Frequent Itemset Mining Using Amended K-Nn Technique Onhetrogeneous Hadoop Clusters

V. Seethalakshmi, V. Govindasamy, V. Akila

Abstract: ***With the improvement of Data Innovation, there is an exponential development in the amount of information that is produced and utilized in the recent years. The necessityof data storage and retrieval ofinformation in order to derive value from the inaccurateinformation has cleared route for parallel, dispersedfunctions like Hadoop. The existingHadoopexecution presumes that the power capacity of all the nodes in a group ishomogeneous, but the cloud framework is composed of different hardware systems. Further, the cloud has distinctive equipment arrangement frameworks. Hence, it is required to modify the data placement strategy, so that the data is organized depending on the handling power of a node. Consequently, it is important to redesign the data placement strategy in Hadoop, where we can spread out the information depending upon the computing capacity of a node. In this paper, we are proposing a dynamic block placement policy in Hadoop to distribute the input information blocks among the various nodes depending on the processing power of every node. The proposed technique is named as Amended k-Nearest Neighbor (AMENDED k-NN) technique. The k-NN technique can modify, balance the data dynamically and rearrange the input data in heterogeneous environment, according to the processing power of every node in the Hadoopheterogeneous environment. The proposed data placement strategy can distribute the stored information in the heterogeneous clusterso as toenhance the data-processing capability. Our methodat the point of the Hadoop Distributed File System (HDFS) does not consider the correlations among application data. Experimental results reveal that Amended k-Nearest Neighbor significantly improves the Frequent Itemset Mining (FIM) efficiency of the currentFiDoop-DPresults by 31 percentage with an average of 18 percentage. The proposed Amended k- NearestNeighbor (AMENDED k-NN) technique provides reduced data execution time, redundant transaction and computation cost. We plan to integrate Amended K- Nearest Neighbor with a data-placement system in HDFS on Heterogeneous sets in order to increase additionally the load balancing system which isdone in HDFS.***

Key Terms: ***Amended k-NN, K-nearest neighbor, Frequent Itemset Mining, FiDoop-DP, Clustering, Data Placement Policy.***

I. INTRODUCTION

Big Data is a word which refers to anenormoussize of both structured and unstructured information. In many cases, the quantity of information is extremely enormous. The available information exceeds the current capacity level of storage [8]. Big Data is to help the software companies to improve the process and to make the intelligent decision much faster.

In Big Data era, the requirement for a customizable technique to work with huge data sets at a sensiblespeed becomes a need.Frequent itemset disclosuremethods help in generating qualitative information which gives business knowledge and helps the decision makers. Frequent Itemset Mining (FIM) is a standout amongst the most basic and time-consumingjobs in Association Rule Mining (ARM). ARM is a regularly utilized data mining algorithm. FIM acts as a vital asset by separating the most critical successive examples that happen frequently in a vast exchange database. A common utilization of ARM is the renowned market load investigation. The major purpose of the FIM is to identify the more essential frequent itemsets. FIM techniques can be partitioned into two types:FP-growth and Apriori systems. Apriori system is a typicaltechniquewhich uses the generate-and-test process. It produces a huge count of itemsets. Apriori system has to testawhole database many times. To decrease the time needed for examining the databases, we have the other category of FIM. This is FP-growth scheme. FP-growth scheme avoids the generation of itemsets. Most parallel FIM techniques are based upon the Apriori system. But, in Apriori-like parallel FIM techniques, every processor needs to check the database many times and to compare an inordinate number of candidatesitemsets with another processors.

Therefore, Apriori-like parallel FIM solutions incur manyproblems-large input output and synchronization overhead. These problems make Apriori like parallel FIM solutions strenuous to scale up. The scalability issues have been addressed by the FP-growthlike parallel FIM methods.

In particular, the current parallel techniques does not have a system that empowers automatic parallelization, load balancing, data circulation, reduced time, reduced cost and fault tolerance on huge computing groups. To provide a solution to the above mentionedproblem,we have designed a parallel FIM techniquenamed as Amended K-Nearest Neighbors (K-NN) technique. Compared with the existing FIMtechnique, Amended K-NN technique has distinctive features. In Amended K-NN technique, the Mappers simultaneouslydecompose itemsets. Further, the Reducers achieve combination processes as well as mining.

In our proposed method, we have designed Amended K-NN technique and incorporated that technique into Hadoop Distributed File System (HDFS). The proposed technique is to initially organize data blocks into heterogeneous nodes in a cluster in order to balance the data load.

I. LITERATURE SURVEY

In this literature survey, various mining techniques like FP-growth, Apriori algorithm, Fi-Doop-DP etc., have been discussed. In these techniques, given a huge dataset, data partitioning techniques incur high communication and mining overhead made by repeated transactions

Revised Manuscript Received on March 25, 2019.

V. Seethalakshmi,Research Scholar, Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India
V. Govindasamy,Associate Professor, Department of Information Technology, Pondicherry Engineering College, Puducherry, India
V. Akila,Assistant Professor, Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

transmitted among computing sets.

In the paper [1], a partitioning approach called FiDoop-DP (Data Partitioning in FIM on Hadoop) is addressed. FiDoop-DP uses the MapReduce type. The performance of FiDoop-DP is boosted by parallel FIM on Hadoop clusters. The implementation results reveal that FiDoop-DP is helpful to reduce network and processing loads by the virtue of removing repeated transactions on Hadoop nodes.

The proposed method [2] uses the inherent parallelism which is present in the FP-Growth technique to such an extent that N-Processing entities can work in a collaborative environment. The Processing Entities (PEs) work in an autonomous way in parallel. Communication among the entities will occur at final stage of each iteration. The proposed technique logically divides the Field-Programmable Gate Array (FPGA) into different PEs.In addition, the workload is distributed among them. In addition to parallelizing the method, PE contains the entire part of an item of the tree. A particular PE is capable to extract its frequent patterns independently of the remaining PEs. Theproposed framework is validated by using the benchmark datasets. The result shows that there is a linear roposed system.

Another approach namely Low overhead and Elastic MapReduce (LEMO-MR) explains about huge- scale dispersed record ordering and database working through MapReduce for the information handling application area [5]. The platform is of great use for data intensive application. It can also be designed to process CPU intensive disseminated application. MapReduce has been related with information and volume [6]. MapReduce when executed through multiple layers ofreflections can experience theill efforts of the high complexity.

Real data-driven approach is used to validate the proposed systems. The paper [7] depicts on a joint scheduling technique in MapReduce, where maps and shuffle stagescan be covered and be led in parallel. The main objective of scheduling technique is to decrease the average task makespan. The key test is that the map and rearranges stages cannot be completely parallelized because of their reliance relationship: the rearrange stage may hold up to exchange the information transmitted by the map stage. To evade I/O underutilization, tasks that can form a strong set should be combined shrewdly and implemented. A few online and web based scheduling arrangements are proposed to execute jobs in a pairwise way. Scheduling optimalities are discussed in the context of a few situations.

The paper [8] highlights that Hadoop has developed into a strong stage to modify huge datasets. By modifying the MapReduce information stream, a few methodologies broaden the relevance of Apache Hadoop to support the iterative functions. Upgrading HDFS and methods to deal with make or create query methodologies are moderately normal. The HDFS file framework was created to have high throughput to store extensive datasets.

Paper [9] presents a MapReduce technique and model that fault tolerated. An exploratory assessment of the execution of a task with their methods results in decreased asset usecontrast. Thisis useto normalize Byzantine Fault Tolerance ideal model.

increase in the speedup of the proposed framework. It is suitable only for large set of data. FiDoop uses parallel FIM technique to incorporate the frequent items ultrametric tree or FP-tree rather than conventional FP trees [3]. FiDoop uses the MapReduce programming model. FiDoop achieves reduced storage. FiDoopavoids the necessity to constructconditional pattern bases. FiDoop easily integrates MapReduce tasks to achieve parallel mining of frequent itemsets. Improved efficiency ofFiDoop is achieved by adjusting I/O load across the data nodes of a cluster. The workload balancing metric is used to visualize the increased the load balance achieved among the cluster computing nodes.

Two methods are proposed with different information parallel functions They are parallel matrix multiplication and numerical simulation of lid-driven cavity flow, on a multimode graphics processing unit (GPU) server and on a heterogeneous GPU-exaggerated multimode group [4]. The multi-dimensional efficient scheme with multiple parameters may be needed to achieve the same efficiency of the scheme. For huge and complicated applications, balancing the workload requires additional effort in the p

The proposed Mappers are transferred through an appropriated meta-information store [10]. Mappers have an understanding of their state and they do preserve their fault-tolerance and scalability. Programming Adaptive Mappers dynamically take various information partitions (splits) to amortize Mapper start-up expenses. Adaptive Combiners enhance the nearby collection by keeping up a store of partial sum for the frequent keys. Adaptive Sampling and Partitioning strategies test the Mapper results. They utilize the obtained data to deliver adjusted segments for the reducers.

II. INFERENCE

The literature survey provides the taxonomy of FIM on HeterogeneousHadoop environment based on the various evaluation metrics such as Redundant Transaction, Execution Time, Computation Cost etc. It discusses how different mining techniques and scheduling techniques help in gaining better outcome in a HeterogeneousHadoop environment. Furthermore, the merits and the demerits of various techniques are analyzed. Thus, it is evident that these techniques are important in mining techniques.

In the area of Big Data, MapReduce isused to innovateequivalent data mining techniques, along withFIMtechniques. Existing parallel FIM calculations give a huge dataset information apportioning procedure. Here, the arrangements endure high correspondence. The existing FiDoop-DPtechniqueaddressesinteraction among connectionsso as to divide a big dataset into multiple data nodes in a Hadoop groups.

FiDoop-DP can be able to

- Divide transactions withmore resemblancecollectively and
- Cluster the repeated items into a group.

FiDoop-DP uses the Voronoi structure-based data partitioning method to achieve information partition, wherein Locality Sensitive Hashing (LSH) is included to provide an investigation of association among connections. Locality

Sensitive Hashing verifies all the transactions to find all the most common sets. In the apportioning procedures, MapReduce stage is in its early stages, prompting genuine execution issues. Accordingly, data dividing in FIM influences organizing activity as well as processing loads. Their confirmation demonstrates that data parceling calculations should focus on system and processing loads notwithstanding the issue of load adjusting. Existing data dividing arrangements of FIM in Hadoopattempt to adjust the calculation stack by similarly appropriating information among hubs. In any case, the connection among the data is frequently disregarded which will prompt poor information region.Further, the information rearranging costs and the system overhead will increase.

A.DISADVANTAGES of EXISTING SYSTEM

The main disadvantages of the existing FiDoop-DPtechnique are

1. It requires large memory space as it deals with large number of itemset generation.
2. It performs multiple scans for generating itemsets.
3. Execution time is wasted in producing itemset every time
4. It performs badly with long pattern data sets
5. It results in huge memory consumption.
6. It becomes inefficient when datasets are sparse.
7.

I.PROPOSED SYSTEM

A.AmendedK- Nearest Neighbor (k-NN)Technique

To alleviate more communication problems and to decrease computing cost in MapReduce basedFIM techniques, Amended k-NN is developed. Amended k-NN is designed to exploit the interaction among transactions to divide a huge dataset over by data sets in a Hadoopenvironment. One of the salient features of Amended k-NN is its potential of reducing network traffic and computing work load by decreasing the count of repeated transactions, which are transferred between Hadoopsets. The Amendedk-NNtechnique is a lazy learning technique. The proposed technique is non-parametric. The inherent data distribution is not under-existing data partition. This is a main advantage of k-NN techniquebecause the major part of the data does not followthe assumptions. Amended k-NNtechnique is also a slow technique. This resultsin faster training assumed. Amended k-NNtechnique can keep all the training data. Amended k-NNtechnique makes decisions whichare dependent on the complete training set. The k-NN Techniquecan be used in Nearest Neighbor based, Gene Expressions, Content Retrieval,3-D Structure predictions and Protein-Protein interaction.

A.Pseudo Code of the Proposed Algorithm

The pseudo-code of the AK-NN algorithm is given below:

Input: (i)The Training Set ‘x’ with the tag designs $\{x_i, i = 1,2,3,..n\}$,
(ii) The Test Set ‘y’.

Output: (i)Class Tag of ‘y’.
(ii) Confidence $C_i$for each Class Tag.

Algorithm:

For i = 1 to n
Compute the distance from x_i to y using the Distance Metric.
If $i \leq k$ where k is centroid point

Include x_i in the set of k- NN.
Else if (x_i is closer to y_i than any previous nearest neighbors)
Remove the furthest of the k-NN.
Include x_i in the set of k-NN.
End If
End For
For j =1 to C
Calculate the class tags
End for
Class tag of ‘y’ is assigned to the class with which it has the highest relationship value
End function

B.The Major Features of Amended k-NN Technique:

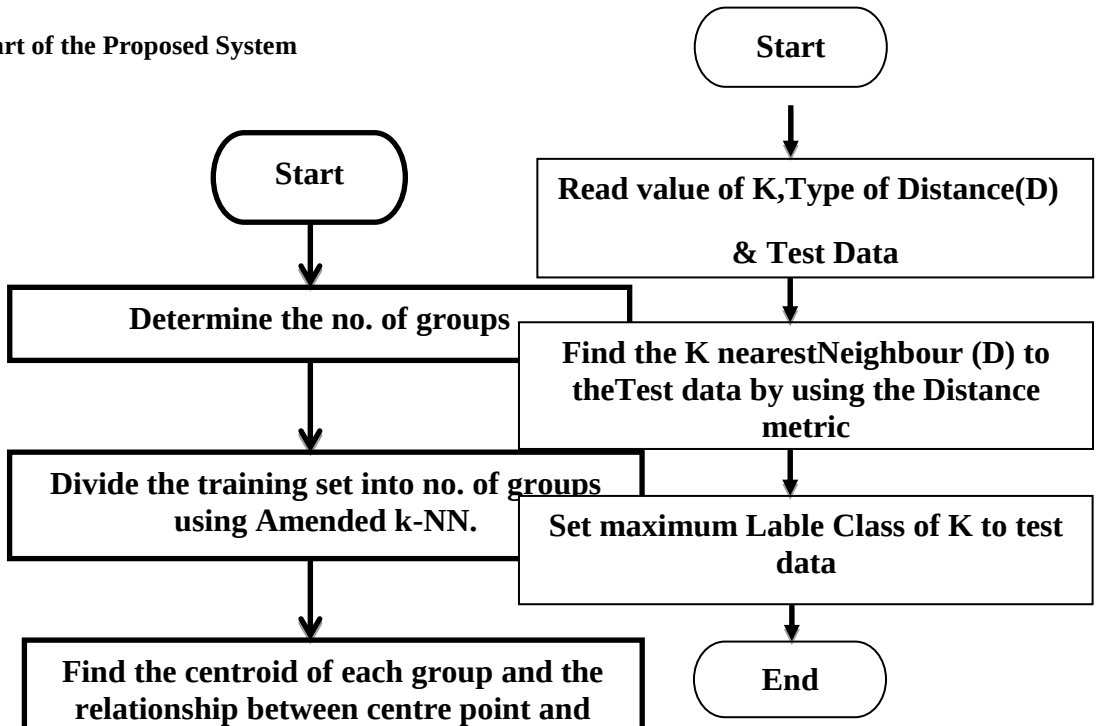
- Amended k-NNtechnique claims that the information is in a feature distance. To be specific, the data sets are in a dimensional set.
- The information can be perhaps even in multi-dimensional directions. Since the sets are in feature distance, they have a view of space. This requirementmay not be needed for Euclidean Distance even it ismostlyutilized scheme.
- Amended k-NN techniqueutilizes training data to categorize the new data setstogether. This training data is called testing dataset.
- The training informationcontains a set of vectors and a class tagconnected with every vector. In the easier situation, it will be either + or –. But,Amended k-NNtechnique can work similarly well with arbitrary amount of classes
- This given number 'k' indicates number of neighbors. The 'k' neighbors are characterized dependent on the distance metric impact the characterization. This is generally an odd number if the quantity of classes is 2. If k=1, then the system is basically termed as the nearest neighbor technique.

Retrieval Number: F1003376S19/19©BEIESP
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

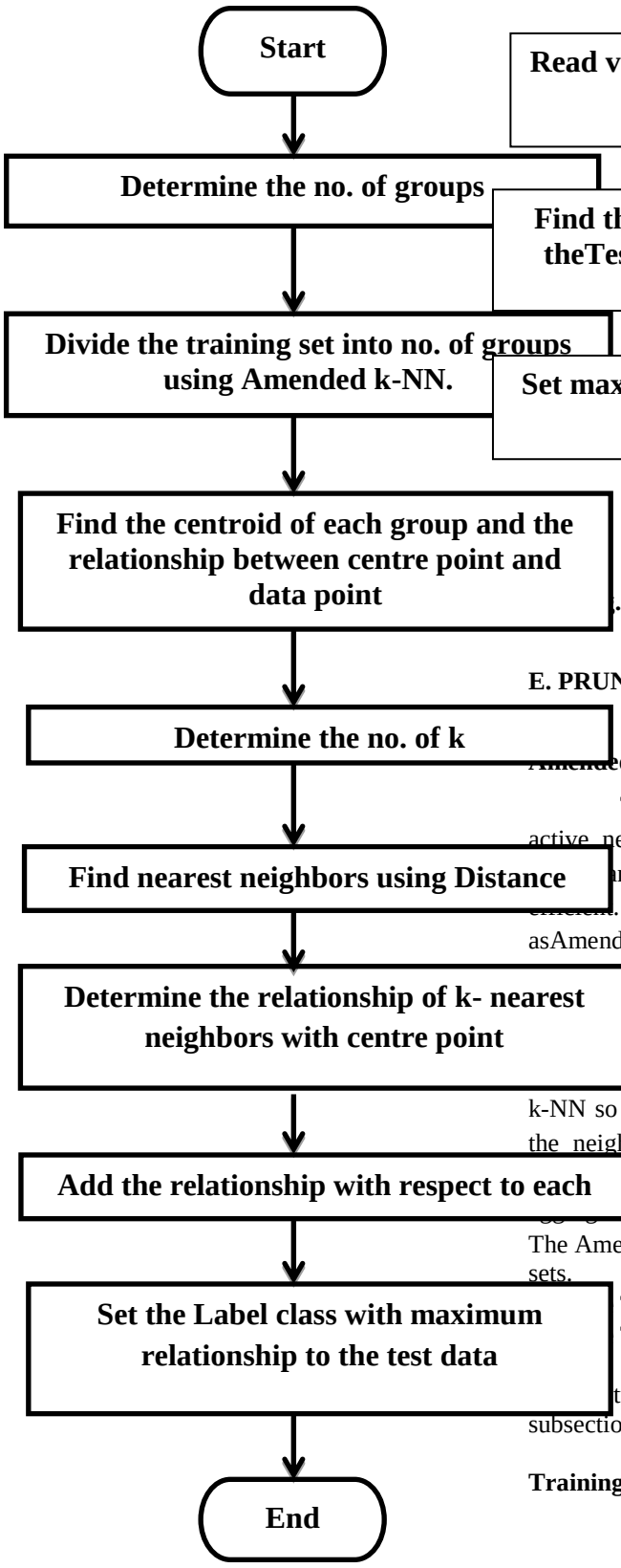
Fig.1. Flow Chart of ProposedSystem

D.Flow chart of Amended k-NN Technique

C.Flow Chart of the Proposed System



.2. Flow Chart of Amended k-NN Technique



E. PRUNING TECHNIQUES

ded k-NN Technique Working

The aimof the Amended k-NN technique is to use active neighbors in training set. This Amended k-NN is n traditional k-NN in both conditions: effective and . The proposed k-NN technique is named asAmended K-Nearest Neighbor. Compared to the k-NN technique, the main aim is to cluster. to its repetition in a set of neighbor tags, emsetis correlated as common. Amended K-eighbor is to be consideredas a type of weighted k-NN so that the query label is approximated by weighting the neighbors of the query. The strategy computes the of the equivalent labeled neighbors to the number of neighbors.

The Amendedk-Nearest Neighbor Technique consists of two sets.

Training Set
Testing Set

two sets are described briefly in the next subsections.

Training Set

Retrieval Number: F1003376S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Amended k-NN techniqueneeds a specific training setto be categorized. The training setincludes capturing the information vector coordinates by the side of the class tag. The class tag is used as an identifier for the information vector. The class tag helps to categorize information vectors during the testing set.

Testing Set

Given the data sets for testing, our idea is to get the class tag for the new set. The techniquestarts with k= 1 condition and then it is extended for k=k condition.

a) k=1

It is assumed that 'x' is the variable to be considered.

- Thevariablenearby to 'x' in the training set is found out. The next closest variableis assumed as 'y'.
- Now, the nearest neighbor conditionis invoked to allocate the tag of 'y' to 'x'.

This analysisproceeds when the amount of data sets is not much huge. If the count of data sets is very huge, then there is more chance that tag of 'x' and 'y'is common.

b) k=K or k-Nearest Neighbor

This is a straightforward additionof k-NNtechnique. 'K' nearest neighbors is to get for a given data set from the testing data and it is categorized by the outcome of election process done in the training set. Usually, 'K' is odd when the quantity of classes is 2.

Generalisingthe K-Nearest Neighbor Testing Set

- The rate of 'K' (input) is determined.
- The training set is preparedbycollecting the coordinates and class tags of the data sets.
- The testing data set is loaded.
- An election process is conducted amongst the 'K' nearest neighbors of the testing set from the training set.Based on a distance metric the node with the highest vote is conducted.
- The class tag of the winning node is assigned to the new data of the testing set.
- This processing is done repeatedly till the data in the testing sets are categorized.

II.MODULES OF AMENDED k-NN ALGORITM:

A. Frequent Itemset Mining

FIM is the most difficult and more time-consuming task in ARM, a most commonly used data mining job. FIM acts as a major tool for decision making by separating the most valued frequent patterns in parallelto happen in a major exchange database. A classic utilization of ARM is the most frequent market basket investigation.

B. K- Means Selection of Pivots

The selection of key points is identified easily for the consistency co-efficient of the left-overitems for Voronoi structure-based partitioning. In specific, we use the K-means based collectionpolicy to choose key points. The key pointselection process is arranged as an information pre-processing level. K-means is a famoustechnique for clustering analysis in data mining. K-means clustering attempts to partition 'n' into 'k' clusters i.e., given a lot of objects (x_1, x_2,..x_n), where each object is a D-dimensional real vector. Here, k-means clustering partitions the 'n' objects into 'k' ($k \leq n$) sets of$C = C_1, C_2,..C_k$ the cluster, in which each object belongs to acluster with the nearest mean. The clustering outputs can be used to partition the information space into Voronoi cells. To decrease the computational cost of k-means, the sampling of the exchange database is performed before running the k-means technique. It is worthwhile to mention that the selection of initial key points plays a critical role in clusters performance. Thus, k-means is assumed to conduct pivot selection. After the 'k'data clusters are created, we selectthe center point of each group as a pivot value for the Voronoi structure-based data partitioning technique.

C.MapReduceFramework

MapReduce is a prevalentstatisticshandlingmodel for efficient and fault tolerant loaddissemination in hugegroups. A MapReduce calculationhas two stages - the Map stage and the Reduce stage. The Map sectionssplit an input statistic into a big range of fragments. The fragments are equitably disseminated to Map jobs over a group of nodes to manner. Each Map job takes in a key-pointmatch and after that, it creates an arrangement of middle of the key-value sets. Then, MapReduce runtime framework gathers and sorts all the middle value related with a similar middle key. The runtime frameworkapplies the middle key to reduce jobs. Each Reduce jobs takes in all middle sets related with a specific key and

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

outputs a final set of key-value sets. MapReduce proceeds with computation being taken closer towards information. Further, MapReduce schedules map jobs to the nearest nodes where the input data is stored. This is done order to increase data locality.

C.ADVANTAGES of AMENDED k-NN TECHNIQUE

- Reduces additional overhead.
- Improves data locality.
- Reduces computing cost.
- Improves MapReduce &Hadoop performance.
- Uses unstructured data.

D.HARDWARE AND SOFTWARE REQUIREMENTS

Ubuntu 14.04 Long Term Support (LTS), Java (JDK 1.8) andEclipse (Mars) is used to implement the technique.

UBUNTU 14.04 LTS

Ubuntu has important features such as stability, speed, open source and reliability. There features motivatedus to use Ubuntu for running the Hadoop.

JAVA (JDK 1.8)

Java JDK is designed to be simple and architecture neutral, so that it can be executed on a variety of hardware platform. Its main characteristics such as easy learning, object oriented, and platform independent, inspired us to write programs and execute them on Hadoop platform.

ECLIPSE MARS

Eclipse Mars provides an Integrated Development Environment (IDE). It is the most commonly used Java IDE. It has been workspace and play-in system for the environment.

S.NO.	Specifications	Name Node	Data Node1	Data Node2
1	Hard Disk	60GB	50GB	40 GB
2	RAM	4GB	2GB	1 GB
3	Processor	Intel Core i5-4200U	Intel Pentium Dual Core, E-6700	Intel Core 2 Duo E-7500
4	CPU	2.30GH	3.20GHz	2.93GHz
		z		
5	Operating System	Ubuntu 14.04 LTS	Ubuntu 14.04 LTS	Ubuntu 14.04 LTS

Table1. System Specifications

III.OUTPUT

Multiple datasets are given as input to test the performance of Amended k-NN techniquein the heterogeneousHadoop cluster. The various datasets are:

1. CPU Scheduling Dataset
2. Breast cancer Dataset.
3. Cancer Dataset.

CPU SCHEDULER DATASET

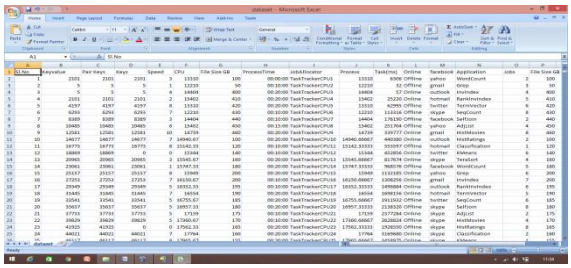


Fig. 3.CPU Scheduler Dataset

CPU Scheduler Dataset is adataset about CPU scheduling taken from the Kaggle website. It consists of key values, pair keys, speed, file size(GB), process time, job allocator, process, task, application, etc., The file size is about 6MB.

CANCER DATASET

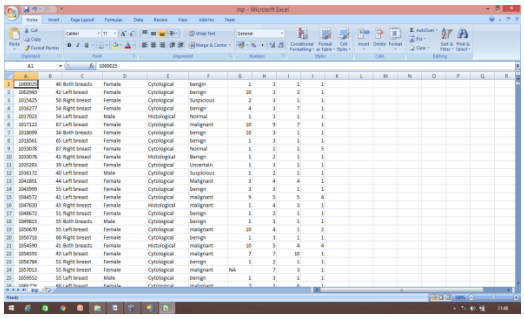


Fig.4. Cancer Dataset

Cancer Dataset describes the details of multiple patients having Breast Cancer. It is taken from the Kaggle website. Itstores the details of the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

patient including their Id, age, cancer location, cytological or histological, cancer condition, etc., The file size is about 3MB.

BREAST CANCER DATASET

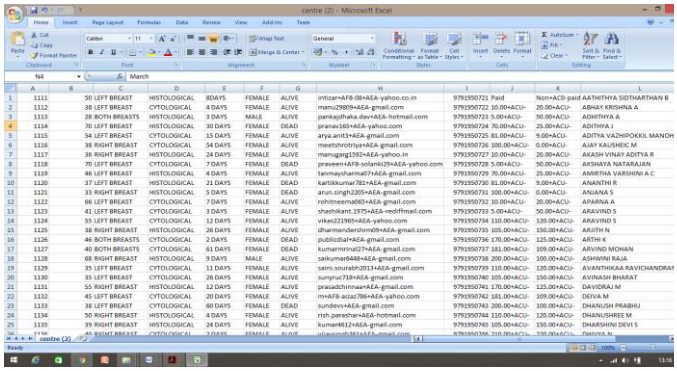


Fig.5. Breast Cancer Dataset

Breast Cancer Dataset is taken from the Unique Client Identifier (UCI) Machine Learning Repository. It has details about the patient, their contact info, etc,. Thefile size is about 4.5MB.

FINAL OUTPUT

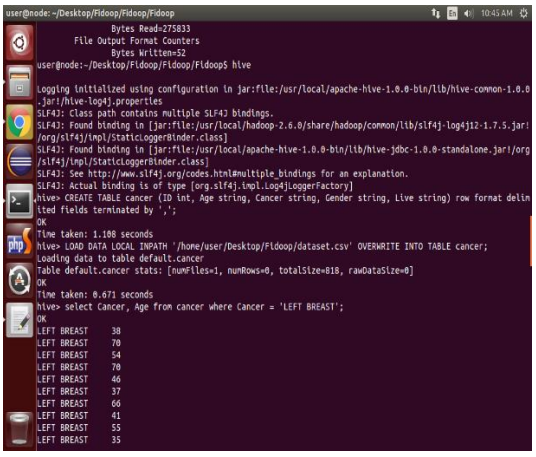


Fig.6. Generating Hive SQL Table

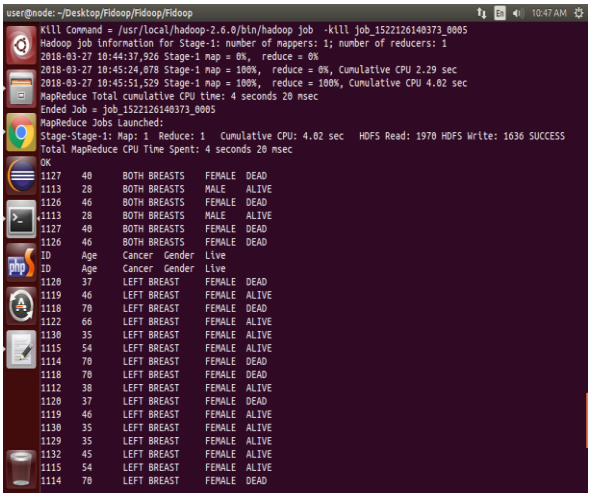
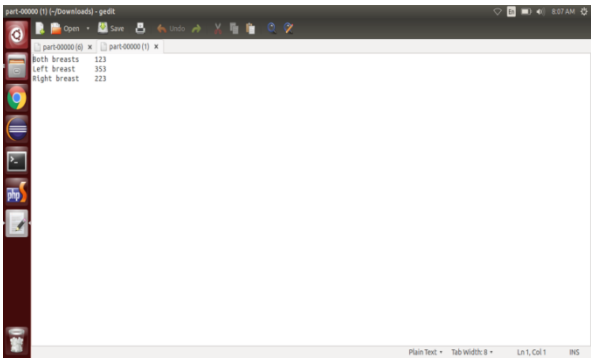
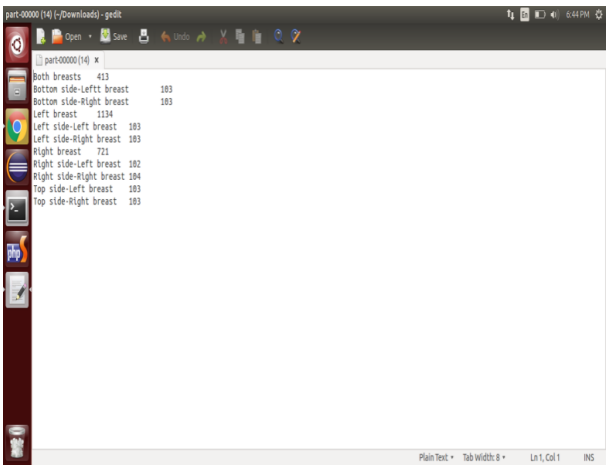


Fig.7. HDFS Output





OUTPUT PARAMETERS

The parameters considered are

- **Execution time-** The time taken to finish the actual execution of job or task.
- **Redundant transactions**
- **Computation cost-** Computational cost is visualized as the execution time for each step during implementation.It ismeasured as the time that the execution it takes for our model to execute on real-time hardware.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This is measured asthe executiontimefor each step during implementation.

Results and Comparison of Existing (FiDoop-DP technique) and Proposed Work (Amended k-NN technique)

This section discusses the experimental results. Here,the existing FiDoop-DP technique is compared with the proposed Amended k-NN technique.The three different datasets- Cancer Dataset, Breast Cancer Dataset and CPU Scheduling Dataset are considered for execution. Each of the dataset has been partitioned into a training setand a testing set.

The efficiency of the proposed Amended k-NN techniquehas been measured withthree different datasets. The details of these datasets are discussed in the previous section. These datasets are taken from Kaggle website and UCI (University of California, Irvine) Machine Learning Repository.

In Hadoopheterogeneous environment, the performance of the existing FiDoop-DP technique and the proposed Amended k-NN technique are analyzed using the three different datasets. A comparison of the existing FiDoop-DP technique and Amended k-NN technique is highlightedin the Table2 for the Cancer Dataset.Table 3 highlights the comparison of the currenttechnique and proposed algorithm for Breast Cancer Dataset. Table 4 highlights the comparison of the currentalgorithm and proposed algorithm for the CPU Scheduling Dataset.

Parameters	Existing (FiDoop-DP Technique)	Proposed (Amended k-Nearest Neighbor Technique)
Execution time	0.629	0.497
Redundant transactions	40%	25%
Computational Cost	30%	25%

Table2. Comparison of FiDoop-DP and Amended k-NN technique for Cancer Dataset

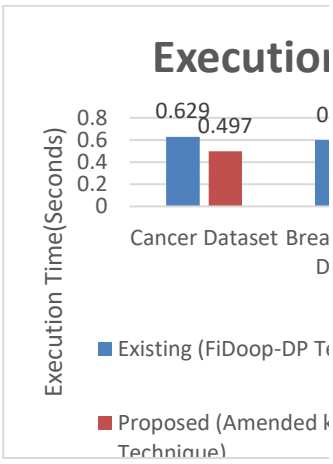
Parameters	Existing (FiDoop-DP Technique)	Proposed (Amended k-Nearest Neighbor Technique)
Execution time	0.600	0.450
Redundant transactions	40%	35%
Computational Cost	35%	28%

Table 3. Comparison of FiDoop-DP and Amended k-NN technique for Breast Cancer Dataset

Parameters	Existing (FiDoop-DP Technique)	Proposed (Amended k-Nearest Neighbor Technique)
Execution time	0.555	0.425
Redundant transactions	35%	25%
Computational Cost	30%	22%

Table 4. Comparison of FiDoop-DP and Amended k-NN technique for CPU Scheduling Dataset

The graphical representation of the current and proposed technique is depicted in Graph-1. It indicates the Execution time comparison of the current and the proposed technique implemented with three different datasets. Graph-2 represents Redundant transaction comparison of the current and the proposed technique implemented with three different datasets. Graph-3 highlights the Computation cost comparison of the current and the proposed technique implemented with three different datasets.

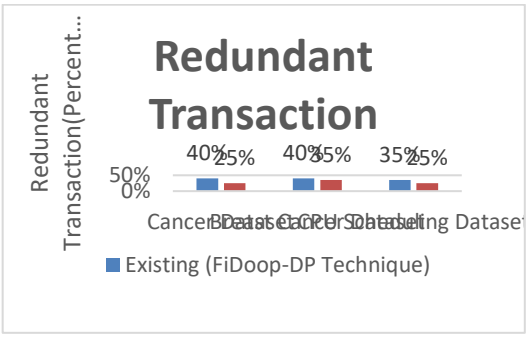


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Graph 1: Comparison of Execution time in both Existing and Proposed techniques

Significance of Graph 1:

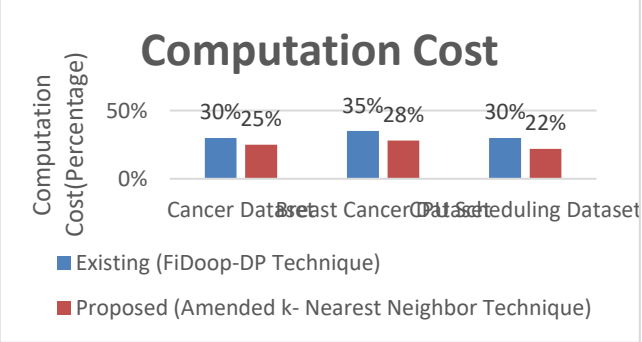
The Graph 1 shows the increase in the percentage of Execution time for three different datasets namely-Cancer datasets, Breast Cancer datasets and CPU Scheduling datasets. In the above Graph, X-axis represents the various datasets and Y-axis represents the Execution Time in seconds. The proposed technique is approximately 0.2 seconds better than the existing technique for all the three datasets. This fact is highlighted in the above Graph. The better performance of the proposed work with respect to the existing work is clearly depicted in the Graph 1.



Graph 2: Comparison of Redundant transactionof both Existing and Proposed techniques

Significance of Graph 2:

The effective increase in the percentage of Redundant transactions for the three different datasets is depicted in Graph 2. In the above Graph, X-axis represents the different datasets and Y-axis depicts the Redundant transaction in percentage. The result of the proposed technique is approximately 10-15% better than the existing technique.



Graph 3: Comparison of Computation cost in both Existing and proposed techniques.

Significance of Graph 3:

This Graph 3 represents the reduced percentage of Computation cost for three different datasets namely Cancer datasets, Breast Cancer datasets and CPU Scheduling datasets. In the above Graph, X-axis shows the various Datasets and Y-axis shows the Computation cost in percentage. The proposed work is approximately 5-10% greater than the existing work.

IV.CONCLUSION

Motivated by the performance decrease caused by homogeneity,We have proposed a data balancing scheme in HDFSon heterogeneousHadoop environment. The proposed data balancing mechanism balances the data blocks of an input file on heterogeneousHadoop clusters according to their processing capabilities. The Amendedk-NN (k-Nearest Neighbors) technique significantly increases the MapReduce performance of Hadoop on heterogeneous clusters.Thus, after balancing the data on heterogeneousHadoop cluster, FIM is implemented using Amendedk-NN (k-Nearest Neighbors) technique on theheterogeneousHadoop cluster.The proposed techniquethus is amended effectively to reduce execution time, redundant transaction and Computation cost.Experimental results reveal that Amended K- Nearest Neighbor significantly improves the Frequent Itemset Mining (FIM) performance of the existing FiDoop-DP solution by 31 percentage with an average of 18 percentage. The proposed system fairs well compared to the existing systems in terms of the parameters-Data Execution time, Redundant Transaction and Computation cost.

V.FUTURE ENHANCEMENT

The present proposed system can beextended by integrating Amended K- Nearest Neighbor with a data-placement scheme in HadoopDistributed File System on Heterogeneous environments in order to further increase the load balancing schemes.

REFERENCE

1. Yalingxun, Jifu Zhang and Xiao Qin, “FiDoop-DP:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Data Partitioning in Frequent Itemset Miningon Hadoop Clusters", IEEE Transactions on Parallel and Distributed Systems, Jan 1 2017, Volume: 28, Issue: 1, Pp 101-113.

2. Shavachko K, Kuang H, Radia S and Chansler R, "The Hadoop Distributed File System", In proceedings of the 26th Symposium on Mass Storage Systems and Technologies., 2017, Washington, DC, USA: IEEE. P. 1-10.
3. Yalingxun, Jifu Zhang and Xiao Qin, "Fidoop: Parallel Mining of Frequent Itemsets Using MapReduce", IEEE Transactions on Systems and Cybernetics Systems, March 2016, Volume: 46, Issue: 3, Pp. 313-325.
4. Ziming Zhong, Vladimir Rychkovand Alexey Lastovetsky, "Data Partitioning on Multicore and Multi-GPU Platforms Using Functional Performance Models", IEEE Transactions on Computers, Sept. 1 2015, Volume: 64, Issue: 9, Pp. 2506-2518.
5. Ivaniltonpolato, Reginaldo Re, Alfredo Goldman and Fabio Kon, "A Comprehensive View of Hadoop Research – A Systematic Literature Review", Elsevier Journal of Network and Computer Applications 2014.
6. Lin M, Zhang L, Wiermanaand Tan J, "Joint Optimization of Overloading Phases in MapReduce", IEEE Transaction on Computers Perform Eval 2013; 70(10): 720-35.
7. Costa P,Pasin M, Bessani A, Correia M.O , "The Performance of Byzantine Fault tolerant MapReduce", IEEE Trans Dependable Secure Computing 2013.
8. Xiej, Tian Y, Yin S, Zhang J, Ruan X and Qin X, "Adaptivepre-Shuffling in Hadoop Clusters", Procedia, Computer Science IEEE Transaction on Computer Science 2013: 18(0):2458-67.
9. Zacharia FadikaandMadhusudhanGovindaraju, "LEMO-MR: Low Overhead and Elastic MapReduce Implementation Optimized for Memory And CPU-Intensive Applications", IEEE International Conference on Cloud Computing Technology and Science, 2010, In Cloudcom, Pages 1–8.
10. Leo S and Zanetti G. Pydoop, "A Python MapReduce And HDFS API for Hadoop",In proceedings of the 19th International Symposium on High Performance Distributed Computing, IEEE Transactions on Systems, Man, and Cybernetics: Systems New York, NY, USA: ACM; 2010. P. 819-25.
11. Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.
12. Rajesh, M., and J. M. Gnanasekar. "Sector Routing Protocol (SRP) in Ad-hoc Networks." Control Network and Complex Systems 5.7 (2015): 1-4.
13. Rajesh, M. "A Review on Excellence Analysis of Relationship Spur Advance in Wireless Ad Hoc Networks." International Journal of Pure and Applied Mathematics 118.9 (2018): 407-412.
14. Rajesh, M., et al. "SENSITIVE DATA SECURITY IN CLOUD COMPUTING AID OF DIFFERENT ENCRYPTION TECHNIQUES." Journal of Advanced Research in Dynamical and Control Systems 18.
15. Rajesh, M. "A signature based information security system for vitality proficient information accumulation in wireless sensor systems." International Journal of Pure and Applied Mathematics 118.9 (2018): 367-387.
16. Rajesh, M., K. Balasubramaniaswamy, and S. Aravindh. "MEBCK from Web using NLP Techniques." Computer Engineering and Intelligent Systems 6.8: 24-26.

Retrieval Number: F1003376S19/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication