A Novel Mapreduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data

Nour E. Oweis
Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava
Ostrava, Czech Republic

Mohamed Mostafa Fouad
Arab Academy for Science, Technology, and Maritime Transport
Cairo, Egypt

Sami R. Oweis
Alumni of Electrical and Computer Engineering,
Oakland University
Rochester, MI, USA

Suhail S. Owais
Department of Computer Science, FIT, Applied Science University
Amman, Jordan

Vaclav Snasel
Faculty of Electrical Engineering and Computer Science,
VSB-Technical University of Ostrava
Ostrava, Czech Republic

***Abstract*—Big Data mining is an analytic process used to discover the hidden knowledge and patterns from a massive, complex, and multi-dimensional dataset. Single-processor's memory and CPU resources are very limited, which makes the algorithm performance ineffective. Recently, there has been renewed interest in using association rule mining (ARM) in Big Data to uncover relationships between what seems to be unrelated. However, the traditional discovery ARM techniques are unable to handle this huge amount of data. Therefore, there is a vital need to scalable and parallel strategies for ARM based on Big Data approaches. This paper develops a novel MapReduce framework for an association rule algorithm based on Lift interestingness measurement (MRLAR) which can handle massive datasets with a large number of nodes. The experimental result shows the efficiency of the proposed algorithm to measure the correlations between itemsets through integrating the uses of MapReduce and LIM instead of depending on confidence.**

Keywords—Big Data; Data Mining; Association Rule; MapReduce; Lift Interesting Measurement

I. Introduction

The recent advances in computers and communications have increased the number of relevant applications associated, such as a Radio Frequency Identification Devices (RFID), Wireless Sensor Networks (WSNs), Internet of Things (IoT), and other applications. Usually, these applications create a huge stream of non-stop data that is currently well-known as big data, denoted as “Big Data” [1]. Big Data is a massive set of data that is too complex to be managed by traditional applications. Nowadays, it includes huge, complex, and abundant structured, semi-structure, and unstructured data as well as hidden data that are generated and gathered from several fields and resources [2]. There are many challenges to manage such sets of Big Data include extracting, analyzing, visualizing, sharing, storage, transferring and searching [3]. These data (Big Data) are stored in powerful computers; include many of hidden patterns indicators that help in decision making. Data mining approaches facilitate decision making through determining and explain those patterns in a meaningful knowledge format [4].

Since, traditional data processing approaches and its applications could not be directly implanted when working with the big data management [5], it is necessary to apply new techniques, algorithms, and frameworks to manage, extract, and execute the big data mining development, and make these data mining techniques very helpful and more efficient.

Frequent pattern mining is one of the well-known data mining techniques that focused on discovering a number of interesting patterns from a large set of data items [6]. The association rule is a frequent pattern mining that is usually applied to find all the frequent co-occurrence relationships from a set of transactions [7]. Usually, the association rule strength is measured through two parameters (Support and Confidence). However these two parameters may not be sufficient to discover some interesting patterns, thus another measuring criteria is used which is the “Lift” [8].

On the other hand MapReduce is a parallel-based approach proposed for parallel processing of large datasets. This paper investigates the efficiency of the integration of both approaches (the Association Rule and the MapReduce). Therefore this paper proposed a parallel-based MapReduce approach for an association rule algorithm based on the Lift interestingness measurement.

This paper presents a MapReduce approach that has been used for defining the association rule importance based on the Lift interestingness measurement. This approach can be easily applied to many commodity machines to deal with big data. Finally, the work presented in this paper is in agreement with the published literature [1, 9, 10], and concludes that the traditional data processing tools and its applications are incapableof handling the current huge and complex data, such as, managing big data mining, and the newest industrial age of the IoT. Therefore, the parallel algorithm is the suitable solution for the big data mining techniques. This paper proposes a solution to solve one of the most critical problems of big data mining by emerging data mining, big data with parallelization and association rule to improve the usage of huge, complex amount of dataset.

Following the first section, this paper is organized as follows: Section two covers the background for association rule, and big data including the MapReduce paradigm. Section three covers the related work topics including the parallel association rule by utilizing several methods, measures, and techniques. Section four presents the proposed algorithm including the dataset, software use, and the novel parallel-based MapReduce approach for an association rule algorithm based on Lift interestingness measurement. Section five covers the experiments results and conclusion.

II. Background

This section presents the background information about the main topics used in this paper and the following subsection defines the association rule including different measured such as: support, confidence, and Lift interestingness measure.

A. *Association Rules*

Data mining techniques contains a variety of applications and notable uses which are designed to work skillfully with a very huge amount of data. These applications and there notable uses cover wide domains of our life, including social networks, health care, financial, communications and many more.

Data mining approaches can be classified into two major models [11]: The descriptive data mining and the predictive data mining models. The descriptive models are unsupervised learning that describe the historical events, and the presumed or real relationship between elements that created them. These models uses a summarization analysis tools including multiple techniques such as, association rule for discovering and extracting relevant data. This model is commonly used in marketing analysis [11].

The predictive models are supervised learning that can accurately predict future outcomes based on existing data that carries out the analysis and extraction in more specifications and classifications. This model is commonly used in marketing predictions to forecast which new products may be popular in the future [12].

The association rules used within a dataset to discover non-trivial hidden patterns between items in a set could utilize either descriptive or predictive models [13]. In many cases, the algorithms generate large number of association rules, often in thousands or millions. It is almost impossible for users to visualize or validate such a large number of complex association rules, which limits the usefulness of data mining results. Therefore, it is important to identify that the components of an association rule are two sets of items: Left Hand Side (LHS) and Right Hand Side (RHS). The LHS is the antecedent (an item found in the transactions) and the RHS is the consequent (an item that is found in combination with the former) .Moreover, there are two well-known measurements for an association rule, the Support of the rule and the Confidence of a rule [13].

1) *The Support Rule*

The Support rule is considered a global measurement of interest for an itemset denoted by *Supp(X)*. The *Supp(X)* is calculated by counting the number of proportion transactions (P) within the dataset as shown in (1) [14]:

$$Supp(X) = P(X \cup Y) \qquad (1)$$

where:

X is an itemset of interest.

Y is an itemset with a defined condition of interest.

The itemset is called "frequent itemset" once its support output value is higher than a given minimum support.

2) *The Confidence Rule*

The Confidence rule is considered a localization measure of correlation between X and Y denoted by $Confidence\ (X \rightarrow Y)$. The confidence rule is calculated as the ratio between the support of the union between X and Y subsets and the support of X as shown in (2) [15].

$$Confidence\ (X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp\ (X)} \qquad (2)$$

These two measures (Support and Confidence) may not be enough to extract some hidden patterns and to determine the correlation rule between LHS and RHS as mentioned in previous research work [16, 17, 18]. Therefore, an additional measurement is used; the Lift interestingness measure (LIM).

3) *The Lift Rule*

Lift interestingness measure defines the number of transactions that contain the items used to find interesting patterns. The Lift measure is denoted by $Lift\ (X \rightarrow Y)$ as shown in (3) [19].

$$Lift\ (X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp\ (X) \times Supp(Y)} \qquad (3)$$

The Lift rule output defines the correlation between the LHS and RHS as follows:

a) Lift > 1 → Positive Correlation.

b) Lift < 1 → Negative Correlation.

c) Lift = 1 → Independent Correlation.

While most of the current algorithms proposed with regards to the association rule, such as Apriori [20], depends on either support, confidence or a combination of both rules, this paper utilized the "Lift measurement" instead of confidence to extract the association rules.

The proposed algorithm depends mainly on the support measurement, in contrast to other algorithms that depend on both support and confidence with Lift to extract association rules. Therefore, it simplifies the Lift measurement because it depends on support only.

B. *Big Data*

Big Data is a complex, heterogeneous, massive, and hidden set of data that is hard to be managed, processed, analyses, and visualize by traditional applications. Gartner defined the term of big data as: "Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [11, 21].

Although the Volume, Variety, and Velocity are the most common dimensions of the big data. There are other dimensions "Vs" that are recently defined such as: Veracity, Viability and value. While, the Volume describes the huge data capacity, the Velocity describes the speed of the data transmission and processing per interval time. The Variety defines the heterogeneity of data types. The Veracity focuses on the data quality (data cleaning from several noises). Finally the Viability and the Value dimensions: While the former defines the different data prediction possibilities, the later tends to gain valuable knowledge [22, 23]. Understanding these dimensions is important for designing big data mining techniques and platforms.

There are multiple big data analytic techniques used to extract, analyze, and visualize the complex and different data types. In the following section will introduce the MapReduce approaches as a main paradigm used in this paper.

C. MapReduce:

MapReduce [24] is a data mining paradigm developed by Google that allows programmers to implement and processing large dataset with parallel and distributed algorithm on a cluster computing by using several programming languages such as: Matlab, C, C++, Java, Perl and more by using several MapReduce libraries [25, 26, 27].

MapReduce consists of two consequence processes, *Map* and *Reduce* and their functionally is defined as follows: The Map process is responsible for dividing, filtering, and sorting data tuples (key/value pairs) within using number of distributed clusters, the Reduce process summarizes the results into a few set of tuples [10, 28].

The advantages of MapReduce approach are many, for example: the big data classification approaches [29], the online Machine Learning for multicore and automatically failure handling [11, 30]. Parallelism also gives some possibilities partial recovery server failures: if the operating portion, which produces a pre-processing operation or convolution fails, its operation may be transferred to another working unit (assuming that the input data for the ongoing operation are available) and in others recent applications [31]. The most popular open source implementation of the MapReduce is the Apache Hadoop [32].

III. Related Work

The most well-known association rule algorithm is Apriori. Agrawal R. et al. [20] proposed the Apriori algorithm to extract relationships between data. This was done through applying a pruning technique to make the number of candidate itemsets much smaller and then find the frequent patterns to generate an association rule.

An improved version of Apriori algorithm was proposed by Aflori, and Craus [33]. It entitled the Frequent Pattern Growth (FP-Growth) algorithm capable for repeatedly reducing the search costs for short patterns. These patterns are linked to long frequent patterns to offer high selectivity mechanism.

Among the current research, there are several proposed parallel algorithms for association rule, such as the Parallel Apriori. This algorithm, proposed by Yang, X. Y., et al [34], is a parallel implementation of Apriori algorithm based on the MapReduce approach. The algorithm gives a solution to the exponential growth of data that encounter the traditional association rule mining techniques. This algorithm shows its benefits to deal with big data without consider the synchronization problem.

Jongwook Woo [35] proposed an Apriori-Map/Reduce algorithm and with both time and complexity, which theoretically shows that the algorithm provide much higher performance than the sequential algorithm as the map and reduce nodes get added. Also the paper shows that the itemsets produced by the algorithm can be adopted to compute and produce association rule for market analysis.

Chen, Y., et. al [36] The Parallel Randomized Algorithm for Approximate Association Rules Mining in MapReduce (PARMA) minimizes the data replication, the communication cost and the runtime improvement over parallel FP-Growth (PFP). The algorithm randomly separates the data into sets of samples. The machines works in parallel with their assigned set to produce deliverables and to be filtered and aggregated into a single output set.

Lin, X, et al. [37] proposed a parallel association rule algorithm called Niche-Aided Gene Expression Programming (NGEP). The advantage of the NGEP over both Apriori and FP-Growth is its efficiency to achieve more association rules with a higher accuracy rate.

Zhou, X., et al. [38] proposed an improved parallel association rules algorithm utilizing Hadoop as the MapReduce distributed programming framework. It has shown that the algorithm achieve well based on parallel performance and could be easily realized with the Hadoop platform.

Based on the previous literature, and briefly say that parallel association rule algorithms is one of the best choices for high performance big data mining techniques. Therefore this study proposes the application of the MapReduce approach for a paralyze association rule algorithm that is based on the Lift interestingness measurement.

IV. The Proposed Algorithm

As stated previously, the correlations between data items using the association rule measurements, is usually based on both confidence and support interests measurements. However the use of confidence is not effective to determine the association rules, since it does not describe the type of correlation(s) between the LHS and the RHS in the association rules.

The proposed algorithm that has been entitled “MapReduce-Based for Lift Association Rule (MRLAR)” is based on the Lift-Based Algorithm (LBA) [16] which is illustrated in Algorithm 1. Where the MRLAR improves the LBA algorithm through parallel executing. In which, the proposed algorithm works to determine the type of correlation between LHS and RHS in parallel association rules. The MRLAR algorithm was illustrated in Fig. 1.

The functionalities of the proposed algorithm MRLAR are discussed as follows:

- ***Map function:*** *This step combines two steps; the data splitting step and the Mapping step [39]. The splitting step performed to distribute the data across each separated Map nodes. The map step consists of a map function that was established to find the association rules for some entities within a large sized database.*

Algorithm 1: The Lift-Based Algorithm (LBA) [16].

```
1.  Scans the database to find frequent items based-on minmumin slections provided by the user
2.  Choose correlation of association rule based-on Lift interestengniss measure
        a.  If Lift > 1, positive correlation, Insert α. (In this choice lift value > 1+ α)
        b.  If Lift < 1, negative correlation, Insert β. (In this choice lift value < 1- β)
        c.  If Lift = 1, dependent correlation, Insert α and β. (In this choice lift value is between (1- β and 1+ α).
        d.  Else, insert α and β.
3.  Second scan of the database
4.  Get a frequent item for 1, 2 and 3 item by sequentially check for per value in item:
        a.  If  >= MinSupport added to frequent items
        b.  Else Ignore.
                i.   Generate candidate association rules from frequent items.
                ii.  Calculate the lift value for each candidate association rules to classify
5.  Generate Association Rule for the choosing correlation of sssociation rule.
6.  If not found result (not found association rules), go to step 1 to edit MinSupport or edit the type of correlation).
```

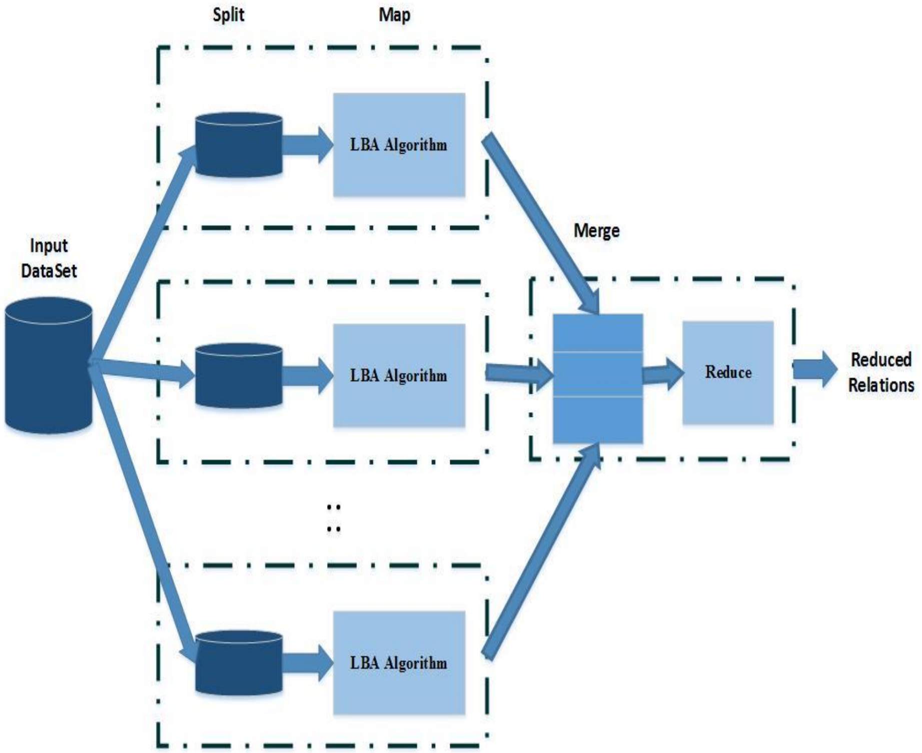

Fig. 1. Illustration of the MapReduce-Based for Lift Association Rule (MRLAR)

The Lift interestingness between some database key values is tended measurement within the MRLAR. Where selecting the key values from the user and based on the choices of the MRLAR can narrow down the search space in order to extract only the association rules of interest.

- ***Reduce function:*** *The reduce process combines the outputs generated by each map node(s) to form the final collected association rules [37]. As it was mentioned previously the rules are weighted by the reduce process not the confident, but as an alternative in this parallel association rules where the authors involved the Lift weight computation to define the correlation between LHS and RHS.*

The algorithm was designed and operated using MATLAB (R2015a) since it is well-known for its ability to support big data enhancement especially by using MapReduce approach [17]. As for experiments the algorithm used a dataset that was provided by the USA domestic airline flights between the period of 1987 and 2008.

The data comes originally from the Research and Innovative Technology Administration (RITA), the dataset is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed with a collection of records consisting of 29 variables of flight information for several airline carries, including arrival and departure times with CSV files format.

These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals. The full dataset can be downloaded from (http://stat-computing.org/dataexpo/2009)

In preparation for the collected data to remove irrelevant incomplete key values prior the startup of the MRLAR algorithm, two main processes were initiated; the data preprocessing and the data decoding.

- ***Data Preprocessing****: Once the data was gathered from the RITA site and before it underwent the decoding process, preprocessing techniques was used to clean it up in order to make sure the data "datasotre" are free of empty cells, Not a Number (NaN) data, or any misrepresented data as strings which potentially could stop the execution of the program and/or result in wrong indication. The NaN and empty cells were assigned to zero.*

- ***Data Decoding****: Then the decoding process started, for Airport codes, which are identified as letters. Using MATLAB that able to find the unique list then assigning a unique digital code to every element of the list as shown in Table I column A. Another way was used to decode the strings by assigning a digit from 1-26 to each letter in the alphabets A-Z as shown in Table I column B.*

TABLE I. ILLUSTRATION OF UNIQUE DIGITAL CODE AND STRING DECODING

A. Unique Digital Code			B. **String Decoding by Digit**	
Code	**Decode**		**xStr**	**x_Code**
ABE	100		A	1
ABI	101		B	2
ABQ	102		C	3
ABY	103		D	4
ACK	104		E	5
ACT	105		F	6
ACV	106		G	7
ACY	107		H	8
ADQ	108		I	9
AEX	109		J	10

Now since a digitalized file with only strings as the column headers, and used the "*datastore*" to read the file into memory. After that, scan the file for duplicated entries using the command "*unique*" when the data processing begin.

- ***Initializing and key generation:*** *One of the most important reasons for designing a MapReduce algorithm based on Lift interestingness measure is to minimize the number of keys needed to be generated. This can be achieved by grouping the items by transactions. For example using "Day of Months" as a key from a collection of tabular dataset.*

The Mapper function used to find the Lift method then pass this key-value pair to the reducer function. The reducer receives key-value pairs by key, and merges multiple cell arrays with the same key into a single cell array. Subsequently can then store the result in a new datastore area. The whole MRLAR algorithm and its processing steps are illustrated in the Fig. 2.

V. EXPERIMENTS AND RESULTS

Many of association rules use "Confidence" and "Support" measures for testing the occurrence of the itemset. Relying solely on both of them may be not sufficient. In this study algorithm, a test was setup as shown in Table II, for the association rules with their support, confidence, and lift measures with one attribute at a time (single-dimension association rule), that was "the delay in flights arrival of more than 60 minutes or more than 120 minutes", also "the delay in flights departure of more than 30, 40, or 90 minutes"

Table II shows the reliability of the relationship between support and confidence at the single dimensionality level. For example, the first two tests ID numbers 1 and 2 in Table II showed that while the support drops from 0.045 to 0.013 the confidence also followed from 0.093 to 0.027 with a positive lift association's value (22.3 and 75.95 respectively). Another point which can be noted is that the confidence increases while the number of itemsets increased.

The same results can be seen in ID numbers 3 and 4 with the same trend between confidence and support as well as the lift measure. However, when the data was challenged at the multidimensional level, the interpretation of the "Support and Confidence" can be misleading. For example, Items 3 and 4 in Table III showed inconsistency in the "support and confidence" trend. While the support has decreased in case 3 (Table III) from 0.0057 to 0.0044 in item 4 (Table III), the confidence has increased between these two cases (0.066 to 0.067). Hence, the use of "support and confidence" model only can survive under the single dimension association rule, but this is not applicable at the multidimensional level. Therefore, there is a need for a novel measure that is being able to be applicable at the multidimensional level, which is in this proposal the lift measure. It is important to state that relying on the lift adds benefits to the prediction process of the future consequence in future datasets with comparing to the current data.

In order to achieve more efficiency with high performance testing, parallel processing based on MapReduce was added in this study. The next experiment integrated multiple attributes with the same dataset. The experiments have been performed with the same two columns of interest (arrival delay and departure delay) but combined them with another attribute that was the month of the year (June and October). Table III shows the measurements for all support, confidence, and Lift measures (multi-dimensional association rule). The results showed that our approach has a high ability to run under several attributes, with the Lift interestingness measures successfully being able to determine the type of correlation between itemsets (positively, negatively, and independent) between LHS and RHS instead of using support and confidence.

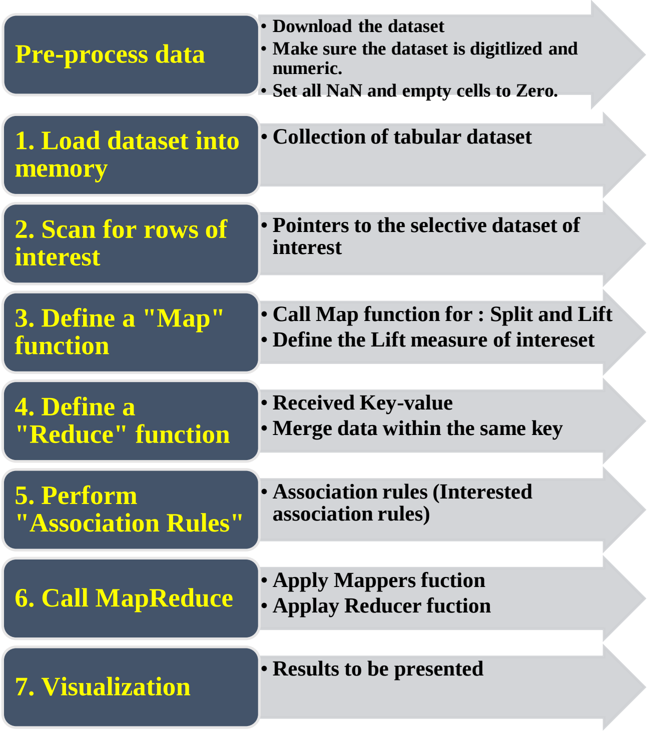
Pre-process data
• Download the dataset
• Make sure the dataset is digitlized and numeric.
• Set all NaN and empty cells to Zero.
1. Load dataset into memory
• Collection of tabular dataset
2. Scan for rows of interest
• Pointers to the selective dataset of interest
3. Define a "Map" function
• Call Map function for : Split and Lift
• Define the Lift measure of intereset
4. Define a "Reduce" function
• Received Key-value
• Merge data within the same key
5. Perform "Association Rules"
• Association rules (Interested association rules)
6. Call MapReduce
• Apply Mappers fuction
• Applay Reducer fuction
7. Visualization
• Results to be presented

Fig. 2. Illustration of the MRLAR steps

TABLE II. ASSOCIATION RULES WITH THEIR SUPPORT, CONFIDENCE, AND LIFT MEASUREMENTS WITH ONE ATTRIBUTE TEST (ARRIVAL DELAY OR DEPARTURE DELAY)

a-	Arrival Delay Cases (in minutes)				
ID	Rule	Itemsets	Support	Confidence	Lift
1	> 60	2946	0.045	0.093	22.3
2	> 120	865	0.013	0.027	75.96
b-	Departure Delay Cases (in minutes)				
ID	Rule	Itemsets	Support	Confidence	Lift
3	> 30	5684	0.087	0.212	11.5586
4	> 40	4260	0.065	0.159	15.4223
5	> 90	1358	0.021	0.051	48.3792

By comparing both approaches single-dimension and multi-dimension association rules, Lift showed the ability to discriminate negative (< 1) and positive (> 1) relationships in the dataset when multi-dimension was applied. But in the other case with single-dimension, Lift always showed high positive values which did not presents a useful knowledge interpretation.

VI. CONCLUSION

In this paper the authors developed a parallel association rule mining algorithm based on MapReduce paradigm by using Lift interestingness measured (MRLAR). The use of the MapReduce approach provided a powerful process over vast amounts of data utilizing parallel approach. Another measurable benefit that the MRLAR algorithm added its capability to directly extract association rule and type of correlation without the need to calculate confidence values, hence eliminating the need of additional calculations.

TABLE III. ASSOCIATION RULES WITH THEIR SUPPORT, CONFIDENCE, AND LIFT MEASURES WITH TWO ATTRIBUTES TEST (ARRIVAL DELAY, DEPARTURE DELAY AND MONTH OF THE YEAR)

a-	Arrival Delay and October Cases				
ID	Rule	Itemsets	Support	Confidence	Lift
1	> 60 & Oct.	182	0.0028	0.0618	0.6678
2	> 120 & Oct.	54	0.0008	0.0624	0.6748
b-	Departure Delay and October Cases				
ID	Rule	Itemsets	Support	Confidence	Lift
3	> 30 & Oct.	376	0.0057	0.0662	0.715
4	> 40 & Oct.	287	0.0044	0.0674	0.7282
5	> 90 & Oct.	87	0.0013	0.0641	0.6925
c-	Arrival Delay and June Cases				
ID	Rule	Itemsets	Support	Confidence	Lift
1	> 60 & Jun.	266	0.004	0.0903	1.1883
2	> 120 & Jun.	99	0.0015	0.1145	1.5063
d-	Departure Delay and June Cases				
ID	Rule	Itemsets	Support	Confidence	Lift
3	> 30 & Jun.	501	0.0076	0.0881	1.16
4	> 40 & Jun.	382	0.0058	0.0897	1.1802
5	> 90 & Jun.	141	0.0021	0.1038	1.3665

The experimental results presented in this paper show that MRLAR performs effectively the detection of associations between itemsets, through integrating the uses of MapReduce and Lift interestingness measured instead of using confidence to determine the correlation between LHS and RHS in association rule.

The proposed approach MRLAR showed a high ability to run under several attributes (multi-dimension association rules), also, Lift interestingness measure successfully were able to determine the type of correlation between itemsets (positively >1, negatively <1, and independent =1) between LHS and RHS instead of using confidence measure in a parallel association rules to help the user to make an important decisions making through determining and explain those patterns in a meaningful knowledge format.

VII. FUTURE WORK

Although there are many benefits of MapReduce but it also have some limitations which cannot be passed over. Some of these limitations which are also specific to MRLAR such as an extracting the association rule with single dimension as shown in the results, also another limitation which is based on MapReduce approach itself that its operates only on data structures of type (key, value) pairs, so that all the input datasets must be adapted into such structure.

Another part of the future work is to develop a parallel data reduction techniques by using singular value decomposition (SVD) and semi discrete decomposition (SDD), or using any other data reduction techniques to remove the unnecessary data with better time based on MapReduce approach. This algorithm lies in the ability to provide pre-processing techniques to reduce the dimensionality of the dataset which reduces the data capacity, for reducing costs. Hence, MapReduce dimensionality reduction algorithm by SVD may will handle another challenge of big data to avoid data dimensionality problems in parallel approach. Therefore, reduce the amount of time and memory required by data mining algorithms, easy visualization of data, and eliminate irrelevant features and noise reduction.

ACKNOWLEDGMENT

This Work is partially supported by Grant of SGS No. SP2016/97, VŠB - Technical University of Ostrava, Czech Republic.

REFERENCES

[1] Oweis, N. E., Owais, S. S., George, W., Suliman, M. G., & Snášel, V. “A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses)”. In Intelligent Data Analysis and Applications Springer International Publishing, pp. 109-119, 2015.

[2] Fouad, M. M., Oweis, N. E., Gaber, T., Ahmed, M., & Snasel, V. “Data Mining and Fusion Techniques for WSNs as a Source of the Big Data”. Procedia Computer Science, 65, ISSN 1877-0509, pp. 778-786, 2015.

[3] Ding, Guoru, Qihui Wu, Jinlong Wang, and Yu-Dong Yao. "Big Spectrum Data: The New Resource for Cognitive Wireless Networking.", arXiv preprint arXiv: 1404.6508, 2014.

[4] Larose, D. T. “Discovering knowledge in data: an introduction to data mining”. John Wiley & Sons, 2014.

[5] Saed Sayad, Data Mining Map, An Introduction to Data Mining, http://www.saedsayad.com/. (Last seen 1–Feb–2016).

[6] Aggarwal, Charu C., and Jiawei Han, eds. “Frequent Pattern Mining”. Springer, 2014.

[7] Aggarwal, C. C. “Association Pattern Mining: Advanced Concepts”. In Data Mining, Springer International Publishing, pp. 135-152, 2015.

[8] McNicholas, P. D., Murphy, T. B., & O’Regan, M. Standardising the lift of an association rule. Computational Statistics & Data Analysis, 52(10), pp. 4712-4721, 2008.

[9] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. “Data mining with big data.Knowledge and Data Engineering”, IEEE Transactions on, 26(1), pp. 97-107, 2014.

[10] Dhanshetti, A., & Rane, T. “A Survey on Efficient Big Data Clustering using MapReduce”. Data Mining and Knowledge Engineering, 7(2), pp. 47-50, 2015.

[11] Menon, S. P., & Hegde, N. P. “A survey of tools and applications in big data”. In Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference, IEEE, pp. 1-7, 2015.

[12] Zaki, M. J., & Meira Jr, W. “Data Mining and Analysis: Fundamental Concepts and Algorithms”. Cambridge University Press, 2014.

[13] Soysal, Ö. M. “Association rule mining with mostly associated sequential patterns”. Expert Systems with applications, 42(5), pp. 2582-2592, 2015.

[14] Chen, X. M., Wang, C. Y., & Cao, H. “Association Rules Mining Based on Minimal Generator of Frequent Closed Itemset”. In Ecosystem Assessment and Fuzzy Systems Management, Springer International Publishing, pp. 275-282, 2014.

[15] Makani, Z., Arora, S., & Kanikar, P. “A Parallel Approach to Combined Association Rule Mining. International Journal of Computer Applications”, 62(15), pp. 7-13, 2013.

[16] Hussein, N., Alashqur, A., & Sowan, B. “Using the interestingness measure lift to generate association rules. Journal of Advanced Computer Science & Technology”, 4(1), pp. 156-162, 2015

[17] Gancheva, V. “Market Basket Analysis of Beauty Products (Doctoral dissertation”, Thesis on Erasmus University Rotterdam), 2013.

[18] McNicholas, P. D., Murphy, T. B., & O’Regan, M. Standardising the lift of an association rule. Computational Statistics & Data Analysis, 52(10), pp. 4712-4721, 2008.

[19] Makani, Z., Arora, S., & Kanikar, P. “A Parallel Approach to Combined Association Rule Mining. International Journal of Computer Applications”, 62(15), pp. 7-13, 2013.

[20] Agrawal, R., & Srikant, R. “Fast algorithms for mining association rules”. In Proc. 20th int. conf. very large data bases, VLDB Vol. 1215, pp. 487-499, 1994.

[21] Srinivasa, S., & Bhatnagar, V. (Eds.), “Big Data Analytics”, First International Conference, BDA 2012, New Delhi, India, December 24-26, 2012: Proceedings (Vol. 7678). Springer, 2012.

[22] Kudyba, S. “Big Data, Mining, and Analytics: Components of Strategic Decision Making”, CRC Press, 2014.

[23] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.", Information Sciences, vol. 275, pp. 314-347, 2014.

[24] Dean, J., & Ghemawat, S. “MapReduce: simplified data processing on large clusters”. Communications of the ACM, 51(1), pp. 107-113, 2008.

[25] Koch, C. “Compilation and synthesis in big data analytics”. In Big Data. Springer Berlin Heidelberg, pp. 6-6, 2013.

[26] MathWorks Documentation, MapReduce http://www.mathworks.com/help/matlab/ref/mapreducer.html?searchHighlight=mapreduce%20and%20matlab, (Last seen 1–Feb–2016).

[27] Fang, H., Zhang, Z., Wang, C. J., Daneshmand, M., Wang, C., & Wang, H. “A survey of big data research”. IEEE network, 29(5), 6, 2015.

[28] Triguero, I., Peralta, D., Bacardit, J., García, S., & Herrera, F. “MRPR: A MapReduce solution for prototype reduction in big data classification neuro computing”, 150, pp. 331-345, 2015.

[29] Wang, B., Huang, S., Qiu, J., Liu, Y., & Wang, G. “Parallel online sequential extreme learning machine based on MapReduce Neuro computing”, 149, pp. 224-232, 2015.

[30] Giakoumakis, P., Chrysos, G., Dollas, A., & Papaefstathiou, I. “Acceleration of Data Streaming Classification using Reconfigurable Technology”. In Applied Reconfigurable Computing, Springer International Publishing, pp. 357-364, 2015.

[31] Bayramli, B. SVD Factorization for Tall-and-Fat Matrices on Map/Reduce Architectures. arXiv preprint arXiv:1310.4664, 2013

[32] Sangavi, S., Vanmathi, A., Gayathri, R., Raju, R., Paul, P. V., & Dhavachelvan, P. An Enhanced DACHE Model for the MapReduce Environment. Procedia Computer Science, 50, 579-584, 2015

[33] Wang, K., Tang, L., Han, J., & Liu, J.. Top down fp-growth for association rule mining (pp. 334-340). Springer Berlin Heidelberg, 2002

[34] Yang, X. Y., Liu, Z., & Fu, Y. “MapReduce as a programming model for association rules algorithm on Hadoop”. In Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference, IEEE, pp. 99-102, 2010.

[35] Woo, J. “Apriori-Map/Reduce Algorithm”. In The 2012 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2012), Las Vegas, 2012.

[36] Riondato, M., DeBrabant, J. A., Fonseca, R., & Upfal, E. “PARMA: a parallel randomized algorithm for approximate association rules mining in MapReduce”. In Proceedings of the 21st ACM international conference on Information and knowledge management ACM. pp. 85-94, October 2012.

[37] Lin, X. “MR-Apriori: Association Rules algorithm based on MapReduce”. In Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference, IEEE, pp. 141-144, 2014.

[38] Zhou, X., & Huang, Y.. “An improved parallel association rules algorithm based on MapReduce framework for big data”. In Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference, IEEE, pp. 284-288, 2014.

[39] Wang, B., Huang, S., Qiu, J., Liu, Y., & Wang, G. “Parallel online sequential extreme learning machine based on MapReduce Neuro computing”, 149, pp. 224-232, 2015.