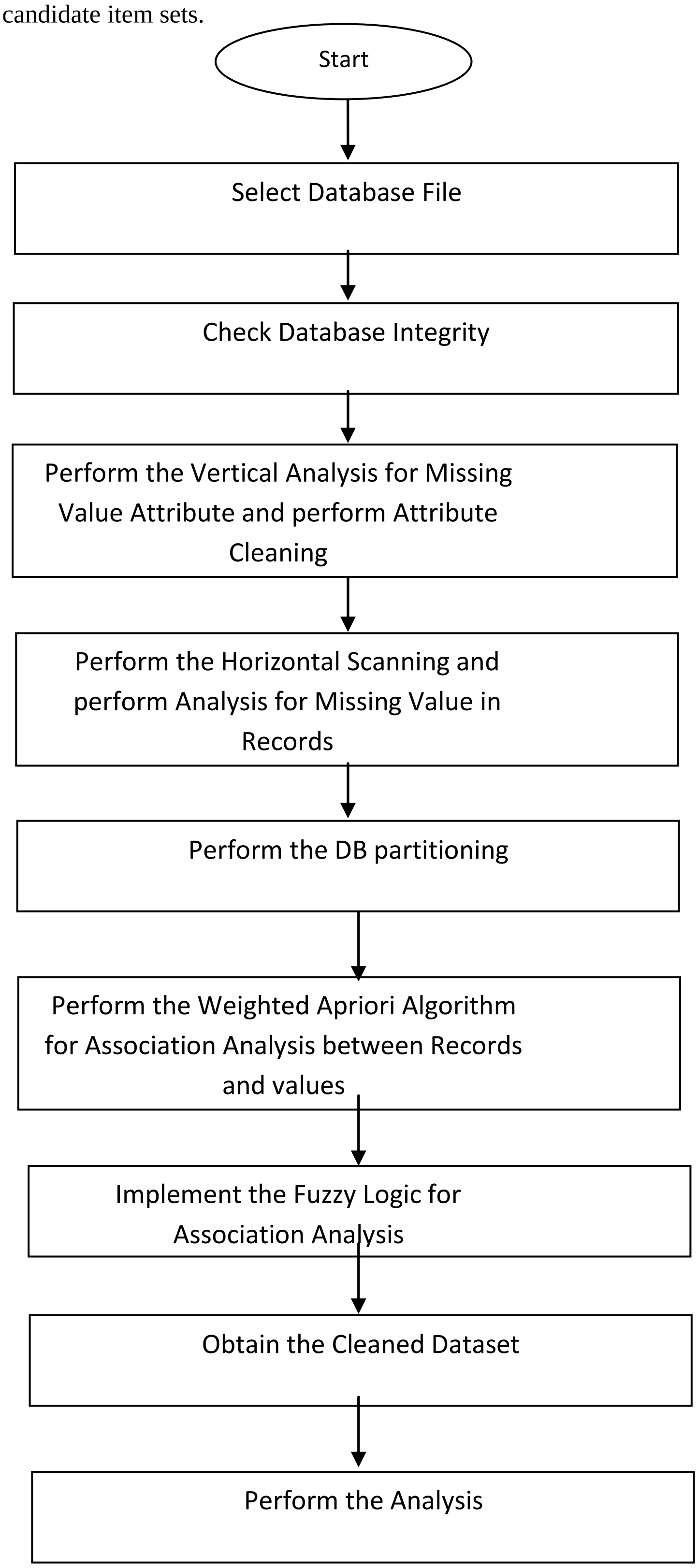
The presented work is about the generation of effective association rules on RFID dataset. The work includes the support counts of candidate item sets after every pass. The association rule first computes support counts of 1-itemsets from each site in the same manner as it does for the sequential Apriority. It then broadcasts those item sets to other sites and discovers the global frequent 1-itemsets. Subsequently, each site generates candidate 2-item sets and computes their support counts.

At the same time, association mining also eliminates all globally infrequent 1-itemsets from every transaction and inserts the new transaction (that is, a transaction without infrequent 1-itemset) into memory. While inserting the new transaction, it checks whether that transaction is already in the memory. If it is fuels increases that transaction's counter by one. Otherwise, it inserts the transaction with a count equal to one into the main memory. After generating support counts of candidate 2-itemsets at each site, mining generates the globally frequent 2-itemsets. It then iterates through the main memory (transactions without infrequent 1-itemsets) and generates the support counts of candidate item sets of respective length. Next, it generates the globally frequent item sets of that respective

Because mining eliminates all globally infrequent 1-itemsets from every transaction and inserts them into the main memory, it reduces the transaction size (the number of items) and finds more identical transactions. This is because the data set initially contains both frequent and infrequent items. However, total transactions could exceed the main memory limit. To deal with this problem, we propose a technique that fragments the data set into different horizontal partitions. Then, from each partition, mining removes infrequent items and inserts each transaction into the main memory. While inserting the transactions, it checks whether they are already in memory. If yes, it increases that transaction's counter by one. Otherwise, it inserts that transaction into the main memory with a count equal to one. Finally, it writes all main-memory entries for this partition into a temp file each local site generates support counts and broadcasts them to all other sites to let each site calculate globally frequent item sets for that pass.3,12 So, the total number of messages broadcast from each site equals (*n* - 1 * |*C*|). We can calculate the total message size using [64]

$$T_{messages} = \sum_{i-1}^{n} (n-1) * C,$$

where *n* is the total number of sites and *C* is number of candidate item sets.


Start
Select Database File
Check Database Integrity
Perform the Vertical Analysis for Missing Value Attribute and perform Attribute Cleaning
Perform the Horizontal Scanning and perform Analysis for Missing Value in Records
Perform the DB partitioning
Perform the Weighted Apriori Algorithm for Association Analysis between Records and values
Implement the Fuzzy Logic for Association Analysis
Obtain the Cleaned Dataset
Perform the Analysis

3. LANGUAGE AND TOOLS

Java was designed to meet all the real world requirements with its key features, which are explained in the following paragraph. Java was designed to be easy for the professional programmer to learn and use efficiently. Java makes itself simple by not having surprising features. Since it exposes the inner working of a machine, the programmer can perform his desired actions without fear. Unlike other programming systems that provide dozens of complicated ways to perform a simple task, Java provides a small number of clear ways to achieve a given task.

3.1 JAVA SWING

Swing components facilitate efficient graphical user interface (GUI) development. These components are a collection of lightweight visual components. Swing components contain a replacement for the heavyweight AWT components as well as complex user interface components such as Trees and Tables.

Swing components contain a pluggable look and feel (PL & F). This allows all applications to run with the native look and feel on different platforms. PL & F allows applications to have the same behaviour on various platforms. JFC contains operating system neutral look and feel. Swing components do not contain peers. Swing components allow mixing AWT heavyweight and Swing lightweight components in an application.

3.2 STRUCTURED QUERY LANGUAGE

SQL (Pronounced Sequel) is the programming language that defines and manipulates the database. SQL databases are relational databases; this means simply the data is store in a set of simple relations. A database can have one or more table. You can define and manipulate data in a table with SQL commands. You use the data definition language (DDL) commands to creating and altering databases and tables.

You can update, delete or retrieve data in a table with data manipulation commands (DML). DML commands include commands to alter and fetch data. The most common SQL commands include commands is the SELECT command, which allows you to retrieve data from the database.

3.3 GUI CREATION

The different GUI components used in IDS monitor are JButton, JLabel, JTextField, JTextArea, JScrollPane, DefaultTableModel and Container. JButton is used. Swing components facilitate efficient graphical user interface (GUI) development. These components are a collection of lightweight visual components. Swing components contain a replacement for the heavyweight AWT components as well as complex user interface components such as Trees and Tables.

Swing components contain a pluggable look and feel (PL & F). This allows all applications to run with the native look and feel on different platforms. PL & F allows applications to have the same behaviour on various platforms. JFC contains operating system neutral look and feel. Swing components do not contain peers. Swing components allow mixing AWT heavyweight and Swing lightweight components in an application.

TABLE 3.3: Application Components

CLASS	DESCRIPTION
JButton	Push Button implementation
Jlabel	It displays the area for a short text string. A label does not react to input events. As a result, it cannot get the keyboard focus.
JTextField	It is a lightweight component that allows the editing of a single line of text.
JTextArea	It is a multi-line area that displays plain text.
JScrollPane	Provides a scrollable view of a lightweight component.
Container	Components added to a container are tracked in a list. The order of the list will define the components' front-to-back stacking order within the container. If no index is specified when adding a component to a container, it will be added to the end of the list.

Figure 3.3 : Graphical Interface

3.4 REFINED PROPOSED WORK

The database cleaning is required to get the reliable results from the Big data. All the analytical and high level decisions are based on such kind of centralized database. As the dataset of warehouse is very large it required an optimized approach to perform the same. In this proposed work to get the accuracy and the efficiency we have combined three approaches to get the desired work of removing the impurities of database. The first dimension will identify the incomplete dataset and remove it. For this work we are using the association based mining along with fuzzy rule set. This work is here presented in the form of an example.

Table 3.4: Sample DataSet

A	B	C	D	E	F	G
1	1	0	0	0	1	1
1	1	1	1	1	1	0
1	1	0	0	1	1	1
1	1	0	0	0	0	0
1	0	1	0	0	0	0
1	1	1	0	0	0	0
1	0	0	0	0	0	0

In table 3.4 a sample dataset is shown with 7 attributes and 7 records. The value 1 here represents the presence of data and 0 represents the absence of data. From this dataset we drive a table of existing dataset.

Table 3.5: Association of Data

{A}	7
{A} & {B}	5
{A} & {C}	3
{A} & {D}	1
{B} & {D}	1
{C} & {D}	1

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

In table 3.5 some association are shown here with the occurrence of values in different attributes. In same way all the possible association will be taken. On the analysis of this association a lower threshold value will be defined such as 2. Now all the association having support less then this defined value will be eliminated. Now the attribute with minimum support will be eliminated from the dataset. In such way the incomplete information fields and the non required fields and record set will be eliminated and incomplete data impurities will be eliminated.

To remove the duplicate data a cluster approach will be used. First of all the data will divide to the clusters. To perform the clustering the K means algorithms is defined. As the clustering will be performed all the data values will be defined in some clustered according to the defined values. If there is some value left that cannot be placed in any cluster, it will represent the case of inaccurate data. It means the data itself is incorrect. Such kind of data will be pruned from the dataset. The clustering process will eliminate the second kind of impurity called invalid dataset.

4. CONCLUSION

The reliability of data is because of its accuracy. Big data contains bulk of data. It includes the data taken from different data centers. Because of this Big data can contain some data impurities. Big data is responsible for all kind of management level decision related to data. Because of this there is the requirement to perform the cleaning on user data. To perform the knowledge acquisition and knowledge discovery we are here presenting an optimized approach to perform the data cleaning along with data association and the classification. The proposed approach is the rule based approach along effective rule generation. In this approach at first duplicate data and other impurities are cleaned from the database. The work is here defined for rule identification for RFID system.

5. FUTURE WORK

In this present work, the work is performed on data cleaning on a centralized dataset. The work can be improved by taking the concept of some other database systems such as Distributed Database or the Mobile Database System. More work can be done in direction of Efficiency.

REFERENCES

[1] Edmon Begoli, 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] Kapil Bakshi, “Considerations for Big Data: Architecture and Approach”, IEEE, 2012.

[3] Lawrence 0. Hall, Nitesh Chawla , Kevin W. Bowyer, “Decision Tree Learning on Very Large Data Sets”, IEEE, Oct 1998.

[4] Edmon Begoli," Design Principles for Effective Knowledge Discovery from Big Data", 2012 Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture 978-0-7695-4827-2/12 © 2012 IEEE.

[5] Aditya B. Patel," Addressing Big Data Problem Using Hadoop and Map Reduce", 2012 NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING

[6] Dan Garlasu," A Big Data implementation based on Grid Computing".

[7] Marcus R. Wigan," Big Data’s Big Unintended Consequences", 0018-9162/13 © 2013 IEEE.

[8] Xin Luna Dong," Big Data Integration", ICDE Conference 2013 978-1-4673-4910- 9/13@ 2013 IEEE.

[9] Xindong Wu,"Data Mining wih Big Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 1041-4347/13 © 2013 IEEE.

[10] Seref SAGIROGLU," Big Data: A Review", 978-1-4673-6404-1/13 ©2013 IEEE.

[11] Yuri Demchenko," Addressing Big Data Issues in Scientific Data Infrastructure", 978-1-4673-6404-1/13 ©2013 IEEE.

[12] Antonia Azzini," Consistent Process Mining Over Big Data Triple Stores", 2013 IEEE International Congress on Big Data 978-0-7695-5006-0/13 © 2013 IEEE.

[13] Zibin Zheng," Service-generated Big Data and Big Data-as-a-Service: An Overview", 2013 IEEE International Congress on Big Data 978-0-7695-5006-0/13 © 2013 IEEE.

[14] P. Eredics* and T.P. Dobrowiecki, Data Cleaning for an Intelligent Greenhouse 6th IEEE International Symposium on Applied Computational Intelligence and Informatics, May 19–21, 2011.