An Effective Data Warehousing System for RFID Using Novel Data Cleaning, Data Transformation and Loading Techniques

Barjesh Kochar[1] and Rajender Chhillar[2]
[1]Department of MCA, Guru Nanak Institute of Management, India
[2]Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India

Abstract: *Nowadays, the vital parts of the business programs are the data warehouses and the data mining techniques. Especially these are vital in the Radio Frequency Identification (RFID) application which brings a revolution in business programs. Manufacturing, the logistics distribution and various stages of supply chains, retail store and quality management applications are involved in the RFID technology in business. A large volume of temporal and spatial data is generated by the ubiquitous computing and sensor networks of RFID and these are often generated with noises and duplicates. The noises and duplicates in the RFID data declare the need of an effective data warehousing system. The warehousing system has the responsibility to provide proper data cleaning technique to clean the dirty data which occurs in the applications. Also, the cleaned data has to be transformed and to be loaded properly so that they can be stored in the database with minimum space requirements. In this paper, we propose a novel data cleaning, transformation and loading technique which makes the data warehousing system employed for any RFID applications more effective. The chosen RFID application is tracking of goods in warehouses using RFID tags and readers, one of the significant RFID applications. The data cleaning is performed based on the probability of each RFID tag's response and the window size which is made adaptive. The window size changes on the basis of the occurrence of the dirty data and hence the cleaning is more effective. The purified data is transformed in a special structure in such a way that the ware house can have only the tag IDs which are under transaction and the time of interrogation in the size of bits. The transformed data are loaded into the warehouse using the proposed loading technique in a dedicated tabular format.*

Keywords: *Data warehousing system, data cleaning, data transformation, data loading, dirty data, RFID, suspicious tags.*

Received September 29, 2009; accepted March 9, 2010

1. Introduction

In recent years, data warehousing and data mining became an important part of many organization's IT infrastructure [17]. Data mining is a recently rising field, linking the three worlds of databases, artificial intelligence and statistics. The information age has enabled many organizations to collect large volumes of data. However, the utility of this data is negligible if "meaningful information" or "knowledge" cannot be extracted from it [13]. Data mining focuses on applying some specific machine learning algorithms that can discover previously unknown regularities and trends in databases [21]. Data mining or knowledge discovery in databases is the automatic extraction of implicit and interesting patterns from large data collections [18]. If expressed alternatively, data mining is an inductive, iterative process that extracts information or knowledge patterns from volumes of data [20]. But data warehousing intends to store clean and reliable data obtained from various systems employed in an organization or an application, in an integrated format. The data thus structured can particularly serve for reporting and analytic requisites. A data warehouse can be defined as the repository of an organization's data that are stored electronically. In general, it is an operational database which supports the processing of day-to-day transactions. Some of the essential components of a data warehousing system are: Extracting, cleaning, transforming and loading of data.

There is a symbiotic relationship between the activity of data mining and the data warehouse-the architectural foundation of decision support systems. The data warehouse sets the stage for effective data mining. Data mining can be done where there is no data warehouse, but the data warehouse significantly improves the chances of success in data mining [8].

Data now can be stored in different types of databases, the data warehouse is one of such database architecture, a storehouse of multiple heterogeneous data sources, organized under a combined scheme at a single site in order to support the management decision-making [25]. Data warehouses have established themselves in the information flow architectures of business organizations for two main reasons: firstly, as a buffer between operational and transactional tasks on the one hand, and analytical

strategic tasks on the other. Secondly, to capture the history of business transactions for purposes of archiving, traceability, experience mining and reuse [9].

A data warehouse contains purified and organized data that allow decision makers to make business decisions based on facts, not on perception; it includes a repository of information which is built using data from the distant and often departmentally isolated, systems of enterprise – wide computing [19]. To make a data warehouse, we have to follow a process known as Extraction Transformation and Loading (ETL) which involves [4] extracting data from various outside sources, transforming it to fit business needs, and ultimately loading it into the data warehouse.

In the extraction phase, operational data are moved into the Enterprise Data Warehouse (EDW). Transformation phase changes the structure of data storage. Loading process represents an iterative process. The data warehouse has to be populated repeatedly and incrementally to reflect the changes in the operational system(s) [19]. Data cleaning is another vital element of the data warehousing system. In data warehouses, data cleaning is usually an off-line, centralized, iterative, and sometimes interactive process which concentrates on a small set of well defined tasks [5]. Data cleaning process is used to ensure the reliability of the data warehouse data (i.e., the fact that these data respect the database constraints and the business rules) [22].

The data warehouse is a large system, which classifies the data so that the most possible information can be extracted from it. It organizes and makes the data available for analysts so that they can make better decisions. It will allow for comprehensive queries of the information sorted [6]. The extreme development of data warehousing is offering organizations with a potential decision support utility that can be successfully utilized to maintain supply chain activities throughout a business or industry. Data warehouses serve as the basis of a company's successful supply chain management. In such supply chain management, the role played by Radio Frequency Identification (RFID) applications in object tracking and supply chain management is essential [3].

RFID has emerged in order to replace a barcode which is used for the object identification till now. Unlike the barcode system, RFID has many different advantages: it can have memory in addition to identification of data and it can be recognized out-of-sight and form a relatively long distance [11]. RFID applications are set to play a vital role in object tracking and supply chain management systems. Imminent, it is expected that every major retailer will use RFID systems to track the movement of products from suppliers to warehouses, store backrooms and ultimately for the points of sale [7]. RFID has been gradually adopted and utilized in a wide area of applications, such as aircraft maintenance, baggage handling, laboratory procedures, security and healthcare. True benefits of RFID technology can be realized only when the tracking information from RFID components is efficiently included into business applications [16]. In the RFID applications, to get the information from the tags, initially, the readers interrogate nearby tags by sending an RF signal. Tags in the area respond to these signals with their unique identifier code. An interrogation cycle is iteration through the reader's protocol that attempts to establish all tags in the reader's area [10]. The results of multiple reader interrogation cycles are typically grouped into what we term epochs [1] and a number of epoch forms a window. The received information from all the tags are then stored in the data warehouse. An efficient data ware housing system is required in this situation of any RFID applications.

As the tags are dynamic in most of the applications, a single reader cannot get the response from a certain tag throughout all the periods. Meanwhile, the response of any tag may be interrupted by any objects, humans or any other interventions. So, the reader may missed the information from the tag by deciding that the tag is absent in its range, results in the storage of dirty data in the warehouse. Hence, cleaning the dirty data can be analyzed in three cases. In the first case, the intrusion may occur at the beginning of the window. In the second case, the intrusion may occur at the middle of the window and the third arises when the intrusion occurs at the end of the window. Though, the reader can't get the response from the tags at the beginning and the middle of the window in the first and second cases respectively, the reader can get the same at the end cycles. As the readers come to a decision only at the end of the window, the possibility for dirty data is very low in the first and second cases. But in the third case, the possibility of getting the dirty data is very high as there are no responses from the tags at the end of the window. The data warehousing system has the responsibility to provide proper data cleaning technique to clean the third case dirty data occurs in the applications. Also, the cleaned data has to be transformed and to be loaded properly so that they can be stored in the database with minimum space requirements.

In this paper, we are proposing a novel data cleaning, transformation and loading technique which makes the data warehousing system employed for any RFID applications more effective. The rest of the paper is organized as follows. Section 2 deals with some of the recent research works related to the paper and section 3 details the proposed data cleaning, data transformation and data loading techniques with necessary mathematical formulations. Section 4 discusses the implementation results and section 5 concludes the paper.

2. Related Works

Lee *et al*. [12] have presented a framework for a service provisioning middleware system that can discover primitive services and compose dynamic complex services according to the context information. They have also described an algorithm of service composition which uses the service history information and an ontology engine with data mining. Finally, they have shown that their experiments enhance the possibility of provisioning services considering user's preference and thus provide users with newly composed optimal services.

Ling *et al*. [14] have discovered the possibility of generating user profiles of fashion preferences from information captured by RFID technology. Proposing a design of a smart wardrobe, they have investigated the suitability of the technology as an identification tool of real objects and to support in detecting and tracking real objects movements. Then they have presented a model of user fashion profile, which was generated through queries and data mining techniques. In order to illustrate the usefulness and real world feasibility of their model, they have constructed a working prototype as a proof of concept. For evaluation purposes, they have created a random generator, which was able to generate random clothing items and dressing events that serve as input to their model for creating user profiles. Their experimental result clearly indicates that RFID technology was suitable to assist in creating smart systems.

Velpula and Gudipudi [23] have focused primarily on the techniques for detecting insider attacks. They have also discussed the processes required to implement a solution. In particular, they have described a behavior-anomaly based system for detecting insider attacks. Their system uses peer-group profiling, composite feature modeling, and real-time statistical data mining. The analytical models are sophisticated and used to update the real-time monitoring process. Finally, they have described an implementation of this detection approach in the form of the IBM Identity Risk and Investigation Solution (IRIS).

Liu *et al*. [15] have given certain insight analysis of RFID system from the view of application perspective. First they have introduced some essential implementation principles of RFID system and also have shown its hot application areas. Then, they have analyzed the limiting factors of RFID system in advanced development, and presented a generalized modeling process from the view of knowledge-embedded data mining. Further, based on the above though of knowledge-embedded data mining, a new forecasting method, KERGM (1, 1), was developed to forecast RFID's potential application prospect. Lastly, several countermeasures and advices for the development of RFID industry were put forward. It was shown that the future RFID industry should be greatly promising, however, some countermeasures should be adopted to support its development, such as reducing the cost of RFID tag with technical innovation, popularizing RFID system in wider fields, and strengthening international cooperation in RFID standard system.

Wadhwa and Lin [24] have reviewed the RFID technology and the components which forms the backbone of the RFID system. Next, they have demonstrated the usefulness of RFID in supply chain and presented some data mining challenges in RFID. Finally a real-life case study was used to demonstrate how organizations are using RFID data.

Bottani [2] has presented a discrete event simulation model reproducing the adoption of RFID technology for the optimal management of common logistics processes of a Fast Moving Consumer Goods (FMCG) warehouse. In their study, simulation was emerges as a powerful tool to replicate the both reengineered RFID logistics processes and the flows of Electronic Product Code (EPC) data generated by such processes. Moreover, a complex tool was developed to examine data resulting from the simulation runs, thus addressing the issue of how the flows of EPC data generated by RFID technology can be exploited to provide value-added information for optimally managing the logistics processes. Specifically, an EPCIS-compliant data warehouse was also designed to act as EPCIS Repository and store EPC data resulting from simulation. Results of his study can provide a proof-of-concept to validate the adoption of RFID technology in the FMCG industry.

Chhillar and Kochar [3] have discussed a novel efficient approach, called as Efficient RFID Data Warehouse Management (ERDWM), for warehousing the RFID data effectively. Their approach has considered the loads of the RFID readers in order to identify overloaded readers and the one without load. Based on the priority, the latter are placed in the area of the former for the effective collection and warehousing of data without any loss.

3. Proposed Data Cleaning, Transformation and Loading Techniques for Data Warehousing

For an effective data warehousing system deployed in any RFID applications, the data extracted from the RFID reader is essentially to be exact, the data to be stored in the repository to be transformed efficiently and to load the data in the right format. As discussed earlier, in any RFID applications, because of any noisy signal or any interruptions the reader might not get the response from the tag under the range of the reader. So, it is necessary to make the system more robust against these kinds of interruptions which are restricting the system to get the tag response. Also, the proficient

transformation of the obtained data and loading of the data into the repository is required for storing the data with minimum space requirement and effective data retrieval respectively, significant pre-requisites of the data warehousing. In order to accomplish all the requirements so as to make the data warehousing system of any RFID employed applications, we are proposing a novel data cleaning, data transformation and data loading techniques for more robust data warehousing system for any RFID applications. The proposed techniques are detailed in the further sub-sections.

3.1. Data Cleaning

Data cleaning is a process which is used in the databases to identify the incomplete, incorrect, inaccurate, irrelevant parts of the data and then it is used for replacing, modifying or deleting this dirty data. Let m be the total number of readers and n be the total number of tags used in the application and the vector representation can be given as: $R_{1\times m}=[R_0 \;\; R_1 \;\; R_2 \cdots R_{m-1}]$ and $T_{1\times n}=[T_0 \;\; T_1 \;\; T_2 \cdots T_{n-1}]$. When each reader broadcasts the RF signal, the tags which are in the range of the corresponding readers responds. In the proposed data cleaning technique, a window of size w_{s_i} which is constituted by a certain number of epochs is utilized for each reader. The w_{s_i} of any reader can be determined as:

$$w_{s\,i}=\sum_{a=0}^{E_i-1} e_a \tag{1}$$

Where, e_a represents a^{th} epoch time in seconds and E_i is total number of epochs per window. The w_s of each reader is made adaptive and changes based on the dirty data obtained in the previous window and thus the data cleaning is accomplished. The data obtained at each epoch of every window takes the binary values based on the probability of response as follows:

$$d_{ij}(k)=\begin{cases}1 & ;if \;\; P_{ij}(k)>0\\ 0 & ;else\end{cases} \tag{2}$$

Where, P_{ij} is the probability of the response from each tags which can be determined as the number of interrogation cycles, the reader receives response from the tag by the total number of interrogation cycles. In the meantime, the reader notifies the central server about the information of newly entered tags at every interrogation cycle except the first interrogation cycle of the process, the core process in the data cleaning. For this purpose, initially, the central sever creates a null set and the reader sends the T_j as the set element to the server. The reader recognizes the newly entered tags as:

$$S_{n_{xyz}}^{(i)}=S_{T_{xyz}}^{(i)}-S_{R_{xy(z-1)}}^{(i)} \tag{3}$$

In equation 3, $S_{R_{xy(z-1)}}^{(i)}$ is a set of tag IDs covered by the i^{th} reader till $(z\text{-}1)^{th}$ interrogation cycle, $S_{T_{xyz}}^{(i)}$ is a set of tag IDs covered by the i^{th} reader at the z^{th} interrogation, $S_{n_{xyz}}^{(i)}$ is a set of newly entered tags in the range of i^{th} reader, where, $0\le x\le n_W-1$, $0\le y\le E_x-1$, $0<z<I$, n_W is the total number of windows, E_x is the number of epochs in the x^{th} window and I is the number of interrogation cycles which remains constant throughout the process. The determined $S_{n_{xyz}}^{(i)}$ is sent to the server and the set elements are appended in the set S_{NT} and in the set $S_{R_{xyz}}^{(i)}$ as follows:

$$\begin{aligned}&\text{a.}\quad \{S_{NT}\}<<S_{n_{xyz}}^{(i)}\\ &\text{b.}\quad \left\{S_{R_{xyz}}^{(i)}\right\}<<S_{n_{xyz}}^{(i)}\end{aligned} \tag{4}$$

In equation 4, the elements of $S_{n_{xyz}}^{(i)}$ is appended with the elements of the set $\{S_{NT}\}$ and with the elements of $\left\{S_{R_{xyz}}^{(i)}\right\}$ respectively. The data cleaning is performed with the aid of the set $\{S_{NT}\}$. When a window of interrogation completes, each reader has the data as mentioned in the equation which directly represents either the presence or absence of the tags by ones or zeros respectively. The data cleaning is initiated by extracting the tags which does not responds at the end epochs i.e. tags which has zeros at the end epochs of the window. The extracted tags are sorted downwards in such a way that the tags which has the maximum number of zeros occupies the top of the newly obtained data vector $d_{ih}^{(r)}$, where, $r\in(0,n-1)$, $0\le h\le X_T^{(i)}$ and $X_T^{(i)}$ represents the total number of suspicious tags. Then a querying process is performed by each readers of the application.

Querying: From the $X_T^{(i)}$ tags, $Q_T^{(i)}$ number of suspicious tags is queried into the server which can be determined as $Q_T^{(i)}=X_T^{(i)}/2$. The server checks the presence of the tags mentioned in the vector $d_{iq}^{(r)}, 0\le q\le Q_T^{(i)}$, in the set $\{S_{NT}\}$ by sending a set of all the r (Tag IDs) values, say $\{r\}$, in $d_{iq}^{(r)}$. If $r_a\in\{S_{NT}\}$, $0\le a\le|Q_T|$, then the server clears the elements in the $\{S_{NT}\}$ as well as the r_a. After the

clearance of the required elements of $\{s_{NT}\}$ and $\{Q_T^{(i)}\}$, the server returns the set $\{r_s\}$ or informs by a HQ signal (i.e., Halt the querying process) if $\{S_{NT}\} \neq \phi$ or $\{S_{NT}\} = \phi$ respectively. The reader cleans the data in the vector d_{ij} based on the response of the server as follows:

- If the server returns *HQ* signal to the reader, the data of the queried tags $d_{iq}^{(r)}$ is correct and the remaining tag information is incorrect. So, the reader assures the absence of the tags in $\{r\}$ and the presence of the remaining $X_T^{(i)} - Q_T^{(i)}$ tags in the reader's range.
- If the server returns the $\{r_s\}$, the reader has to check the following two possibilities of the $\{r_s\}$ and makes the decision based on its outcome:
  - *Case 1*: If $\{r_s\} = \phi$, the reader assures the absence of the queried tags $\{r\}$ in the reader's range $\{r\}$ and the querying process is repeated again for the remaining $X_T^{(i)} - Q_T^{(i)}$ tags as similar as done earlier.
  - *Case 2*: If $\{r_s\} \neq \phi$, the server assures the absence of $\{r\} - \{r_s\}$ and presence of $\{r_s\}$ tags in the reader's range. In the similar fashion, the querying process is repeated for the remaining tags $X_T^{(i)} - Q_T^{(i)}$ also.

Once the reader went for the second case, it has to change it's w_{s_i} of the upcoming interrogations, but only after the cleaning of the data obtained in the current window as follows:

$$w_s^{new} = \begin{cases} w_s & ; \; if \; -w_s/2 + w_{s_{th}} \leq \alpha \leq w_s/2 - w_{s_{th}} \\ 2\alpha - 1 & ; \; if \; \alpha < -w_s/2 + w_{s_{th}} \\ 2\alpha & ; \; if \; \alpha > w_s/2 - w_{s_{th}} \end{cases} \quad (5)$$

Where, w_s^{new} is the new window size, $w_{s_{th}}$ is the threshold for window size and α is the weightage of the window which can be determined as:

$$\alpha = \sum_{a \in \{C_T\}} \left[d_{ij}^{(a)} \left(\frac{w_s}{2}\right) * \left(\frac{w_s}{2} + 1\right) + \sum_{b=1}^{\frac{w_s}{2}-1} d_{ij}^{(a)} \left(\frac{w_s}{2} + b\right) * \left(\frac{w_s}{2} - b\right) \right] \quad (6)$$

$$0 \leq j \leq |C_T|$$

In equation 6, $\{C_T\}$ is a set of total number of tags whose data are cleaned and $|C_T|$ is the cardinality of the set $\{C_T\}$. With the aid of the window of size w_s^{new}, the same process is repeated. Repeating the process is done by changing the window size by equation 5 at the end of every process, the probability of the occurrence of the dirty data gets reduced in the upcoming interrogations. The window size is maximized when the probability of the occurrence of the dirty data is felt small from the current process result and it is minimized in the case of occurrence of the dirty data. Hence, the technique cleans the data by modifying the dirty data by performing the technique and also looks for effective operation of the cleaning using adaptive window. At the end of the proposed cleaning technique, all the readers obtain the cleaned data which means the tag information that are actually available in their range. After the cleaning technique, data transformation is performed for consistent representation of the data warehouse.

3.2. Data Transformation

After the cleaning process, a set of tag ids $\{T_j^{(r)}\}_i$; $0 \leq j \leq n_w^{(i)}$ are obtained which are in the particular reader's range at the time of the interrogation of every window and so $n_w^{(i)}$ sets are obtained for each reader. It is to be noted that the sets used here have the elements with possible repetitions. The information to be warehoused are the tag IDs and the time of completion of each window at which the tag IDs in all the reader's range. Hence, the proposed data transformation technique is constituted by two sections of transformations. The first section of transformation transforms the structure of tag IDs to be stored which are available in the range of each reader at the completion of every window. The second section of transformation technique transforms the structure of the time of completion of the interrogation windows of each reader. In the proposed data transformation technique to keep the tag IDs of every window, the readers generate two sets of tag IDs which are newly arrived and left as $\{T_{in_j}^{(r)}\}$ and $\{T_{out_j}^{(r)}\}$, respectively. The set elements are appended as follows:

$$\{T_{in_j}^{(r)}\} << \begin{cases} \{T_j^{(r)}\}_i & ; \; if \; j = 0 \\ \{T_j^{(r)}\}_i - \{T_{j-1}^{(r)}\}; \; else \end{cases} \quad (7)$$

$$\{T_{out_j}^{(r)}\} << \begin{cases} \phi & ; \; if \; j = 0 \\ \{T_{j-1}^{(r)}\} - \{T_j^{(r)}\} \; ; \; else \end{cases} \quad (8)$$

The generated sets by equations 7 and 8 are constituted of subsets which have the tag IDs that are newly arrived and left respectively at the particular window of interrogation. Thus, instead of keeping all the tag IDs responded at each window, only the newly arrived and left tag IDs are kept. For transforming the set of time of completion of the interrogation windows $\{t_j\}_i$, initially, a set of variation of each time of interrogation window from the time of its previous interrogation window termed as $\{t_j^{\text{var}}\}_i$ is determined as:

$$\{t_j^{\text{var}}\}_i = \begin{cases} t_{j_i} & ; if \; j = 0 \\ t_{j_i} - t_{j-1_i} \; ; else \end{cases} \quad (9)$$

The elements of the set $\left\{t_j^{\text{var}}\right\}_i$ are rounded up and then indexed based on their frequency of occurrence in the set. With the aid of the frequency of elements in the set, transformation is applied and the set of elements obtained after transformation is constituted by two subsets. One of the subset consists of time of interrogation window which exhibits maximum number of occurrences in the set $\left\{t_j^{\text{var}}\right\}_i$ in such a way that $f(t_0) \geq f(t_1) \geq \cdots \geq f(t_{n_r})$. Hence, n_r number of sets are obtained, where, n_r represents the actual number of elements determined from the set $\left\{t_j^{\text{var}}\right\}_i$ without considering the repetition of each elements. The second subset consists of the availability index indicating the occurrences of the element in the set $\left\{t_j^{\text{var}}\right\}_i$. The availability index of $\{t_{n_r}\}$ in the set $\left\{t_j^{\text{var}}\right\}_i$ can be transformed as:

$$\left\{t_a^{Tr}(b)\right\} = \begin{cases} 1 \; ; \; if \;\; t_a = \left\{S_{t_a}^{(b)}\right\}_i \; ; \; 0 \leq a \leq n_r \text{ and } 0 \leq b \leq \left|S_{t_a}\right| \\ 0 \; ; \; else \end{cases} \quad (10)$$

Where,

$$\left\{S_{t_a}\right\}_i = \begin{cases} \left\{t_j^{\text{var}}\right\}_i & ; \; if \;\; a = 0 \\ \left\{S_{t_{a-1}}\right\}_i - t_a & ; else \end{cases} \quad (11)$$

From equation 10, it can be obtained that the time at which each interrogation window of every window has been completed is transformed into a certain structure. Because of the proposed data transformation technique, instead of keeping the time of completion of each interrogation window in byte format, they are kept as bit format. Thus transformed bit sized time of interrogation information and significant Tag information is subjected for loading the repository.

3.3. Data Loading

For loading the data into the warehouse, tables are created and the transformed data are stored in the tables in the proposed format. In the RFID applications, the transformed data to be stored are the Tag IDs which are newly entered at each reader's range at every window of interrogation and the Tag IDs which are left from each reader's range at every window of interrogation. Also, the time at which the interrogation of each window is completed by all the readers is also to be stored in the warehouse in the transformed format. Hence, the warehouse is organized in such a way that the information from each reader is stored in separate files and each file consists of the time of interrogation and tag information obtained at that time. The transformed data are stored in the form of table in each file.

For tag information, a table of n_r number of records with two fields is generated. Two fields have the newly entered tags and left out tags at each time of completion of interrogation window stored at every record of the table. Similarly, for time of interrogation window, the set elements $\left\{t_j^{\text{var}}\right\}_i$, frequency of each element $f(t_{n_a})$ and the availability index $\left\{t_a^{Tr}\right\}$ which is converted to integer from its binary form are stored. Instead of storing the huge information that is in terms of bytes is reduced to bits and stored in a prescribed format and so it can be retrieved exactly as it was in the past. Hence the proposed data cleaning, data transformation and loading technique makes the data warehousing system more effective in any RFID applications.

4. Results and Discussion

The proposed data cleaning, data transformation and data loading techniques for a data warehousing system have been implemented in the working platform of JAVA (version JDK 1.6). The proposed techniques for effective warehousing system are dedicated to perform the tracking of goods in warehouses. In such warehouses, the goods exhibit mobility from one location to other. The reader present in every location which covers a certain range depends upon the RF strength. Four readers are assumed to be present in the warehouse and initially 100 tags are assumed to be available under each reader's range. As the tags changes their locations in the mentioned application frequently, the availability of tags under each reader's range also changes. As per the application, each reader of the warehouse has interrogates on window basis and obtains the tag response for every window. The data thus obtained has to be stored in the data warehouse so that the location of the required goods at that particular instant could be identified easily. The proposed techniques are responsible for effective data warehousing system which is utilized in the application in keeping the exact tag information under each reader's range and the time of completion of interrogation of window. Initially, the window size w_s have been set as 6 and the response received from the tags are tabulated in Table 1. From the response of the tags as given in Table 1, the suspicious tags are identified and cleaned. The suspicious tags thus identified, the dirty data corresponding to the suspicious tags and the final cleaned data are tabulated in Table 2.

In the next process, called data transformation, the tag IDs and the time of interrogation window which are transformed, using the proposed data transformation technique are tabulated in Tables 3 and 4, respectively.

Table 1. A sample result details the response of some of the tag's response which are under the range of R1 at every epoch.

Tag IDs	E_0	E_1	E_2	E_3	E_4	E_5
1	1	1	0	1	0	1
2	0	1	1	1	0	0
3	0	1	0	1	0	1
4	1	1	1	1	0	1
5	1	1	0	1	1	1
6	1	1	1	1	0	1
8	0	1	1	1	0	1
9	0	1	1	1	0	0
10	1	1	1	1	1	0
22	0	1	0	1	0	0

Table 2. Sample output for suspicious tags, the dirty data from the readers and the cleaned data by the proposed data cleaning technique.

Suspicious Tags	Dirty Data	Cleaned Data
[2, 9, 10, 22]	[1, 3, 4, 5, 6, 8]	[1, 3, 4, 5, 6, 8, 2, 9]
[15]	[11,12, 13, 14, 16, 17, 18, 19, 20, 7]	[11, 12, 13, 14, 16, 17, 18, 19, 20, 7, 15]
[21, 22, 23, 24, 30]	[25, 26, 27, 28, 29]	[25, 26, 27, 28, 29, 22, 23]
[33, 36, 37, 38, 39, 40]	[31, 32, 34, 35]	[31, 32, 34, 35, 36, 38, 40]

For data loading, the format which has been suggested in the proposed data loading technique is also given in Tables 3 and 4. For comparison, the data without transformation is given in Table 5.

Table 3. The tag information to be stored after transformation: newly entered tags and left out tags in the range of R1 for different time of interrogation.

$Tags_{in}$	$Tags_{out}$
[2, 19]	[3, 9, 22]
[40]	[1, 4, 8, 9, 10, 22]
[35, 2]	[3, 40]
[3, 37, 32, 4]	[19, 35]
[19, 40, 13, 35]	[8, 1, 10]
[1, 10]	[40, 37, 32, 4]
[1, 2]	[10, 35, 13]

Table 4. The transformed data of time of each interrogation window by R1 which is constituted by the distinction between each interrogation time, its frequency of occurrence and the availability index.

Distinction Between Interrogation Time	Frequency of Occurrence	Availability Index
3.1	6	[5, 5, 12]
3.11	5	[86, 32]
4.59	1	[64, 0]
5.1	1	[64, 0]
0.0	1	[128]
4.6	1	[64]
6.1	1	[32]

In Table 4, the availability index [5, 5, 12] indicates binary value of 5, 5 and 12 each represented in 8-bits. The 1s in the binary value represents the availability of the distinct time of interrogation throughout the sample periods taken for transformation, say, 100 in our test case. The size incurred for loading the results in data warehouse after transformation Tables 3 and 4 is 4 KB when compared to the results before transformation Table 5 that requires 8.57 KB. Hence it can be observed that the size incurred for loading the data in the data warehouse obtained from the proposed data transformation and loading techniques are comparatively smaller than (compression ratio is approximately 53% for the considered 10 samples) loading the cleaned data as such. When the samples are increased and the number of tags which exhibits mobility gets decreased, a good compression ratio (> 60%) can be accomplished using the using the proposed techniques. Thus the results we have obtained for the proposed data cleaning, data transformation and loading have been discussed clearly. From the results obtained, it can be claimed that the proposed techniques makes the data warehousing system more effective which leads to proficient and accurate storage of RFID data.

5. Conclusions

In order to achieve efficient data warehousing system in the RFID application to monitor the goods kept in a warehouse, we have proposed a novel data cleaning, data transformation and data loading techniques. As the cleaning has been performed with the concern of the probability of response of the tags and the window size, the technique shows good performance. The window size has been increased if the strength of the dirty data is less and decreased if it is high in the past window of interrogation. In the data transformation, the purified data have been transformed into a structure in such a way that only the tag IDs which shows mobility per interrogation of window are considered. Similarly, the time of each interrogation window has been transformed so that the data in byte size have been reduced to bit size and makes the data loading simple. By means of data loading, the cleaned and transformed data have been loaded into the warehouse in the proposed format so as to mine the data easily in the future. From the results of the proposed techniques for data warehousing system, it can be concluded that the performance of the proposed techniques is good enough to meet the system requirements by storing the RFID data efficiently in the warehouse.

Table 5. A sample of the cleaned data from R1 and R2 obtained before transformation.

R_1		R_2	
Time of Interrogation	**Tag IDs**	**Time of Interrogation**	**Tag IDs**
0.0	[1, 3, 4, 5, 6, 8, 2, 9]	0.0	[11, 12, 13, 14, 16, 17, 18, 19, 20, 7, 15]
3.11	[1, 2, 3, 4, 5, 6, 8, 9, 10, 22]	3.12	[12, 14, 17, 18, 19, 7, 16, 11]
4.59	[1, 4, 9, 10, 22, 5, 6, 8]	3.1	[12, 14, 16, 18, 19, 20, 11]
3.11	[1, 2, 3, 4, 5, 6, 8, 9, 10, 22]	3.11	[11, 12, 13, 14, 16, 17, 18, 19, 20, 7]
5.1	[1, 2, 3, 4, 6, 8, 9, 10, 22, 5]	3.1	[11, 12, 13, 14, 17, 18, 16, 19]
3.1	[3, 8, 9, 4, 6, 10]	3.1	[11, 12, 13, 14, 16, 18, 19, 20, 7]
3.11	[1, 2, 3, 4, 6, 8, 9, 10, 22]	5.61	[12, 13, 14, 18, 19, 11, 16]
3.1	[1, 3, 6, 8, 9, 10, 4, 22]	3.1	[11, 12, 13, 14, 16, 18, 20, 7]
3.11	[1, 2, 10, 19, 6, 4, 8]	6.11	[11, 12, 13, 14, 20, 7, 18, 16]
4.6	[1, 2, 3, 4, 6, 8, 9, 10, 22, 19]	6.1	[12, 13, 18, 7, 11]

References

[1] Abowd G., and Mynatt E., "Charting Past, Present, and Future Research in Ubiquitous Computing," *ACM Transactions on Computer-Human Interaction*, vol. 7, no. 1, pp. 29-58, 2000.

[2] Bottani E., "Reengineering, Simulation and Data Analysis of an RFID System," *Journal on Theoretical and Applied Electronic Commerce Research*, vol. 3, no. 1, pp. 13-29, 2008.

[3] Chhillar R. and Kochar B., "A New Efficient Approach for Effective Warehousing of RFID Data: Readers Load Sentient Scheme," *American Journal of Scientific Research*, vol. 3, no. 4, pp. 85-95, 2009.

[4] Chhillar R. and Kochar B., "Extraction Transformation Loading –A Road to Data Warehouse," *in Proceedings of 2nd National Conference on Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries*, Bangkok, pp. 358-362, 2008.

[5] Derakhshan R., Orlowska M., and Li X., "RFID Data Management: Challenges and Opportunities," *in Proceedings of IEEE International Conference on RFID*, TX, pp. 175-182, 2007.

[6] Flies D. and Lopez D., "Building a Classroom Data Warehouse," *in Proceedings of the 34th Instruction and Computing Symposium*, University of Minnesota, USA, pp. 1-7, 2001.

[7] Gonzalez H., Han J., Li X., and Klabjan D., "Warehousing and Analyzing Massive RFID Data Sets," *in Proceedings of 22nd International Conference on Data Engineering*, Champaign, pp. 83, 2006.

[8] Inmon W., "The Data Warehouse and Data Mining," *Communications of the ACM*, vol. 39, no. 11, pp. 49- 50, 1996.

[9] Jarke M., List T., and Koller J., "The Challenge of Process Data Warehousing," *in Proceedings of the 26th International Conference on Very Large Data Bases*, USA, pp. 473- 483, 2000.

[10] Jeffery S., Franklin M., and Garofalakis M., "An Adaptive RFID Middleware for Supporting Metaphysical Data Independence," *The International Journal on Very Large Data Bases*, vol. 17, no. 2, pp. 265-289, 2008.

[11] Lee H., Choi D., Lee S., and Kim H., "A Study on RFID Privacy Mechanism using Mobile Phone," *in Proceedings of World Academy of Science, Engineering and Technology*, USA, pp. 75-78, 2005.

[12] Lee S., Lee J., and Lee B., "Service Composition Techniques Using Data Mining for Ubiquitous Computing Environments," *International Journal on Computer Science and Network Security*, vol. 6, no. 9B, pp. 110- 117, 2006.

[13] Lindell Y. and Pinkas B., "Privacy Preserving Data Mining," *Journal on Cryptology,* vol. 29, no. 2, pp. 36- 54, 2000.

[14] Ling S., Indrawan M., and Loke S., "RFID-Based User Profiling of Fashion Preferences: Blueprint for a Smart Wardrobe," *International Journal on Internet Protocol Technology*, vol. 2, no. 3, pp. 153-164, 2007.

[15] Liu B., Zhang R., and Liu S., "RFID System and its Perspective Analysis with KERGM(1,1) Model," *Journal on Computers*, vol. 3, no. 7, pp. 9-15, 2008.

[16] Mylyy O., "RFID Data Management, Aggregation and Filtering," *in Proceedings of the 31st International Conference on Very Large Data Bases Seminar on RFID Technology*, USA, pp. 6-154, 2007.

[17] Rob M. and Ellis M., "Case Projects in Data Warehousing and Data Mining," *Issues in Information Systems*, vol. 8, no. 1, pp. 1-7, 2007.

[18] Romero C., Ventura S., and Garcia E., "Data Mining in Course Management Systems: Moodle Case Study and Tutorial," *Journal on Computers and Education*, vol. 51, no. 1, pp. 368-384, 2008.

[19] Santillo L., "Size and Estimation of Data Warehouse Systems," *in Proceedings of the 4th European Conference on Software Measurement and ICT Control FESMA DASMA*, Germany, pp. 173-184, 2001.

[20] Stankovski V., Swain M., Kravtsov V., Niessen T., Wegener D., Kindermann J., and Dubitzky

W., "Grid-Enabling Data Mining Applications with DataMiningGrid: An architectural Perspective," *Journal of Future Generation Computer Systems*, vol. 24, no. 4, pp. 259-279, 2008.

[21] Telbany M., Warda M., and Borahy M., "Mining the Classification Rules for Egyptian Rice Diseases," *The International Arab Journal of Information Technology*, vol. 3, no. 4, pp. 303-307, 2006.

[22] Vassiliadis P., Quix C., Vassiliou Y., and Jarke M., "Data Warehouse Process Management," *Journal of Information Systems*, vol. 26, no. 3, pp. 205-236, 2001.

[23] Velpula V. and Gudipudi D., "Behavior-Anomaly-Based System for Detecting Insider Attacks and Data Mining," *International Journal on Recent Trends in Engineering*, vol. 1, no. 2, pp. 261-266, 2009.

[24] Wadhwa V. and Lin D., "Radio Frequency Identification: A New Opportunity for Data Science," *Journal on Data Sciences*, vol. 6, no. 3, pp. 369-388, 2008.

[25] Wikramanayake G. and Goonetillake J., "Managing Very Large Databases and Data Warehousing," *Sri Lankan Journal on Librarianship and Information Management*, vol. 2, no. 1, pp. 22- 29, 2006.

Barjesh Kochar is pursuing his PhD under the guidance of Dr. Rajender Singh Chhillar and currently working as the head of MCA in GNIM, New Delhi. He has published a number of research papers in international, national level conferences and journals.

Rajender Chhillar a Head of Department of Computer Science and Applications in M.D.U Rohtak, India. He has published a number of research papers on various topics of computer science in international, national journals and conferences.