OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

FREQUENT ITEMSET MINING USING DISTRIBUTED FRAMEWORK

Sonal Rajabhau Londhe[1], Prof. Varsha R. Dange[2]
Student, Dept of Computer, Dhole Patil College of Engineering, Savitribai Phule Pune University, Pune, India[1]
Guide, Dept of Computer, Dhole Patil College of Engineering, Savitribai Phule Pune University, Pune, India[2]

Abstract: **Data mining is that the extraction of hidden prognostic data from massive databases, that could be a powerful new technology with nice potential to assist firms likewise as analysis specialise in the foremost necessary data in their information warehouses. Data processing tools predict future trends and behaviours, permitting businesses to create proactive, knowledge-driven choices. Frequent Itemset Mining is one among the classical data processing issues in most of the information mining applications. It needs terribly massive computations and I/O traffic capability. Resources like single processor's memory and processor area unit terribly restricted, that degrades the performance of algorithmic program. During this paper we've planned associate degree algorithmic program which can run on Hadoop – one among the recent hottest distributed frameworks that chiefly specialise in Mapreduce paradigm. The planned approach takes into consideration inherent characteristics of the Apriori algorithmic program associated with the frequent itemset generation and thru a block-based partitioning uses a dynamic work management. The algorithmic program greatly enhances the performance and achieves fault tolerance compared to the present distributed Apriori based mostly approaches. Planned algorithmic program is enforced and tested over a multinode cluster.**

Keywords: Frequent Itemset Mining, Mapreduce, HDFS

I INTRODUCTION

Information mining is that the extraction of hidden prognostic facts from massive databases, that is a effective new era with excellent potential to help firms likewise as analysis specialise in the major necessary facts in their information warehouses. information processing gear predict destiny developments and behaviors, permitting groups to create proactive, expertise-driven alternatives. common Itemset Mining is one some of the classical facts processing issues in most of the data mining applications. It needs terribly huge computations and that I/O visitors functionality. assets like unmarried processor's memory and processor place unit extraordinarily confined, that degrades the overall performance of algorithmic program. in the course of this paper we've planned accomplice degree algorithmic application which could run on Hadoop – one some of the recent most up to date dispensed frameworks that mainly focus on Mapreduce paradigm. The planned approach takes into consideration inherent characteristics of the Apriori algorithmic program associated with the frequent itemset era and via a block-based partitioning uses a dynamic paintings control. The algorithmic application substantially complements the overall performance and achieves fault tolerance compared to the present distributed Apriori based totally ordinarily processes. deliberate algorithmic program is enforced and examined over a multinode cluster.

II RELATED WORK

The authors of "affiliation Rule mining extracting common itemsets from the large database" have presented a trouble of locating the common items from immoderate database. The authors have advanced the rules that have minimum transactional assist and minimal confidence. For this an set of rules is used that carefully estimates the itemsets for one pass. It adjusts the information among the number of passes and itemsets which can be measured in a skip. This process makes use of pruning system for avoiding sure itemsets. therefore this gives specific common itemsets from excessive databases [1]. range of parallelization methods is used to growth the overall performance of Apriori-like algorithms to discover common itemsets. MapReduce has not only created but additionally exceeds inside the mining of datasets of gigabyte scale or more in either homogeneous or heterogeneous corporations. Mapping management of Mapreduce technique will minimize the overhead of every Mapreduce segment. The authors have implemented three algorithms, DPC, FPC, and SPC [2]. The authors have supplied a balanced parallel FP-increase set of rules BPFP

[3], a revised model of PFP set of rules. FP- increase set of rules is applied with the MapReduce approach named as Parallel FP-boom set of rules. BPFP balances the weight in PFP, which enhances parallelization and mechanically complements execution. BPFP offers an awesome performance by way of utilizing PFP's grouping system [3].

FIUT shows a brand new technique for mining common itemsets known as as common Itemset Ultrametric Tree (FIUT).It's far a sequential set of rules. It consists important degrees to scan the database. First stage calculates the aid be counted for all itemsets in a large database. 2nd degree uses pruning approach and gives most effective common itemsets. While calculating frequent one itemsets, degree will construct small ultrametric bushes. Those effects may be proven with the aid of building small ultrametric trees [4].

Dist-Eclat, BigFIM are two FIM algorithms used with MapReduce Framework. Dist-Eclat focuses on velocity by way of load balancing procedure the use of k-FIS. BigFIM concentrates on hybrid technique for mining immoderate information[5]. Apriori set of rules is additionally used to create k th FIS itemsets. The k th FIS is used to search common itemsets primarily based at the Eclat device. those three algorithms are used with round robin technique which achieves a better statistics distribution [5].

PARMA makes use of parallel mining approach with the blessings of Randomization for extracting frequent itemsets from good sized amount of databases. This divides its functionality into two components, first off it gathers the arbitrary facts samples and secondly it uses parallel computing approach this is utilized to growth the mining velocity. This technique avoids the replication that is very steeply-priced. this is done by using making quantity of little arbitrary segments of the transactions. After that a mining set of rules applies on each section individually. It applies parallel technique via utilizing MapReduce programming paradigm[6]. in the processing of k -Nearest Neighbor Joins makes use of MapReduce and distributes the immoderate statistics at the quantity of machines. this is carried out via the mappers and the reducers supply the outcomes in terms of the KNN be part of. KNN join is the important thing aspect to search the kth-nearest neighbor. Mapreduce is applied for powerful computing the statistics to achieve the great overall performance end result [7]. To diagnose the Heterogeneous Hadoop Cluster and to search number one faults, this paper is used Hadoop schedulers to provide green Hadoop clusters although they're in heterogeneous clusters [8]. to differentiate and extract frequent and rare itemsets from the huge database, phase of scanning may be accomplished right here. In first scan it accepts enter and distributes it into mappers and finds out infrequent itemsets the usage of minimum aid. The reducer gives mixed end result and sends it to 2d section. in this phase it scans first segment output and offers the final end result [9].

FIUT is used with MapReduce to discover frequent itemsets. MapReduce is a popular programming technique used to compting massive datasets [10]. It divides into three MapReduce stages. the primary mapreduce section reveals out common one itemset. here the database is divided into number of enter documents and given to each mapper. second MapReduce segment scans the frequent one itemsets and prunes the infrequent itemsets. This phase generates k-common itemsets . 0.33 mapreduce segment uses FIUT algorithm. This section decomposes the itemsets after which it'll create ultrametric tree[10].

III PROBLEM STATEMENT

The proposed work first investigate problem of Frequent Itemset on conventional systems. The system proposes a new data partitioning method to well balance computing load among the cluster nodes. System also focuses on database security like SQL injection and find efficient way for security as well execution in HDFS framework.

IV SYSTEM ARCHITECTURE

In Proposed System a new data partitioning method to well balance computing load among the cluster nodes is developed.

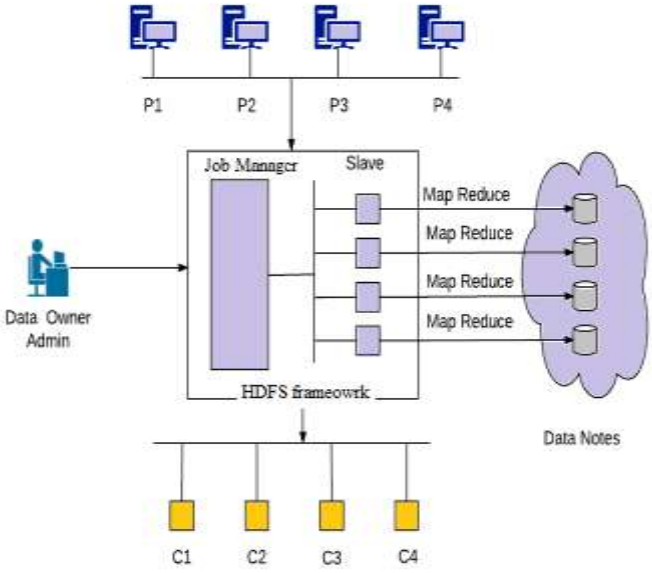

Figure 1: Proposed System Architecture

Proposed Algorithm

Algorithm for SQL injection and Prevention

Phase : 1

1: Procedure SPMA(Query, SPL[ ])
INPUT: Query=User Generated Query
SPL[ ]=Static Pattern List with m Anomaly Pattern
2: For j = 0 to m do
3: If (AC (Query, String.Length(Query), SPL[j][0]) = =0)then
4: Calc anomaly score
5: If ( ) Score Value Anomaly = Threshold
6: then
7: Return Alarm .. Administrator
8: Else
9: Return Query .. Accepted
10: End If
11: Else

12: Return Query .. Rejected
13: End If
14: End For
End Procedure

V EXPERIMENTAL RESULTS

A Multinode cluster is setup for evaluation purpose. The performance of the proposed system against the Frequent Itemset Ultrametric Tree (FIUT) method is tested. The system begins by first uploading the file from HDFS.

The frequent itemset process needs to be provided a minimum support with a start and end date. Fig. 2 demonstrates this process.

Figure 2: Frequent Itemsets Generation Process

The performance of this system against the Frequent Itemset Ultrametric Tree (FIUT) method is as shown in fig 3. Modified Apriori algorithm is a good algorithm to give the correct results as compared to existing systems.

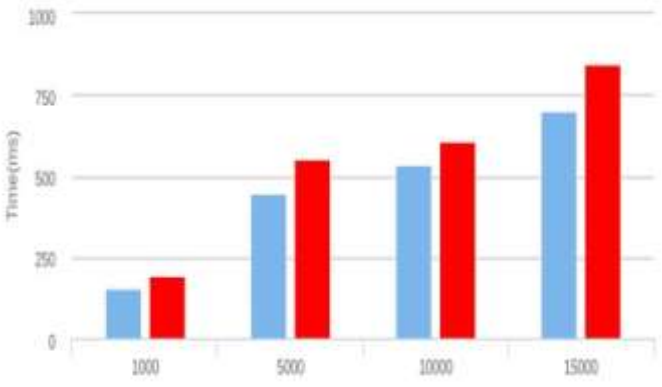

Figure 3: Execution Times of FIUT and Proposed Algorithm (MA)

The SQL Injection and Prevention Algorithm detects any malicious query and raises an error.

A MultiNode Cluster is setup for evaluation of the proposed system. The proposed system was first tested on this cluster. The replication factor is set to 2. Even if one of the slaves goes down, the system is still able to acquire the input and process it using the other Live Node. The system was then run with 1 Dead node to check for fault tolerance. The proposed system acquires the input from HDFS and processes it using this Live Node to produce the desired outputs.

VI CONCLUSION

Frequent Itemset Mining has attracted plenty of interest but much less attention has been given to mining infrequent Itemsets. The mining of common itemset is an vital research area in the field of facts mining. The association guidelines are formed using the common itemset mined. Many unique strategies had been proposed for mining the frequent itemsets.

This device proposes an set of rules so that you can run on Hadoop – one of the latest most popular disbursed frameworks which specially recognition on MapReduce paradigm. The proposed approach takes into consideration inherent traits of the Apriori set of rules associated with the frequent itemset era and via a block-based partitioning makes use of a dynamic workload management. The algorithm is applied and tested on a Multi-Node Cluster. The study revealed that the algorithm considerably enhances the computing overall performance. Fault tolerance has also been verified by effectively executing the proposed machine within the presence of one useless Node.

REFERENCES

[1] JW.Han, J.Pei and YW.Yin, ―Mining Frequent Patterns without Candidate Generation‖, International Conference on Management of Data, vol. 29(2), 2000, pp. 1-12.

[2] H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Chang, ―PFP: Parallel FP-growth for query recommendation‖, Proceedings of the 2008 ACM Conference on Recommender Systems, 2008, pp. 107-114.

[3] Zhigang Zhang, Genlin Ji, Mengmeng Tang, ―MREclat: an Algorithm for Parallel Mining Frequent Itemsets‖, 2013 International Conference on Advanced Cloud and Big Data.

[4] Hui Chen, Tsau Young Lin, Zhibing Zhang and Jie Zhong, "Parallel Mining Frequent Patterns over Big Transactional Data in Extended MapReduce‖, 2013 IEEE International Conference on Granular Computing.

[5] Xinhao Zhou, Yongfeng Huang, "An Improved Parallel Association Rules Algorithm Based on MapReduce Framework for Big Data‖, 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery.

[6]Jinggui Liao, Yuelong Zhao, Saiqin Long, ―MRPrePost-A parallel algorithm adapted for mining big data‖, 2014 IEEE Workshop on Electronics, Computer and Applications.

[7] Sheela Gole, Bharat Tidke, ― Frequent Itemset Mining for Big Data in social media using ClustBigFIM algorithm‖, International Conference on Pervasive Computing.

[8] Siddique Ibrahim S P, Priyanka R, ―A Survey on Infrequent Weighted Itemset Mining Approaches‖ , 2015, IJARCET, Vol.4, pp. 199-203.

[9] Surendar Natarajan, Sountharrajan Sehar, ―Distributed FP-ARMH Algorithm in Hadoop Map Reduce Framework‖, 2013 IEEE.

[10] Xiaoting Wei, Yunlong Ma , Feng Zhang, Min Liu, Weiming Shen, Incremental FP-Growth Mining Strategy for Dynamic Threshold Value and Database Based on MapReduce‖, Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design.