International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1399-1403 www.ijaem.net ISSN: 2395-5252

Improving Big Data Miningenvironment Using Hadoop Mapreduce Technique

Ejimofor, IhekeremmaA.U.
[1]Department of Computer Science NnamdiAzikiwe University, Awka
[2]Ifeagwu E.N.[2]Department of Electrical/Electronics Engineering, Michael Okpara University of Agriculture, Umudike.

Submitted: 10-08-2021 Revised: 22-08-2021 Accepted: 25-08-2021

ABSTRACT: This research paper focused on improving Big Data Mining systems using Hadoop MapReduce technique.The hardware cluster was created by connecting all the commodity hardware (3 Data Nodes and 1 Name node) in a Local Area Network. The commodity hardware consists of one Name node (200GB) and 3 Data Nodes (1TB) each. The data was uploaded in the Hadoop Data Base File System(HDFS), the Name Node divides the data across the three Data Nodes, and performs data integrity and safety measures in order to preserve the integrity of either nodes failing. The files are also replicated across the cluster. Results showed that the Euclidean distance and the pseudo F-statistic validated Hadoop's high scalability and performance thus the preferred data mining solution was achieved.

KEYWORDS: Hadoop, MapReduce, Pseudo F-statics, Euclidean distance, Big Data.

I. INTRODUCTION

[1].Conventional data storage and data mining systems were not built keeping in mind the needs of big data and hence no longer easily and cost-effectively support today's large datasets.

In a broad range of application areas, data is being collected at unprecedented scale. [2].Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such big data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

[3].Big data is very difficult to deal with. It requires proper storage, management, integration, federation, cleansing, processing, analyzing, etc. [4].With all the problems faced with traditional data management, big data exponentially increases these difficulties due to additional volumes, velocities, and varieties of data and sources which have to be dealt with.

Owing to the size of big data, sophisticated tools are required to discover useful patterns from the vast number of potential relationships — hence the role for knowledge discovery through advanced analytics using data mining techniques. Instead of relying on expensive, proprietary hardware to store and process data, Hadoop MapReduce technique enables distributed processing of large amounts of data on large clusters of commodity servers. Hadoop [5].MapReduce Technique is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

Therefore, this paper focuses on providing a roadmap or framework for improving big data mining using Hadoop MapReduce technique which can encompass the previously stated difficulties.

II. CLUSTERING

[6]. Clustering refers to the grouping of records,observations or cases into classes of similar objects.[7]A cluster is a collection of records that are similar to one another and dissimilar to records in other clusters.Clustering algorithms segment the entire data set into relative homogenous subgroups or clusters where the similarity of the records within the cluster is maximized and the similarity to records outside this cluster is minimized.

[8].The centroid (median) is the centre.of a cluster. For finding clusters in data, the nearest criterion used is usually Euclidean distance as is stated in equation (1):

$$d_{Euclidean}\ (x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (1)$$

where $x = x_1, x_2,\dots x_m$,and $y = y1_1, y_2, \dots, y_m$ represent the m attribute values ofthe records.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1399-1403 www.ijaem.net ISSN: 2395-5252

The algorithm terminates when the centroids no longer change. In other words, the algorithm terminates when for all clusters C_1, C_2, …,C_k, all the records owned by each cluster center remain in that cluster. [8].Alternatively, the algorithm may terminate when some convergence criterion is met, such as no significant shrinkage in the mean squared error (MSE) given in equation (2):

$$MSE = \frac{SSE}{N-k} = \frac{\sum_{i=1}^{k}\sum_{p\in C_i} d(p,m_i)^2}{N-k} \quad (2)$$

where SSE represents the sum of squares error, $p \in C_i$ represents each data point in cluster i, m_i represents the centroid (cluster center) of cluster i, N is the total sample size, and k is the number of clusters. Recall that clustering algorithms seek to construct clusters of records such that the between-cluster variation is large compared to the within-cluster variation. [8].Because this concept is analogous to the analysis of variance, we may define a pseudo-F statistic as follows:

$$F_{k-1,N-k} = \frac{MSB}{MSE} = \frac{SSB/k-1}{SSE/N-k} \quad (3)$$

[8].where SSE is defined as above, MSB is the mean square between, and SSB is the sumof squares between clusters, defined as:

$$SSB = \sum_{i=1}^{K} n_i \,.\, Distance^2(m_i, M) \quad (4)$$

where n_i is the number of records in cluster i, m_i is the centroid (cluster center) for cluster i, and M is the grand mean of all the data.MSB represents the between-cluster variation and MSE represents the within-cluster variation. Thus, a "good" cluster would have a large value of the pseudo-F statistic, representing a situation where the between-cluster variation is large compared to the within-cluster variation. Hence, as the algorithm proceeds, and the quality of the clusters increases, we would expect MSB to increase, MSE to decrease, and F to increase.These statistics indicate that one has achieved the maximum between-cluster variation(as measured by MSB), compared to the within-cluster variation (as measured by MSE).

The flow chart diagram of the clustering system used in this paper is shown in Figure 1. During the clustering process, the number of cluster centre is selected and the initial cluster centre is set at random. The object is inserted closest to the cluster centre of the system and the new cluster centre is recalculated. The cluster based on smallest distance is created.

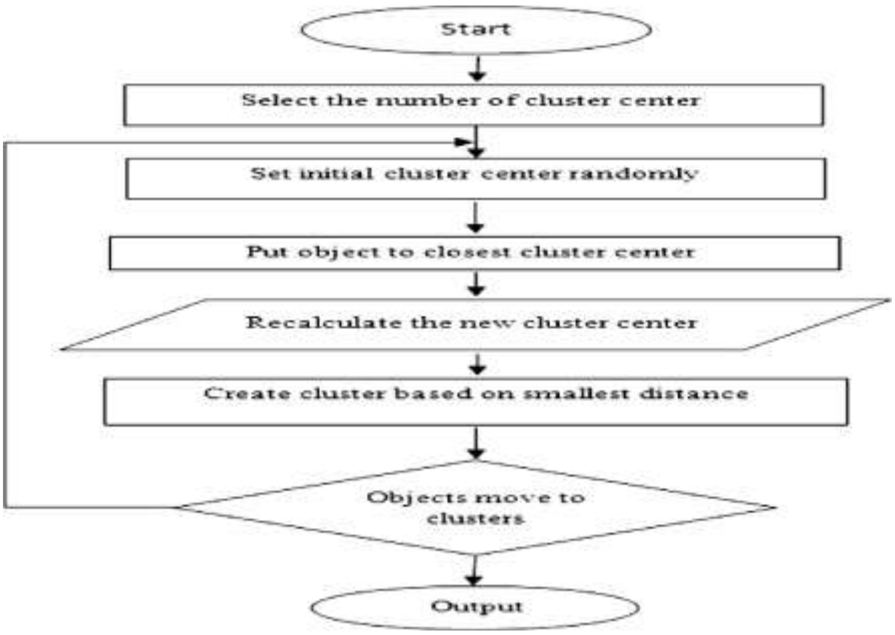

Figure 1: Flow Chart Diagram of Clustering.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1399-1403 www.ijaem.net ISSN: 2395-5252

III. MATERIALS AND METHOD

Materials

The materials used in this paper are grouped into the hardware requirements and software requirements. In the hardware requirements, a cluster of at least four machines is used and each machine has 200GB Ram and 1TB of Disk Space. The Systems used have Linux operating system with each node installed with latest version of java and latest version of Apache Hadoop. The other software requirements; Hyper Text Mark Up Language (HTML)/Cascading Style Sheets (CSS), Hypertext Preprocessor 8 (PHP 8), Ubutu Server v16.

Apache Hadoop is used to process the big data while YARN spreads the data across the cluster. The Hadoop Distributed File System (HDFS) is a distributed database where the data is saved to become accessible to YARN and Hadoop. Ubuntu Server v16 is used to host the program. HTML/CSS is used to present mined/analyzed data in the frontend.PHP 8, used to transfer mined/analyzed data from the server to the client. R,language for statistical analysis.

Method

The hardware cluster was created by connecting all the commodity hardware (3 Data Nodes and 1 Name node) in a Local Area Network (LAN). The commodity hardware consists of one Name node (200GB) and 3 Data Nodes (1TB) each. The data (is uploaded in the HDFS, the Name Node divides the data across the three Data Nodes, the HDFS performs data integrity and safety measures in order to preserve the integrity of either nodes failing. The files are also replicated across the cluster.

IV. DATA PRESENTATION AND ANALYSIS

In Table 1, the descriptive statistics for the clusters is presented.Table 2 shows the Cluster Centroids. Table 3 shows the Euclidean Distances between Cluster Centroids. Table 4 shows the p-value for the clusters. The data shown in Table 1 indicated that out of the fiveclusters used in the paper, cluster 2 has the lowest density of 402 and the minimum distance of 20492.981.

In Table 4, Mean Square Between Clusters (MSB) represents the between-cluster variation and Mean Square Error (MSE) represents the within-cluster variation. Thus, a “good” cluster would have a large value of the pseudo-F statistic, representing a situation where the between-cluster variation is large compared to the within-cluster variation and as the quality of the clusters increases, we would expect MSB to increase, MSE to decrease, and F to increase.

The pseudo-F method selects k = 2 for the preferred clustering solution. The smallest p-value occurs when k = 2, thus the preferred clustering solution.

In Figure 2, the data is loaded across by running a Hadoop job that checks the customer age and saves the total amount paid by this customer to be merged (reduced) with that of other customers of the same age group.

From Figure 3, in terms of revenue generated by age, it is important to note that customers between the ages of 66-80 were generating money more than all other age groups, this was followed by customers between 26 and 45 years.

Table 1: Descriptive Statistics for the Clusters

Number of clusters: 5	**Density of Clusters**	**Within cluster Sum of Squares**	**Average distance from Centroid**	**Maximum distance from Centroid**
Cluster1	10637	1.93656E+12	11764.448	84585.175
Cluster2	402	1.66461E+12	53278.165	57214.704
Cluster3	40485	3.77603E+12	8777.251	20492.981
Cluster4	13479	2.89512E+12	12439.202	186590.538
Cluster5	456	6.76204E+12	118745.953	289785.580

Table 2: Cluster Centroids

Variable	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Grand centroid**
Age	51.2180	49.4934	51.1403	51.3717	52.2587	51.1960
Rate	43.0549	40.8186	37.7261	43.3248	40.0690	39.7808
Units Consumed	336.1804	312.0965	285.8805	337.1285	302.8159	304.8936
Month Due	14523.9199	12845.6092	10851.1873	14650.5179	12217.4933	12252.6255

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1399-1403 www.ijaem.net ISSN: 2395-5252

Arrears	10434.0033	78984.3817	1654.6247	13546.0687	244793.6954	7561.7620
Vat	726.1960	642.2805	542.5561	732.4873	610.8747	612.6213
Total Due	25684.1191	92472.2713	13048.3468	28932.3377	257622.0634	20427.6677
Amount Paid	8459.4517	74726.7095	9405.8384	28358.5923	212506.4719	14857.0403
Arrears Deducted	0.0000	4994.6789	91.5921	1352.3541	15901.4930	467.5669
Arrears Balance	17224.6674	22740.2407	3733.8302	1925.9432	61017.0845	6037.9949

Table 3:Euclidean Distances between Cluster Centroids

	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**
Cluster1	0.0000	116659.1353	20812.7893	25536.3968	390551.9544
Cluster2	116659.1353	0.0000	130169.6568	104495.2930	274472.0682
Cluster3	20812.7893	130169.6568	0.0000	27789.3164	404618.3063
Cluster4	25536.3968	104495.2930	27789.3164	0.0000	378674.3629
Cluster5	390551.9544	274472.0682	404618.3063	378674.3629	0.0000

Table 4: P-value for the Clusters

Value of K	**MSB**	**MSE**	**Pseudo-F**	**p-value**
2	1.1533	0.4317	17.14	0.008
3	0.6532	0.407	15.51	0.012
4	0.4625	0.288	16.01	0.023
5	0.3597	0.0181	19.82	0.048

Fig 2: Power consumed by House Type.

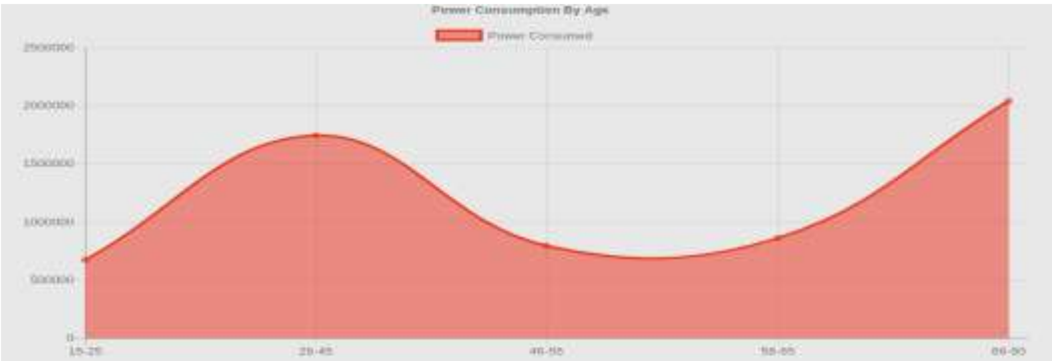

Fig 3: Power consumed by Age

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1399-1403 www.ijaem.net ISSN: 2395-5252

V. CONCLUSION

This paper has shown improving big data mining systems using Hadoop MapReduce technique.

The new system was be designed and built on Hadoop MapReduce.Instead of relying on expensive, proprietary hardware to store and process data, Hadoop enabledparallel distributed processing of large amounts of data on large clusters of commodity servers.

Hadoop has many advantages, and this feature make Hadoop particularly suitable for big data management and analysis. Hadoop can recover the data and computation failures caused by node breakdown or network congestion.

REFERENCES

[1]. Tranfield, D., Denyer, D., and Smart, P. 2013, "Towards a methodology for Developingevidence-informed management knowledge by means of systematic review,". British Journal of Management, March 2013.

[2]. Gandomi, A., and Haider, M., 2015,"Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management", April, 2015.

[3]. Larose Daniel T. and Larose Chantal D., 2015,"Discovering Knowledge in Data: An Introduction to Data Mining," pp. 4-5.

[4]. Borthakur, D., "The Hadoop Distributed File System: Architecture and Design", 2013,pp. 24- 39.

[5]. Wikipedia. (2013). Scientific Instrument [Online].Available: http://en.wikipedia. org / wiki/Scientific-instrument. Accessed 20th February 2019.

[6]. Gar, X., Jager, J., Kriegel, H.P., 2011, "A Fast Parallel Clustering Algorithm for Large Data", 2014, pp. 67-70.

[7]. Jiang, H., Chen, Y., Qiao, Z., Weng, T. H., & Li, K. C., 2015, " Scaling up MapReduce-based bigdata processing on multi-GPU systems. Cluster Computing",pp.369–383.

[8]. Larose Daniel T. and. Larose Chantal D.,2015, "Data Mining and Predictive Analysis," pp. 590-594.