Abstract - All the algorithms of Data Mining evaluate the desired interesting patterns from the transactional database where the data is precise. But in order to extract data from Big Data i.e. huge transactional database is still a problem. A large dataset applications require outlier detection methods for detecting and analyzing accident through wireless sensor networks (WSNs) which have gained achievement in traffic detection and avoiding road congestion. In this paper we propose a fast parallel outlier detection i.e a MapReduce paradigm for parallel programming which offers traffic balancing. MR-AVF(MapReduce-Average Value Frequency) is considered highly scalable with respect to the number of nodes.

Keywords – Big Data; MapReduceModel; programming skill for Big Data mining; Big Data analysis; Searching and mining Big Data; Frequent Pattern; Constraints; Uncertain data; Dataset(WSNs).

I. INTRODUCTION

Data Mining is a domain of Computer Science Engineering where data to be extracted is useful data from a huge database is becoming problematic. One of the important Data Mining technique from uncovering data from large transactional database is outlier detection. Data Mining help online shop owners to know purchasing behavior of a customer. The concept BigData which is getting important and is called as a big game changer is mainly due to vast amount of data generated by social media platform. So BigData is nothing but interesting high volume, high value and/or high variety data with volumes beyond the ability of commonly used software.

Detecting outliers in data such as accidental data of wireless sensor networks (WSNs) for controlling congestion in roadside ways. Year by years use of vehicles has been tremendously increased.

A. Dataset

There are two types of dataset called precise dataset and uncertain dataset. In uncertain dataset each item in a transaction is defined by its probability. While in precise dataset, each item is defined by its unique identity in the form of an item name.

II. PREVIOUS WORK

In previous work [2] we focussed all the existing data mining algorithms for searching information from transactional databases of precise data.

But the space for searching and mining from uncertain data is greater because of probability called existential probability. This problem is worst if we go in the era of BigData. Also users are interested in tiny portion of large search space in BigData Mining. So by providing opportunities for users to express their interest action but the data mining algorithm returns number of patterns out of which only few are of our interest. So we proposed algorithms as:-

(i) 1st algorithms allows users to express their interest in terms of constraints.

(ii) 2nd algorithms uses the MapReduce model to mine uncertain BigData for frequent pattern for satisfying the user specified constraints.

For using the MapReduce model many algorithms have been proposed to mine information from a large space. But an important BigData Mining and analytics task is frequent pattern which get the frequently coming items with consideration of parallel and distributed computing [8] on large grids of nodes. As the name suggests MapReduce involves two keys functions: “Map” and “Reduce”. Mining BigData encounter concept drift problem where statistical properties of the attribute and their target classes shift over time. For example people preferences for product change over time. i OVFDT has a better performance for synthetic and real world concept drift datastreams.

Author in this paper [6] also proposed to extract useful data in real time. IT world is coping with BigData problems. We present implementation details and performance results for “online” Association Rule Mining (ARM) for Big and fast data streaming. Actually we added Apriori and two different FP – growth algorithms inside Esper Complex Event Processing(CEP) engine and compared their performances using last FM social music site data. Our most important findings shows that online ARM gives:-
(a) More unique rules
(b) more throughput
(c) Lower latency

We hope our findings can help on the design and implementation of other fast data analytics.

III. PROPOSED WORK

A parallel programming paradigm called MapReduce which was actually launched by google whose task is to ease the processing of large datasets on simple clusters of nodes called computers. These clusters always contain hundreds or thousands of nodes that first store datasets and then use the datasets for processing as per requirement in a distributed environment. Typically a single master node is used to schedule the storage and computation on the nodes facilities as a master server. Google File System (GFS) [12],[13] was the original MapReduce System which was efficient to tolerate often node failures due to replication of data.

In our work we entirely dependent on Map and Reduce functions. The Map function takes inputs in the form of key-value pair K1 and V1 .Map functions also define the job of these keys and values. The Map functions returns another set of keys and values like K2 and V2. The Reduce function sorts the key value pairs by K2. The related values of all V2 are then reduced to V3 values.

\[ \text{Map}(k1,v1) \rightarrow (K2,V2) \] ; \[ \text{Reduce}(K2,V2) \rightarrow (K2,V3) \].

IV. CONCLUSION

Big Data mining is the data mining techniques to find users pattern from huge amount of data. MapReduce-AVF is simple and easy to develop which ensures load balancing and fault tolerance. The map function takes total of 448 attributes called different keys K1 with its different values V2 and Reduce function reduces all the key values into some values which has been shown in snapshot of output.

ACKNOWLEDGEMENT

I thank (Leung et al.,2014,Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data) and (Karolien Geurts,Traffic Accidents Dataset) for assistance with [MapReduce Technique , Dataset , User , Uncertain Data , Constraint , Big Data Processing , Frequent Pattern].I also thank my Guide (L.H Patil,Aassist. Prof. ,PIET) and co.Guide(U.K Thakur,Aassist prof.,PIET) for their comments that greatly improved the manuscript.

REFERENCES

[1] Carson Kai-Sang Leung, Christopher L. Carmichael “Efficient Mining of Frequent Patterns from Uncertain Data” Seventh IEEE International Conference on Data Mining – Workshops DOI 10.1109/ICDMW.2007 IEEE
Figure 1: The flow of data in a MapReduce/GFS architecture for file storage and MapReduce operations. Dashed lines indicate control messages, and solid lines indicate data transfer.

Figure 2: The snapshot of applying MapReduce Model on accident data.