ABSTRACT

Data mining is defined as a way of gaining knowledge from raw data. Huge quantity of data are captured and stored due to World Wide Web and associated information services and it is necessary to come up with techniques to collect tremendous data. In the current scenario, Knowledge Discovery in Databases from distributed databases with minimum communication cost is a thrust area in the field of computer science technology. Next innovative thoughts are moved on to the algorithms which work higher on distributed data. Map Reduce is an emerging programming model for distributed atmosphere, that is the key model for hadoop framework and it enables resilient, distributed process of huge datasets. This research work explores the apriori algorithms in an exceedingly distributed, hadoop environment and formulates a novel hybrid algorithm which incorporates integrity and security of data.

The objectives of this research work are to design and implement weighted and hash tree apriori algorithms in a distributed environment in order to reduce time and memory complexity in execution, to effectively mine the data with a novel hybrid algorithm and to address the integrity and security issues of distributed data. The novel algorithm is developed by integrating the weighted and hash tree approaches. The novel algorithm is proposed by using tree based apriori with correlation, reachability identification and relevancy computation.
The redundant generation of sub itemsets throughout the pruning stage of the frequent itemset mining can be avoided with the correlation between a set of nodes. Security and data integrity issues in distributed nature databases, are handled in the hybrid algorithm with appropriate techniques. Hash based Message Authentication Code (HMAC) is applied so as to prove data integrity.

Data is distributed in Hadoop Distributed File System (HDFS) using MapReduce framework, whereas the algorithm is defined as map tasks to be executed on the distributed data which is available on various clusters of hadoop nodes. Then the results are combined using reduce task. Data is encrypted before distributing as well as decryption is done after receipt of the results. HMAC (Hash based Message Authentication Code) is generated before sending and moreover after receiving data so as to confirm that no data is missed or manipulated. Weighted apriori uses a link based measures for calculating weight of items between transactions. Hash tree apriori uses a tree approach to reduce number of scans in the iterative pruning. The objective of implementing weighted apriori algorithm as well as hash tree apriori algorithm individually in distributed environment paves the way for the choice of MapReduce.

The MapReduce system orchestrates the different tasks by simultaneously takes care of data transmissions and ensures fault tolerance. Hadoop MapReduce is a processing model for distributed nature data and execution environment which works on large clusters of machines. In the proposed work, experimental data set is given to a Hadoop Distributed File
System and Map Reduce processing is started using weighted apriori and hash tree apriori algorithms. This research work designs the weighted apriori algorithm as mapper tasks in eclipse environment and dataset is maintained in Hadoop Distributed File System DataNodes and jobs are executed using JobTracker. Same procedure is also applied with hash tree apriori algorithm as map and reduce functions. The dataset has been distributed to various clusters and Map task is applied on HDFS to perform algorithm execution and the results are consolidated by the reduce task. VMware environment is employed for the experimental setup which simulates a cloud. This research work primarily focusses on working of association rule mining on a hadoop environment and analyses how it works.

This research work proposes a unique Hybrid Weighted-HashT Apriori algorithm which is applied on HDFS clusters as map and reduce tasks. Initially, Number of HDFS clusters to be partitioned is obtained by user choice. Then similarity is calculated between transactions of dataset and the particular cluster for each transaction will be determined. The data will be moved onto HDFS clusters. After that, HDFS takes over the control and splits the data to various nodes within the cluster and map as well as reduce tasks are executed with the algorithm coded in java. In this research, the map and reduce functions are working in a distributed environment with tree based mining approach so as to scale back the consumption of the memory usage in the process of association rule prediction.
The interpretation is, hybrid apriori algorithm works better, compared to weighted and hash tree apriori algorithms in a distributed environment. Hybrid apriori algorithm is tried with 4, 8 and 12 clusters respectively. The hybrid algorithm is reformed to fit in the data integrity and data security issues so as to facilitate the distributed environment. The novel hybrid apriori algorithm outperforms weighted apriori and hash tree apriori algorithms in terms of time and memory. Another important factor to be proved here is, the amalgamation of security solution into hybrid algorithm which reflects in execution time but does not affect overall performance. Addressing security is considered as a necessity though it consumes time.

The research work designed and implemented weighted and hash tree apriori algorithms as MapReduce tasks in HDFS clusters. A hybrid novel algorithm is formulated by combining weighted and hash tree apriori algorithms and implemented in a Eucalyptus cloud hadoop platform. The work proves the performance of novel hybrid algorithm without, with hadoop and eucalyptus platforms through the experimental readings. The result shows that, the time is considerably reduced in hadoop and eucalyptus platform. The novel algorithm which is discussed in this research work gives a better direction in the field of cloud data mining. The integrity and security assurance are also addressed in the novel hybrid algorithm. The results of various data sets are also taken into analysis so as to prove the integrity and security assurance.