A Cloud mining model for SMEs in E-commerce in Zimbabwe

[1]Tatenda D Kavu, [2]Paidamoyo Mastara, [3]Luckson Phiri, [4]Bethany Jari

Abstract— Cloud mining is an important research topic in the field of data mining and knowledge discovery. Due to the fact that a number of organizations have opted for cloud storage, it is important to use efficient and effective data mining methods to mine the large databases in the cloud so as to extract the previously unknown interesting patterns and relationships. The data patterns are of importance in business intelligence thus for forecasting, market basket analysis, cross selling, market segmentation and customer retention purposes which helps to increase the profits of business organizations. The cloud is a large repository of data units on the World Wide Web which contains social, cultural, political, educational, academic and commercial data which is usually accessed via a web browser. The data repositories can be used by all forms of business enterprises including Small and Medium Enterprises for knowledge discovery to gain business insight and intelligence. This paper seeks to investigate current cloud e-commerce systems in the world and design a model that simulates knowledge extraction for business intelligence from a cloud which store SMEs' data in Zimbabwe.

Index Terms— Business Intelligence, Cloud mining, Data mining, E-commerce, SMEs (Small and Medium Enterprises)

1 Introduction

Cloud is defined as an elastic computing model from which the users can lease the resources from the rentable infrastructure over the internet [2]. Cloud computing is gaining popularity due to its lower cost, high reliability and huge availability [3]. Many companies have opted to the use of business applications and databases hosted in the cloud due to the relatively cheaper cloud services. E-commerce has become common and the transactions are also stored in large databases in the cloud. The databases have a collection of information about all the products bought by the customer.

Cloud mining is a new approach to apply data mining, to the customer data in the cloud [5]. In this paper the data was extracted for analysis from databases in a simulated cloud platform so as to extract knowledge and patterns that help make decisions for the future thus bringing in the concept of business intelligence. Business Intelligence has become the prevalent decision support systems in organizations [2]. Companies especially SMEs for their customer base growth need to keep track of their customers, prospect potential customers interact with them and try to forecast what they will buy in future. With an increase of companies that offer cloud services in Zimbabwe such as Yo Africa, Utande and Realtime BYTES, big organizations and SMEs have all the room to install their services in the cloud due to its advantages stated before. This means more structured and unstructured data streams will continue to grow which when exploited can give rich business knowledge and insights [6]. Many organizations have collected and stored large repositories of data about their current customers, potential customers, suppliers and business partners. However, the inability to discover valuable information hidden in the data repositories especially for SMEs in Zimbabwe prevents the organizations from transforming these data into valuable and useful knowledge. [7]. "We are drowning in data but starving for knowledge" [8].
Cloud computing is currently being adopted by SMEs in Zimbabwe. The customer data contains useful patterns (that is association model, classification model, class model, sequence pattern and so on) [9] which are of relevance to commercial decision making such as market analysis and management, risk analysis and management. This paper aimed to design a model for cloud mining and analysis in SMEs e-commerce databases so as to come up with optimization in business intelligence, and in particular analysis of large collections of transactional databases in the cloud so as to depict customer buying patterns and behaviors. This model offers a unique way of extracting knowledge in SMEs databases due to their informal way of doing business and offering payments, whereby SMEs can offer different methods of payments on the same products.

This study provides answers to the following business questions: 1) which kinds of products are often bought by customers? 2) What combinations of products are usually bought by customers 3) what kinds of products are likely to be preferred by customers in the future?

2 Related Work

Data mining is a technique of extracting previously unknown and potentially useful patterns and knowledge from large amounts of data [12]. Cloud mining is a process of applying data mining techniques to large repositories of data in the cloud. There are relatively new and emerging mining techniques that are known collectively as reality mining. Reality mining is the collection of transactions made daily by individuals to realize how they live and react [2]. The paper mainly focused on mixed methods of mining in large repositories which are: Frequent pattern mining, clustering and classification.

2.1 Frequent pattern mining

It was first proposed by Agrawal in 1993 for market basket analysis in the form of association rule mining. [12]. Frequent

patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold [13]. Frequent pattern mining is a data mining technique that searches for recurring relationships in a given dataset. It analyses customer buying habits by searching for associations between the different items that customers place in their "shopping baskets". Frequent pattern mining has become an important data mining task as it plays an essential role in mining associations, correlations and many other data mining tasks such as data classification and clustering [12].

2.2 Current methods for mining frequent patterns

Frequent item set was first introduced for mining transaction databases, that is Let $\boldsymbol{I} = \{i_1, i_2, \ldots, i_n\}$ be a set of all items. A k-itemset α, which consist of k items from $\boldsymbol{I}$, is frequent if α occurs in a transaction database $\mathbf{D}$ no lower than $\theta|\mathbf{D}|$ times, where θ is a user specified minimum support threshold (min_sup), and $|\mathbf{D}|$ is the total number of transactions in $\mathbf{D}$. The three basic frequent itemset mining methodologies are Apriori, FP-growth and Eclat. [13]

2.2.1 Clustering

Clustering aims to divide a data set into groups that are very different from each other and whose members are very similar to each other .It has been used for grouping users with common browsing behavior, as well as grouping web pages with similar content [17].User profiles can be extracted from the clusters designed, and it is clustering that will form the basis of personalization [16]. Some of the clustering known methods are fuzzy clustering, partitioning, model based and hierarchical.

2.2 Classification

The goal of classification is to identify the distinguishing characteristics of predefined classes, based on a set of instances; on this case are the users [17] .Classification is supervised learning and this information will be used for predicting how the instances in classes and new instances will behave. Classification consists of different algorithms and methods which are: decision tree induction, LCS algorithm, neural networks, rough set theory and Bayesian classifier

One problem of data mining in the cloud has been investigated from the data mining algorithm perspective. Wang utilized the powerful and huge capacity of cloud computing into data mining and machine learning. In the experiments, three algorithms, which are: global effect (GE), K-nearest neighbor (KNN) and restricted Boltzmann machine (RBM) were performed in cloud computing platforms, which use the S3 and EC2 of Amazon Web Services [15]. And they built two predictors based on KNN model and RBM model respectively with the order to testify their performance based on cloud computing platforms. [6]

The Mapreduce programming model was designed for processing massive data sets in a parallel network. Based on this programming model, Wang adapted the SPRINT algorithm which is an ideal tool for data classification. SPRINT has been designed to be easily parallelized. Due to the parallelism, the original SPRINT was modified to be implemented in Hadoop architecture [6]. The algorithm divided data sets in vertical direction and horizontal direction respectively, in accordance with the "Map" step in Mapreduce. The vertical partition separated datasets by attribute, while horizontal partition produced many itemsets. They applied the revised SPRINT algorithm to classify customer groups with different credit grades. [6]

It is hypothesized that the existing techniques have either only improved the efficiency but do not give any effect to the change in the complexity collectively. Moreover the existing published work has not targeted the untapped SMEs e-commerce businesses in the cloud. The paper proposed a technique by which a database can be scanned only once and start working on the tables which are created by the user only at the time of analysis. The main objective focused on adopting algorithms which are both efficient and have better space complexity, and apply them in mining large repositories in the cloud for SMEs. This scheme gives good response and better performance, which means less time, is needed for execution thereby increasing throughput. Hence when the execution time is reduced the cost is also reduced, which become more favorable to SMEs.

3 METHODOLOGY

The investigation made use of eight different transactional databases which are of different small retailers. The different databases represented the organization's databases being hosted in the cloud. A point of sale was created which imitated real world retail shop point of sales and store the transactions in the created databases. The databases contain the customer's details and the transactions done by each customer. The collections of transactions in the databases are mined so as to find the interesting patterns by making use of the selected algorithms. The item sets in each transaction of a customer are chosen at random, thus assuming how different customers behave when purchasing in a shop. Different algorithms were applied to data set in the database so as to find interesting patterns such as frequent patterns, associations and clusters. Some of these algorithms vary in performance depending on the size of the databases.

3.1 Development tools

The following development tools were used which included: Microsoft SQL server 2008, SQL Server Management Studio, Business Intelligence Development Studio, Microsoft Visual Studio 2008, ADOMD.NET and the .NET Framework.
SQL server 2008 was preferred because of a number of reasons such as backup support, including differential back, distributed database, distributed queries, GUI support for managing databases and its ability to serve large quantities of data. It also has the capacity of integrating multiple databases into a

vast single and secure data warehouse. It also includes Business Intelligence and Development Studio which is used for developing data analysis and Business Intelligence solutions by utilizing its Analysis Services, Reporting services and integration services.

3.2 INTERVIEWS

Some interviews were carried out to investigate current systems which are being used by SMEs in Zimbabwe; this was a qualitative approach which was done to help the design of the model. The population consisted of selected SMEs in Zimbabwe. For the purpose of the research the sample consisted of ten SMEs in the capital city Harare, where many people use Information Communication Technology (ICT) on a daily basis.

The information gathered from SMEs highlighted that many of them are appreciating technology to the extent that 70% of them are selling their products online and using POS (Point Of Sale) technology. From the gathered information not one from the approached SMEs are using cloud technology though all acknowledged the presents of cloud services in Zimbabwe. Results from the interviews supplied relevant information which helped much in the development of the model.

3.3 Framework for the proposed algorithm

Input: A Product Database
Output: All high frequent Item sets
Method:
1. Distribute large database from master node to all the Slave nodes.
2. Each slave node scan local database.
3. Start selecting items from the datasets.
    a. Select multiple items
    b. Add them to the transaction list
    c. Maintain a transaction dictionary of the transaction id and the items chosen for that transaction.
    d. Loop
4. Enter the minimum local support and confidence.
5. The frequent itemsets are generated using the Apriori Algorithm.
6. Compute the support value of item sets of each node.
7. Send the support value of each node to the Master node.
8. Master node compute support values of item sets of order **k+1**.
9. Broadcast the support values to all slave nodes.
10. Master node builds the global transaction utility table and prunes the items that do not satisfy the given threshold *min_support.*

The Algorithm was incorporated in the design of the real model using VB.Net.

4. RESULTS AND ANALYSIS

The system imitated the normal point of sale and allowed capturing of customer transactions which were then stored in the databases in the cloud. The data mining techniques were then applied to the stored transactions that include Apriori algorithm for finding frequent patterns and associations and the K-means for finding the clusters. This cloud mining process aimed to produce an analysis of customer patterns from the past and present transactional data and provides information that help retailers to identify opportunities and to help them to decide if the retail market is a potentially profitable segment, after the evaluation of the responsiveness of the customers. The results obtained helps firms and organizations to enhance their marketing strategies such as to target marketing thus knowing which type of customers prefer what products, arranging totally related products on the same shelf, such as cereals and milk and market segmentation.

The model designed involves simulation of the POS and administrative tasks; Figure 2 illustrates the basic moves done by the administrator to view the current trends happening in SMEs world of business**.**

The model was implemented based on a two-tier architecture with two different layers of the system that is the back-end and the front-end (graphical user interface) .The two layers directly interact with each other. The two tier architecture has a greater advantage of that it masks all the operations that occur at the back end to a customer.

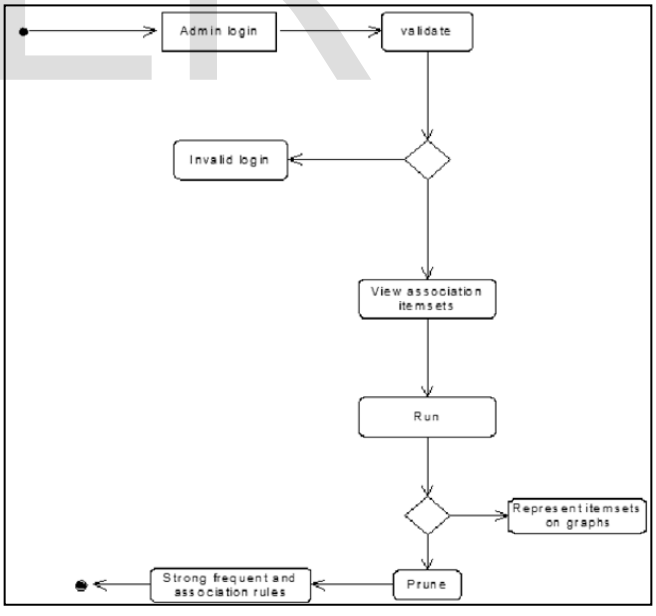

Figure 1: Activity diagram for the administrator

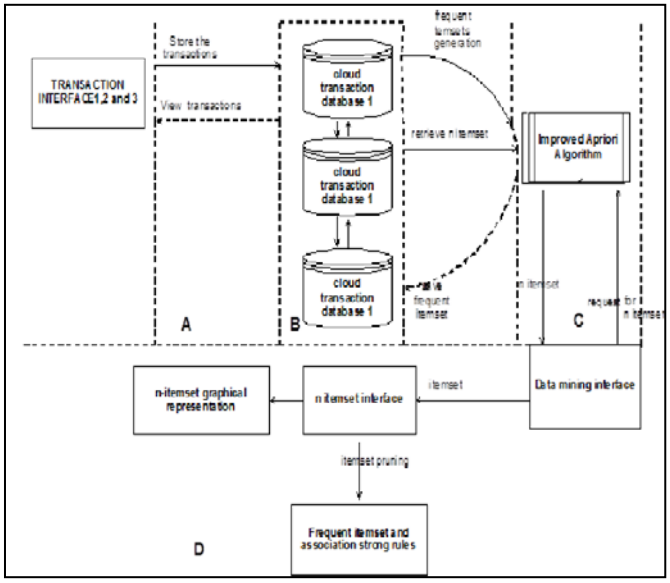

Figure 2: System architecture

4.1 FINDINGS

A figure below (Figure 3) is the central main form that allows the user to navigate through the whole system with the use of different menu tabs. The main menu dashboard is synchronized to the cloud such that all the information which is input or output involves interaction with cloud databases.

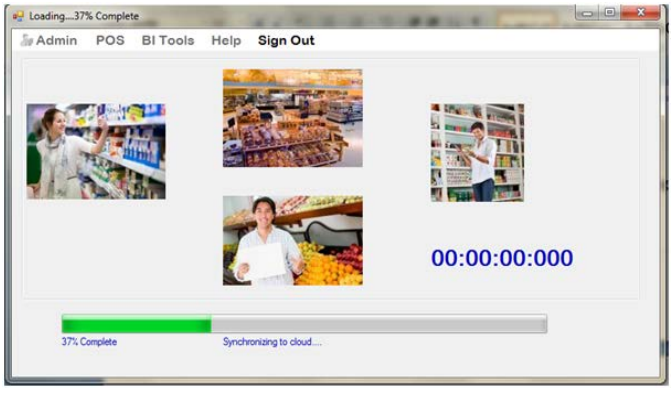

Figure 3: The main dashboard

The transactional form is one of the points of sales out of three where a customer can purchase some products by adding them to cart. A customer can view all the products in his or her shopping basket.

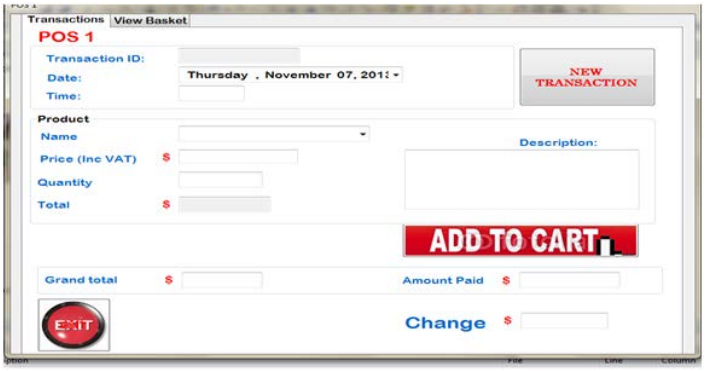

Figure 4: Transaction form

Figure 5 below displays updated relationship patterns of 2 frequent patterns. From this interface the administrator can find the most selling two products and determine from the products those with a support.

Figure 5: Frequent patterns for 2 itemsets

Figure 6 displays frequent patterns for 6 itemsets and their support. It is the maximum frequent pattern item sets which have been implemented by the model.

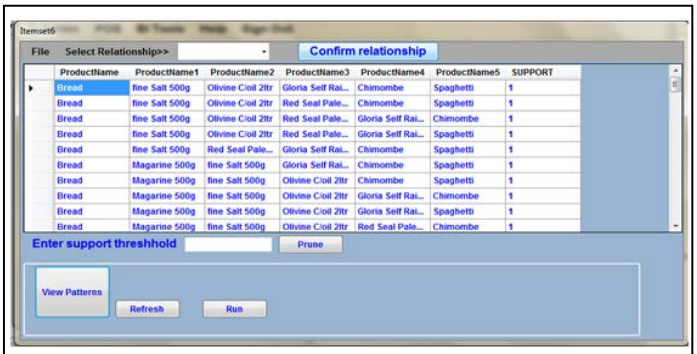

Figure 6: Frequent pattern for six itemsets

Figure 7; represent sales of products discovered from the da-

tabases. The administrator has the choice of selecting POS of sales of which to check business trends.

Figure 7: Graphical representation of trends on POS 1

5.0 CONCLUSION AND RECOMMENDATIONS

The paper recommends that the researches proposed for the cloud transactional databases mining be solved using real world transactional databases. The mining of the transactional databases for frequent and association patterns were based mainly on the Apriori algorithms. However the Apriori algorithms have also their limitations. The researcher therefore recommends future research advancements to make use of some better algorithms such as those based on the FP-Growth techniques which are much faster. The mining of transactional databases can be applied different data mining techniques such as classifications and clustering. The different data mining techniques can best be applied by making use of a combination of different algorithms. No single data mining algorithm can best solve all the problems.

Sequential patterns tell us what items are frequently bought together and in what order, but they cannot provide much information about the time span about which they are frequently bought together for further decision support. Although we know which items will be bought after the other items, we have no idea when the next purchase will occur. So the researcher recommends for further improvements on that.

6.0 References

[1] J.S. Bridle, "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition," *Neurocomputing—Algorithms, Architectures and Applications*, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)

[2] Abdul-Aziz Rashid Al-Azmi, "International Journal of Data Mining & Knowledge Management Process (IJDKP)," *DATA, TEXT, AND WEB MINING FOR BUSINESS INTELLIGENCE: A SURVEY*, vol. Vol.3, no. 2, March 2013.

[3] H Nuseibeh, "Adoption of cloud computing in organisations," , 2011, pp. 1-8.

[4]

[5] Guido Deutsch. (2013) Cloud Mining - CRM Data Mining in the Cloud. [Online]. http://www.data-mining-blog.com

[6] Shanlin Yang Jing Ding, "Classification Rules Mining Model with Genetic Algorithm in Cloud Computing," International Journal of Computer Applications, vol. 48, no. 18, June 2012.

[7] Li Xiu,D.C.K Chau E.W.T. Ngai, "Expert systems with applications," Application of data mining techniques in customer relationship management:A literature review and classification, vol. 36, pp. 2592–2602, 2009.

[8] Motaz K. Saad, Data Mining & Business Intelligence: Practical Tools, Oct. 2007, Dept. of CS –College of IT.

[9] N.C.Mahanti Kanhaiya Lal, "A Novel Data Mining Algorithm for Semantic Web Based Data," International Journal of Computer Science and Security (IJCSS), vol. 4, no. 2, pp. 160-175.

[10]

[11]

[12] Jiawei Han, Data mining concepts and techniques, 3rd ed., Jian Pei Michelin Kamber, Ed. Tokyo,London,Oxford, Paris: Morgan Kaufmann.

[13] Hong Cheng, Dong Xin, Xifeng Yan Jiawei Han, "Frequent pattern mining: current status and future directions," Data Min Knowl Disc, no. 15, pp. 55-87, January 2007.

[14]

[15] Jinhua Fan, "Mining Classification knowledge based on cloud models," Nanjing Communications Engineering Institute, Nanjing,China, 210016, 1999.

[16] Olfa Nasraoui, M.S.S.B.a.R.G., 2008. A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 20(2), pp.202-04.

[17] Pierrakos, D., PALIOURAS, G.O., PAPATHEODOROU, C. & SPYROPOULOS, C.D., 2003. Web Usage Mining as a Tool for Personalization. *Kluwer Academic Publishers.*, p.314.