Big Data Mining For Interesting Pattern Using MapReduced Technique

Aguguo Ihechukwu.C[#1], Matthias Daniel[*2],E.O Bennett [#3]

[#]*Computer Science Department, Rivers StateUniversity*
Nkpolu, Portharcourt, Rivers State Nigeria

Abstract

In the past few years, huge data are need to be stored, access and retrieved, that has increased drastically all over the world, this fast growth of data results in the need to analyse the huge amount of data. Due to lack of proper tools and programs, data remains unused and unutilized with important useful knowledge hidden. This study has carryout data mining interesting patterns in big data. Object-oriented design methodology was used. Frequent pattern growth algorithm on Hadoop using MapReduce has been used and particularly applied it to analyze maximum flight time in flight transaction data store of 108MB. MapReduce program consists of two functions Mapper and Reducer which runs on all machines in a Hadoop cluster. System was implemented in matlab. Computation has been performed to analyzed the actual flight time using user constraints, the constraints are arrival delay and actual elapse time. Airpeace carrier has the longest flight time, the analyzed carrier (Air peace) space was 20000x6 contained 712316 bytes. Thus, the execution time of the entire mining process was 1615 milliseconds.

Keywords — *Big data, pattern mining, itemsets,reducer, mapper.*

I. INTRODUCTION

In these recent years, data usage has increased rapidly worldwide. So storing and accessing the large amount of data becomes a big problem. The data generated from different generating factors like machine logs, human generated data are being stored by companies and are used for decision making

Thinking about the Volume, Variety, and Velocity are the basic components of data the Volume defines the humongous storage volume, the Velocity defines the data transmission rates. The Variety describes data types to be heterogeneous. The Veracity focuses on the accuracy of data (the processing of data from different noises). At length, the attribute of viability and value: while the previous determines the various likelihood of data prediction, the latter appears to acquire worthly knowledge (Kudyba, 2014). In order to design big data mining techniques and platforms, understanding these dimensions is essential.

Data mining is used to find patterns that are useful, hidden and unknown pattern from large databases. Pattern mining is a set of problems that seek to identify combinatorial patterns (or features) from databases.There are numerous known data algorithms which explore interesting patterns of accurate data from databases. There is situation, however, in which data is unknown. Items of uncertain data bases in each transaction having probabilities, that shows the odds of these things being present in the database Data mining in various genuine applications. Without furnishing clients with offices to communicate the intriguing examples to be abused, many existing information mining calculations return bunches of examples out of which just a couple are of intrigue. So every exchange incorporates objects and their existential probabilities when the information is unsure. As far as existential likelihood, the likelihood of such component can be unequivocal.

Data mining is the extraction of knowledge from great deal of observational data sets, to seek out surprising relationship and pattern hidden in data, summarize the data in novel ways that during which to make it comprehendible and useful to the data users. Web usage mining is that the appliance of data mining technique to automatically discover and extract useful information from a particular computer (Bounch et al., 2001). There are many major data processing techniques Classification: Classification may be a one in all the classical data processing techniques that is predicated on machine learning. Essentially classification is applied to classify every item during a set of information into one in all predefined set of categories. Classification technique makes use of many mathematical techniques like linear programming, call trees, neural network and statistics

II. RELATED WORK

In section. We summarize previous works based on frequent dataing mining

Naik and Mankar (2013) proposed the approach for mining of frequent pattern itemsetsshow in dubious database with the assistance of probabilistic back values. Here, aproricalculation utilized with the probabilistic bolster values for finding visited designs. Ordinarily, aprioriecalculationover and overproduces the probabilistic visited itemsetsemploying a bottom-up procedure. Each cycle was performed by the taking

after two steps. To begin with connect step, utilized for creating modern candidates and following pruning step, utilized for calculating the probabilities frequentness of things and calculating the probabilistic visititemsets from the created candidates.

Leung, C.K. (2007) proposed the approach for efficiently mining the patterns from the huge probabilistic data. He elaborated the strategy of UF-growth algorithm to find the frequent pattern. The UF-growth algorithm was actually motivated from two basic techniques that are frequent-pattern growth algorithm and U-apriori algorithm. This algorithm first scan database and calculate the expected support count for particular items. Arrange in descending order of expected support count. While doing this, it also checks for the minimum support count and user specified constraints. Once it was done with this algorithm again, it has to scan the database and construct the UF tree for finding the frequent pattern. The tree will grow according to the scanning of transactions.

Riondato et al. (2012) proposed 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 work in parallel with their assigned set to produce deliverables and to be filtered and aggregated into a single output set.

III. MATERIALS AND METHODS

The research method adopted is constructive research. It is a science of studying how research is carried out. It is also defined as the study of methods by which knowledge is gained. Its aim is to give the work plan of research.

A. Constructive Research Procedure:

i. **Practical relevant problem that has research potential:**Massive sets of data such as big data having data sets that may be structured, semi-structured or unstructured make data mining difficult.

ii. **Obtaining a general understanding of the topic:**This study deals with Big data mining for interesting patterns. MapReduce and association rules are used to generate frequently repeating item for user interest.

iii. **Innovating- designing of a new construct:** Object Oriented Design Analysis (OODA) is adopted in the new construct; use-case and class diagram are used to model the system

iv. **Demonstrating that the new solution works:** The frequent item set data mining was implemented in MatLab.

B. DESIGN METHODOLOGY

Objected-oriented design method was adopted. This concerned with developing an object oriented system to implement requirements. Requirements planning; through this phase we identify the inputs and outputs of the system, and identify frequently repeating item based on minimum support. Our actual software development methodology uses an Object-Oriented Method and Recursive Method (OOM/RD) for which the entire system is broken down into subsystems and modules.

In the OOM/RD model software processes are basically the same, with early parts of the process defining a topmost level structure, and these processes reapplied to parts of the main structure in turn to define much greater details (Ramirez et al., 2011). The steps in a typical OOM/RD model will generally include the following details: Identification of critical objects of the main systems design by breaking them down into modules (smaller blocks) or subsystems, performing software processing on identified objects, re-applying software processing on the identified objects. The steps are crucial to almost any object oriented engineering design and must be performed in a recursive manner to arrive at reasonable performance estimates if OOM/RD is to be used in the systems design.

User design; System Development Life Circle (SDLC) class diagram and sequence diagram to show the interaction between user and systems. SDLC implementation phase; using MatLab.

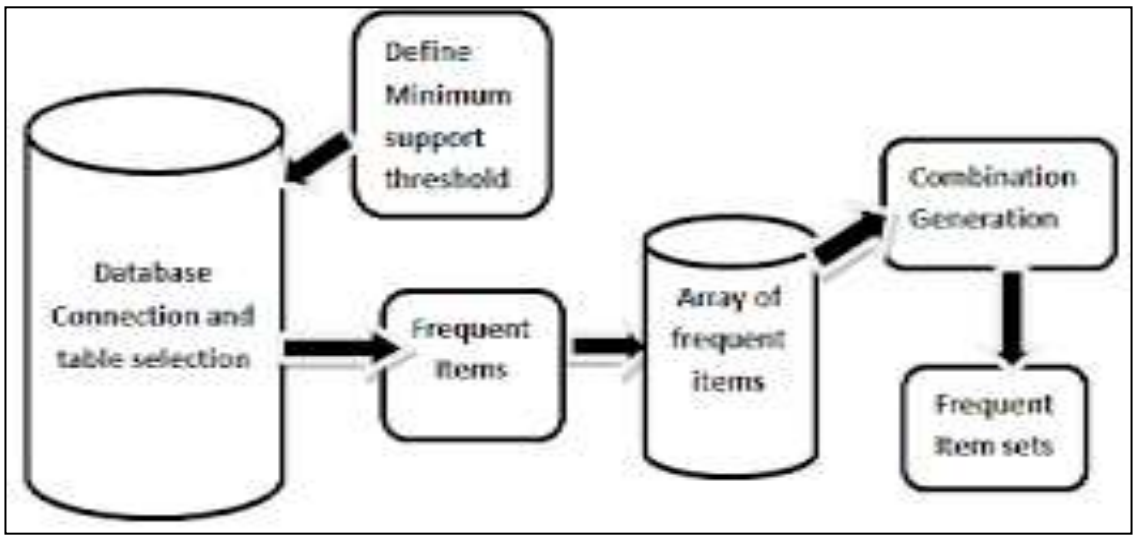


Fig 1: Architectural of Existing System

The database server used for this system was the MySQL database server. To access the transactions_db database, a connection had to be established with the database server.

C. PROPOSED SYSTEM ARCHITECTURE

Architecture specifies the structure, views and action of a system. The proposed system uses MapReduce technique to identify frequently repeating item. Figure 2 shows the component of MapReduce data mining system.

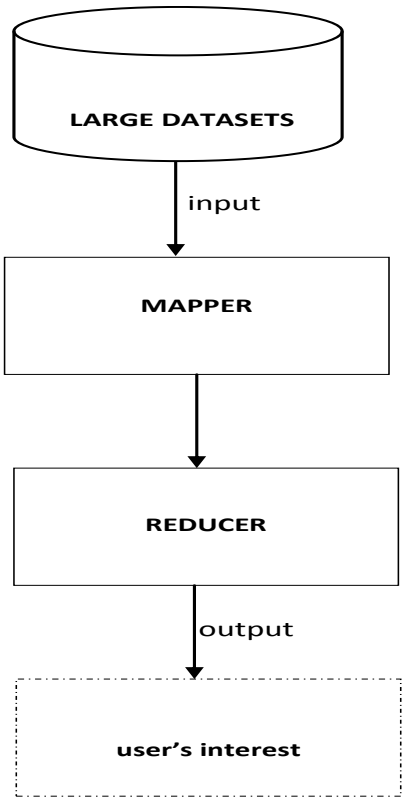


Fig 2: Architecture of the Proposed System

There are four mechanism of the proposed system: large datasets, mapper, reducer and the outcome (user's interested item). The proposed system accepts large datasets and performs linear execution of jobs on the datasets using mapper component followed by execution of a user defined parameter used by the component called reducer which reduces the large datasets to produce user's interest on dataset.

D. DATASET

Dataset is a triplet D = (A, B, C, E, F, G), where A, B is a finite set of objects, C, E is a finite set of items and F, G is a binary (incidence) relation. Therefore, [A, B, C, E] ϵ F, [A, B, C, E] ϵ G if the object a ϵ A, b ϵ B, c ϵ C and e ϵ E. Table 3.1 illustrates an item in a given dataset.

Datasets are converted in the form of (key1, value1) pairs. These (key1, value1) pairs are input into the mapper. Where key is the Transaction Identity (TID) and value is the list of items i.e. transaction. Mapper reads one transaction at time and output (key', value') pairs where key' is each item in transaction and value' is 1. The combiner combines the pairs with same key' and makes the local sum of the values for each key'. The output pairs of all combiners are shuffled & exchanged to make the list of values associated with same key2, as (key2, list (value2)) pairs.

Transactions	List of Items
T1	A,B,C
T2	A,C
T3	B,C,D,E
T4	A,D
T5	E
T6	C,D,E,F,G
T7	B
T8	D
T9	C, B

Table 1 : Dataset

E. MAPPER

MapReduce processes large volumes of data by breaking the job into independent tasks. MapReduce program consists of two functions Mapper and Reducer. The input and output of these functions must be in form of (key, value) pairs. Mapper takes the input (k1, v1) pairs and produces a list of intermediate (k2, v2) pairs as shown in Figure 3.4. Output pairs of mapper are locally sorted and grouped on same key, shuffled and group all the pairs with the same key to a single reducer.

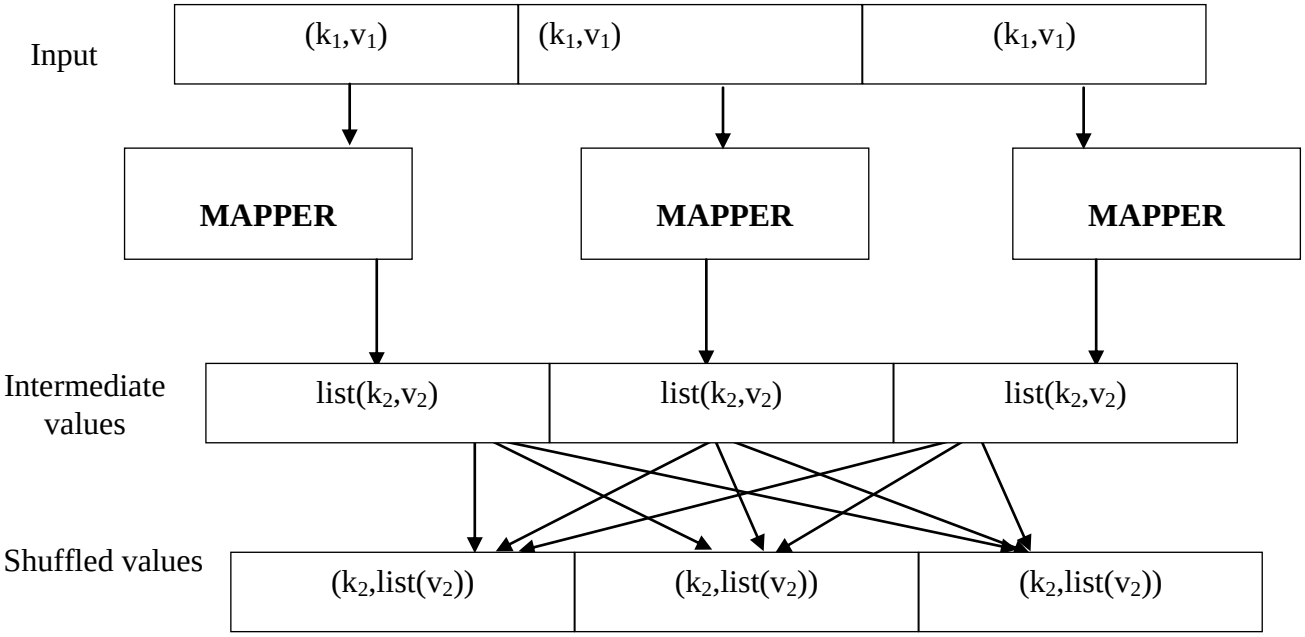
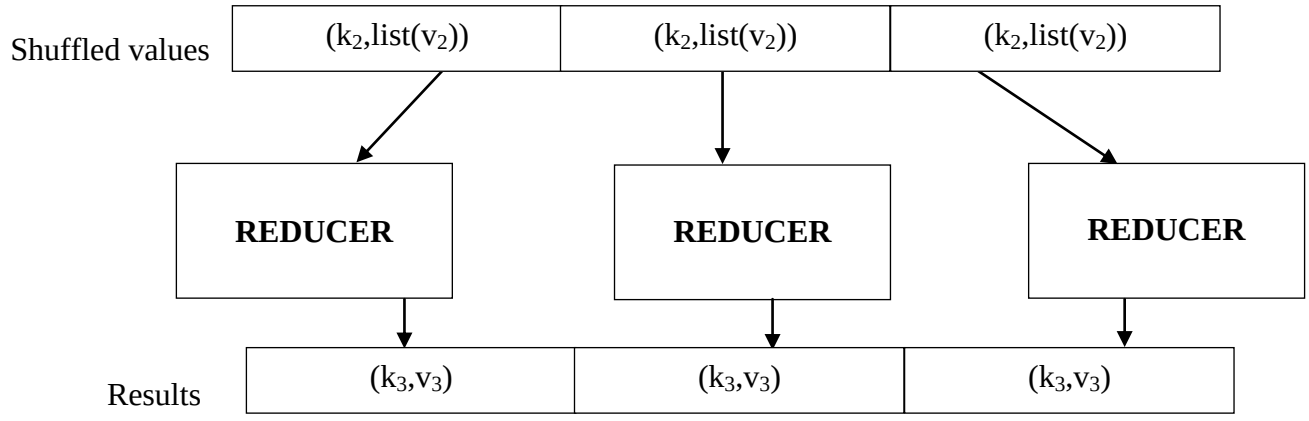


FIG 3: MAPPER

F. REDUCER

Reducers take these (key2, list (value2)) pairs and sum up the values of respective keys. Reducers output (key3, value3) pairs where key3 is item and value3 is the support count ≥ minimum support of the items. Mapper and Reducer are demonstrated in Figure 3.6.



Reducer takes (k2, list (v2) values as input, make sum of the values in list (v2) and produce new pairs (k3, v3) as depicted in Figure 3.5.

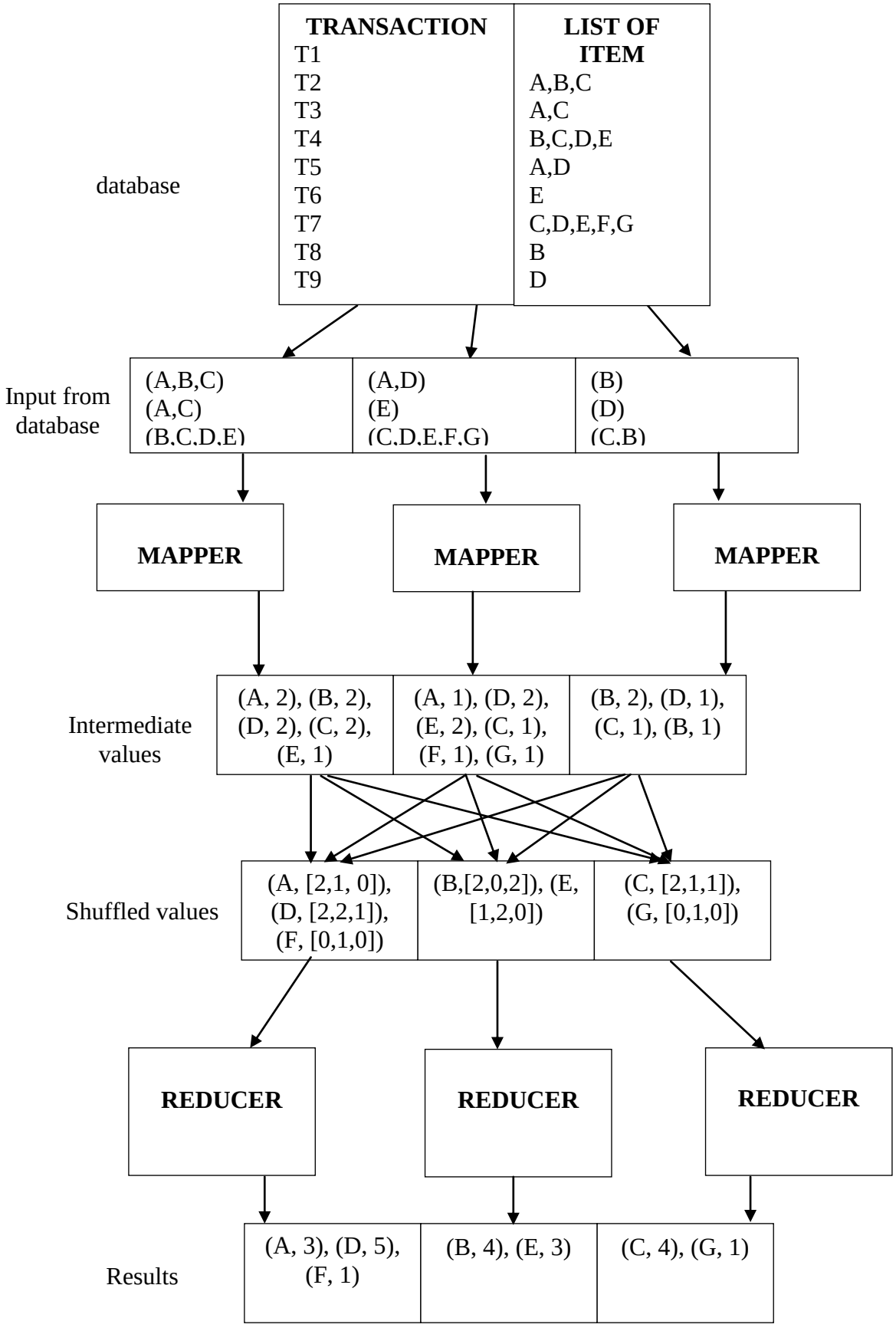


Fig 4: Producing User Interested Frequent Item

Database consists of transaction records having Transaction Identify (TID) and List of Items (LI). The TID ranges from T1 to T9, these datasets are arranging in three columns and supplied into the mapper. The mapper processes the items and displays the items together with the occurrence of each item in column, i.e. (A, 2), this means item A has 2 occurrences in first column, this becomes "intermediate values". The intermediate values are shuffled and exchange in the three columns. In the shuffled values all items and their occurrences in each column are identified, i.e. (D, [2,2,1]) means that item D has 2 occurrences in first column, 2 occurrences in second column and 1 occurrence in third column. The reducer sums the occurrences of respective items, i.e. (A, 3) means item A occurs 3 times.

Table 2: Maximum Flight Time

Algorithm: MapReduce on Itemsets

1. Mapper (key, value)
2. // key: TID
3. // value: itemsets in
4. transaction Ti
5. for each transaction Ti
6. assigned to Mapper do
7. for each itemset in Ck
8. do
9. if itemset∈ Ti
10. output (itemset, 1);
11. end if
12. end for
13. Combiner (key, value)
14. // key: itemset
15. // value: list (1)
16. for each itemset do
17. for each 1 in list (1) of
18. corresponding itemset do
19. itemset.local_sup + = 1;
20. end for
21. output (itemset,
22. itemset.local_sup);
23. end for
24. Reducer (key, value)
25. // key: itemset
26. // value: list (local_sup)
27. for each itemset do
28. for each local_sup in
29. list (local_sup) of
30. corresponding itemset do
31. itemset.sup + = local_sup;
32. end for
33. if itemset.sup ≥ minimum
34. support;
35. output (itemset, itemset.sup);
36. end for

H. EVALUATION OF RESULTS

Mining frequent data from large amounts of data is followed by testing the amount of time and space required to find interesting patterns for users. For the top 3 carriers (airpeace, arik, dana) the number of flights per day was evaluated. 108 MB flight records are structured in tabular text files that were saved on a local disk. Using map reduction algorithm, broad data were analyzed. To analyze the actual flight time, computation was performed using the variable "actual elapsed time," which served as user constraint. The flight transaction data store's maximum flight time is shown in Table 1.1

The carrier with the longest flight is depicted in Table 4.2. (Is it table 2 you are referring)

Fligh tNum	TailNum	ActAPl Elapsed Time	CRSE Lapsed Time	Air Time	Arr Delay
335	'N712SW'	128	150	116	-14
3231	'N772SW'	128	145	113	2
448	'N428AIRPEACE'	96	90	76	14
1746	'N612SW'	88	90	78	-6
3920	'N464 AIRPEACE '	90	90	77	34
378	'N726SW'	101	115	87	11
509	'N763SW'	240	250	230	57
535	'N428 AIRPEACE '	233	250	219	-18

Table 2: Maximum Flight Time

The carrier with the longest flight is depicted in Table 4.2. (Is it table 2 you are referring)

UniqueCa rrier	Flight Num	ArrDelay	DepDelay	Origin	Destination
AIRPAEC E	448	14	8	LOS	OW
AIRPAEC E	1746	-6	-4	LOS	OW
AIRPAEC E	3920	34	34	LOS	FCT
AIRPAEC E	378	11	25	LOS	ACR
AIRPAEC E	509	57	67	LOS	ACR
AIRPAEC E	535	-18	-1	LOS	ACR
AIRPAEC E	11	2	2	LOS	OW
AIRPAEC E	810	-16	0	LOS	PH
AIRPAEC E	100	1	6	LOS	ENU
AIRPAEC E	1333	80	94	LOS	PH

Table 3: Longest Flight Time

Graphical representation of flight arrival and departure intervals shown in Figure

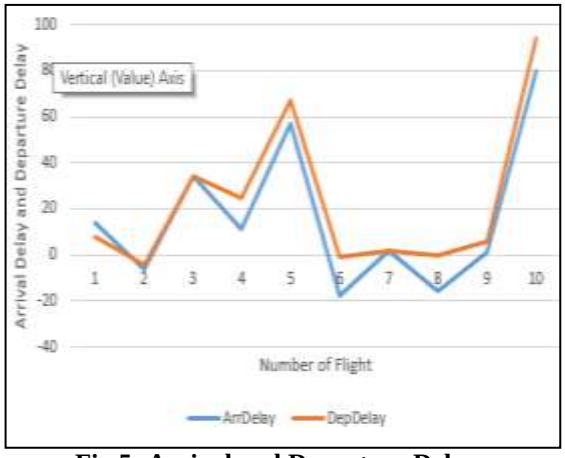


Fig 5: Arrival and Departure Delays

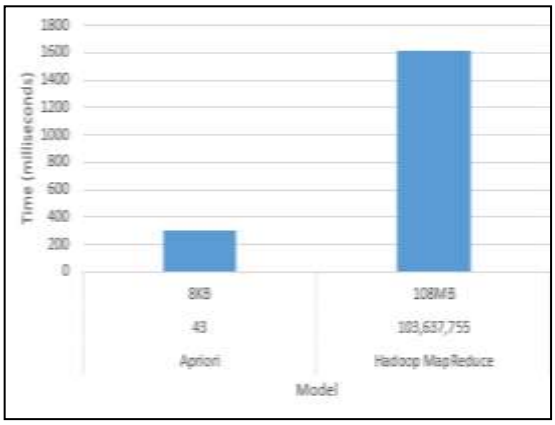


Fig 5: Execution Time

IV. CONCLUSION

In the past few years ago data are need to be stored that has increased drastically all over the world. (Check) This incredibly fast growth of data results in the need to analyze the huge amount of data. Due to lack of proper tools and programs, data remains unused and unutilized with important useful knowledge hidden.This study has carryout data mining on interesting patterns in big data. Frequent pattern growth algorithm on Hadoop using MapReduce has been used and particularly applied it to analyze maximum flight time in flight transaction data store. MapReduce program consists of two functions Mapper and Reducer which runs on all machines in a Hadoop cluster. The inputand output of these functions must be in form of (key, value) pairs. The run time system of MapReduce parallelizes the execution of mapper and reducer on a number of machines. It partitions the datasets into fixed sized blocks and replicates with some replication factor to provide high availability and zero data loss.Computation has been performed to analyzed the actual flight time using user constraint.

ACKNOWLEDGMENT

My profound gratitude goes to my supervisor DR. M. Daniel, for his guidance, patience and most importantly, for his intellectual knowledge which guided me in the achievement of this project.

My gratitude also goes to my Head of Department (HOD) Dr.V.I.EAnireh, for his advice toward this project.I am also indebted to my lectures Dr. E.O Bennett and Dr.Nwiabu for their contributions.

I also want to knowledge my parents Sir/Lady Benjamin Iwueze, for their financial support towards the success of this project, and to my siblings, my course mates, may thanks to you all.

REFERENCES

[1] Naik, R. R. and Mankar, J.R. (2013). "*Mining frequent Item sets from uncertain databases using probabilistic support*". International Journal Emerg. Trends Technology Computer Science, 2(2), pp. 432-6.

[2] Leung, C.K. (2007). "*Efficient mining of frequent patterns from uncertain data*". In: Seventh IEEE International Conference on Data Mining, pp. 204.

[3] Ramirez, U. Heinzelman, J., & Waters, C. (2011)."*Crowd sourcing crisis information in disaster-affected Haiti (US Institute of Peace)"*.In Proceedings of the International Conference on Unmanned Aircraft Systems, 16.

[4] Riondato, M., DeBrabant, J. A., Fonseca, R. and Upfal, E. (2012). "*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, pp. 85-94.

[5] Sangavi, S., Vanmathi, A., Gayathri, R., Raju, R., Paul, P. V. and Dha-vachelvan, P. (2015). "*An Enhanced DACHE Model for the MapReduce Environment*". Procedia Computer Science, 50, pp. 579-584.

[6] Tanbeer, S.K. and Leung, C.K. (2013). "*PUF-tree: A compact tree structure for frequent pattern mining of uncertain data"*. Pac Asia Conference Knowledge Discovery Data Min LNCS, 7818, pp. 13-25.

[7] Toivonen, H. (1996). "*Sampling large databases for association rules*".In*VLDB*, 96, pp. 134-145.

[8] Woo, J. (2012). "*Apriori-Map/Reduce Algorithm*". In International Conference on Parallel and Distributed Processing Techniques and Ap-plications (PDPTA), 45.

[9] Wu, G., Li, H., Hu, X., Bi, Y., Zhang, J. and Wu, X. (2009). MReC4.5: C4.5 Ensemble Classification with MapReduce. *ChinaGrid Annual Conference*, 4, 249-255.

[10] Yahya, O., Hegazy, O. and Ezat, E. (2012). "*An Efficient Implementation of Apriori Algorithm Based on Hadoop-MapReduce Model*". International Journal of Reviews in Computing, 12, pp. 59-67.

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

[12] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S. and Stoica, I. (2012). "*Resilient distributed datasets: A fault tolerant abstraction for in-memory cluster computing*". In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, pp. 2-2.

[13] AggarwaL, C. C. and Han, J. (2014).Pattern Design Mining. Springer, 23.

[14] Agrawal, R. and Srikant, R. (1994). "*Fast calculations for mining association rules*". In Procedures 20th worldwide conference exceptionally expansive information bases, VLDB, 1215, pp. 487-499.

[15] Agrawal, R., Imieliński, T. and Swami, A. (1993). "*Mining affiliation rules between sets of things in huge databases*". SIGMOD Record, 22(2), pp. 207–216.
[16] Amartya, S. and Kundan, K.D. (2007). "*Application of Data mining Techniques in Bioinformatics*", B.Tech Computer Science Engineering thesis, National Institute of Technology, (Deemed University), *Rourkela*.
[17] An, A., Khan, S. and Huang, X. (2003). "*Objective and Subjective Algorithms for Grouping Association Rules*". *Proceedings Third IEEE International Conference on Data Mining (ICDM)*, pp.477-480.
[18] Apache, M. (2013).Algorithms - Apache Mahout.Retrieved May 2019, from https://cwiki.apache.org/confluence/display/MAHOUT/Algorithms.
[19] Baffour, K. A., Osei-Bonsu, C. and Adekoya, A. F. (2017). "*A Modified Apriori Algorithm for Fast and Accurate Generation of Frequent Item Sets*".International Journal of Scientific & Technology Research, 6(8).
[20] T.K.Das , Arati Mohapatro."*A Study on Big Data Integration with Data Warehouse*". International Journal of Computer Trends and Technology (IJCTT) V9(4), 2014
[21] Bounch, F., Giannotti, F., Gozzi, C., Manco, G., Nanni, M., Pedreschi, D., Renso, C. and Ruggier, S. (2001). "*Web log data warehourseing and mining for intelligent web caching",* Journal Data Knowledge Engineering, 36, pp. 165–189.
[22] Chu, C. T., Kim, S. K., Lin, Y. A., Yu, Y. Y., Bradski, G., Ng, Y. A. and Olukotun, K. (2006). "*Map-Reduce for pattern Learning on Multicore*". Advances in Neural Information Processing Systems, 19, pp. 281-288.
[23] Crikovic, G. D. (2010). "*Constructive Research and info-computational knowledge Generation*", *Springer*, 314.
[24] Dhanshetti, A. and Rane, T. (2015). "*A Survey on Efficient Big Data Clustering using MapReduce*". *Data Mining and Knowledge* Engineering, 7(2), pp. 47-50.