*Available online www.jsaer.com*

***Journal of Scientific and Engineering Research*, 2021, 8(6):172-178**


**Research Article**

**ISSN: 2394-2630**
**CODEN(USA): JSERBR**

# Logs Data Reduction - Remove Unnecessary Data to Save Costs

**Aakash Aluwala**

Email: akashaluwala@gmail.com

**Abstract** This paper aims to deal with the reduction of log data by removing unnecessary data to save costs. Since logs are continuing to grow, developers and system administrators are becoming aware of automating analysis through data mining techniques to reduce unnecessary data. The paper concentrates on ways of reducing log data and aims at transforming natural language log data into organized log events using log parsers and Principal Component Analysis (PCA) techniques to reduce unnecessary data and save costs. In addition, these methods eliminate unnecessary data, thereby saving costs in managing log data while simultaneously increasing efficiency.

**Keywords** Logs Data, Principal Component Analysis, Log Parsers, Advanced Persistent Threats, Unnecessary Data, Save Costs.

## 1. Introduction

Many fields such as science and engineering have large sets of log data that are multi-dimensional, for which it is necessary to analyze them and identify the most significant characteristics [1]. The use of logs in system management for dependability assurance is quite common today since, in many cases, there is a lack of other information that specifies various behaviors of the system during its runtime in production. As the sizes of logs grow bigger constantly, developers and also operators plan to automate their examination with the help of data mining methods, which is why organized input data such as matrices are needed [10].

This led to the initiation of several studies in log parsing whose main goal is to convert the natural language log messages to events and reduce unnecessary data to save costs. However, these log parsers and the comparative benchmark data for these are not available in the public domain. Therefore, it is improbable that developers should be alert to the efficiency of the present log parsers and their drawbacks after applied practically. Also, to implement one and redesign it, or to develop one and then recreate it, which is rather inefficient [10].

Traditionally, researchers were engaged in trying to establish the existence of relationships in the pattern of the data whereby they undertake a series of simple plots of two variables at a time. However, the amount of such schemes that were compulsory for such a duty was n n!f2!(n- 2)! [2]. Unfortunately, this type of data analysis is not possible for big datasets. Therefore, the Principal Component Analysis (PCA) method was also incorporated which includes a sufficient number of the principal components according to variations of the data set that make up to 70−80% accuracy results by reducing the unnecessary data and improving efficiency [2]. Therefore, the purpose of this paper is to analyze through secondary qualitative method for logs data reduction techniques by removing unnecessary data to save costs.

## 2. Literature Review

Today's organizations face APT (Advanced Persistent Threats) attacks, which are typically multi-staged and covert. To mitigate these outbreaks, creativities frequently depend on causalities exploration of organization movement data obtained through pervasive systems observing to classify the first idea of access. Nevertheless,

the existence of huge amounts of unnecessary data generated by all-pervasive system monitoring is a great challenge for causality analysis and it is too expensive to host such vast quantities of data [12].

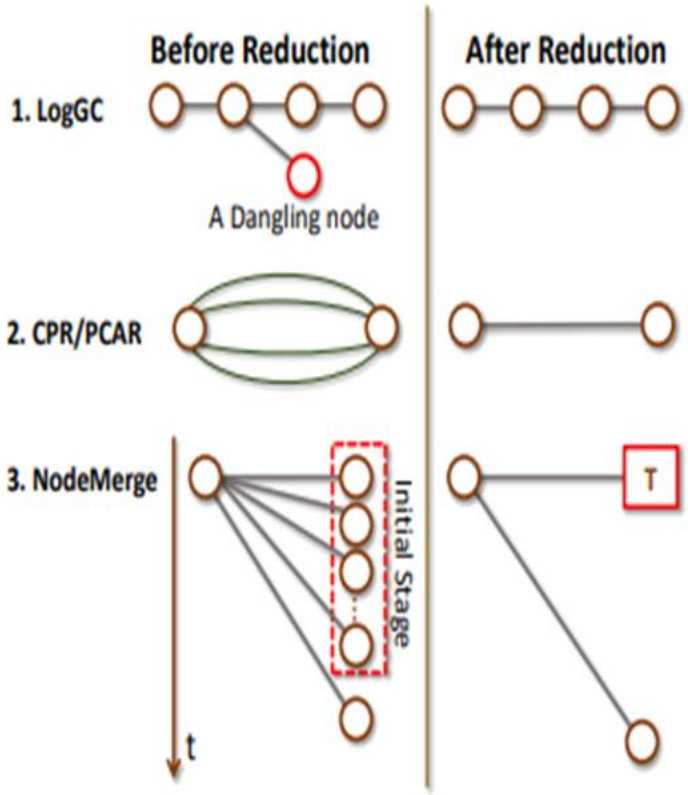


*Figure 1: NodeMerge Reduction Technique*

(Source: Tang et al., 2018) [12]

Therefore, the above figure shows that adopted techniques have reduced the storage amount required for causality analysis while retaining high causality quality. Through the secondary data source, it was found that template-based online system event storage named NodeMerge is a solution for this issue which has worked with the creek of addiction data within operating systems and reduces file read-only actions based on their admittance designs. A similar budget to save costs has helped in reducing expenses and improved performance during cause-effect examination. Since the adjustable quantity of resources was available for the unnecessary data reduction it has preserved causality analysis and reduced log data form to save costs.

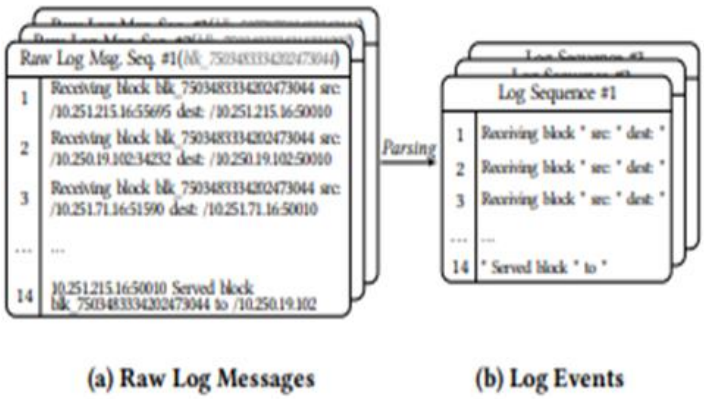


*Figure 2: Raw Log Messages and Log Events*

(Source: Zhang et al., 2019) [13]

It can be seen from the above figure that logs store a massive amount of information (for example, events, parameters, and execution details) on the running status of a software-intensive system, and they are essential in maintaining online service systems. After a problem/anomaly arises, engineers mostly use logs for more elaborate analysis of the unnecessary data [13].

**3. Monitoring Tools Impacted**

Log data reduction techniques for unnecessary data to save costs affect all the monitoring tools in system administration. Tools like Spelunk get faster record processing and storage when log sizes are reduced. They can give more accurate and up-to-date views of the system and security when they don't have to process and store as much data. PCA and advanced log parsers in these tools also help detect anomalies and events, so the system is more reliable and operational.

**4. Tasks**

The following task for reducing unnecessary data through log parsing is executed in five stages/phases of Parallel Log Parsing and the five numbered arrows are the collaborations between the key driver and the flash cluster. The foremost database is consecutively within the Spark driver, which equals the coordinator in the Spark cluster, where it distributes the Spark tasks to the workers in the following steps. Step 1 of a POP Spark application in general contains the following for a text file, where the objective is to read the log data from a distributed case scheme as an RDD. Then use a map function with the input 'describe the preprocessing logic in terms of what operation is to be done on the single log message'.

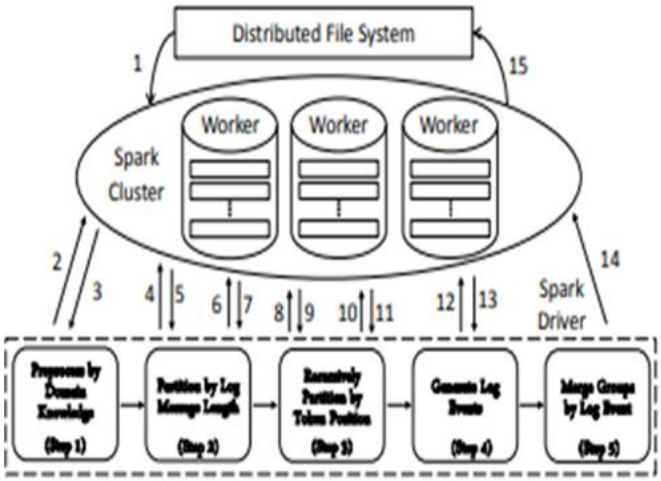


*Figure 3: Parallel Log Parsing Implementation*

(Source: He et al., 2017) [10]

On the one hand, it stores the log messages produced by the reprocessed log in reminiscence and revenues an RDD as an orientation. Then in stage 2, by using the combined function, all the distance of the possible log message is to be calculated and the result is to be displayed in the form of a list. Following this, in step 3 each RDD researcher used combined to constitute the token sets for all the token positions (arrow 8∼9). According to different demonstration sets and the specific thresholds defined earlier, the driver program determines whether the currently obtained RDD should be further partitioned and then uses a screen to create new RDDs and put these new RDDs into the RDD list (solid line 10~11).

Otherwise, delete it and in the same fashion pass the RDD to phase 4 of the general result. Step 4 of the process involves a reduction to create log events for all the RDDs which is represented by the projectile 12∼13. After all log happenings have been obtained, and were extracted, POP applies categorized grouping to them in the foremost program. then merge them using the clustering result and use union to combine the RDDs (arrow 14). Last, after merging operations, the resulting RDDs are saved as a text file into the distributed file system (arrow 15) and ultimately reduce the unnecessary data.

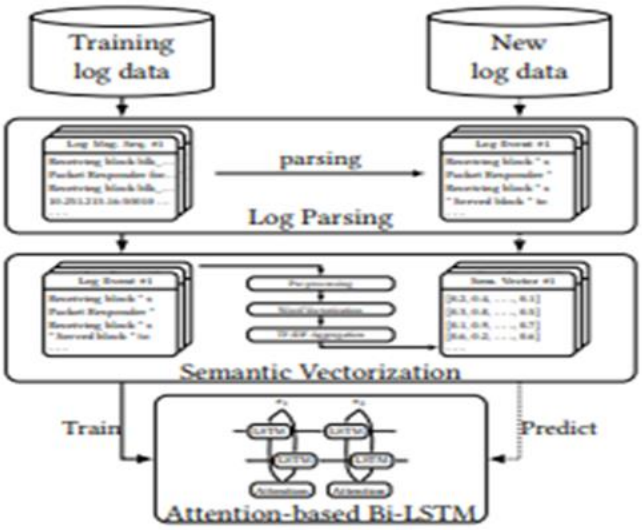


*Figure 4: Unnecessary Data Detection by Log Parsing*

(Source: Zhang et al., 2019) [13]

The above figure displays how LogRobust utilizes the attention-based Bi-LSTM neural network to identify the anomalies, which adopted the contextual knowledge of log sequence and learned the function of assigning different weights/importance of log events. Thus, the approach of a researcher was to manage unstable log sequences. Since LogRobust requires the resolution of each log message to obtain its log event through the abstraction of the parameters in the message. Thus, the message logs turn into usable data that can be utilized in further analysis. Based on precision and speed, LogRobust uses the Drain method for log parsing. Also, the drain is proposed and describes the approach based on its rather high capacity for parsing and working speed to remove unnecessary data [13].

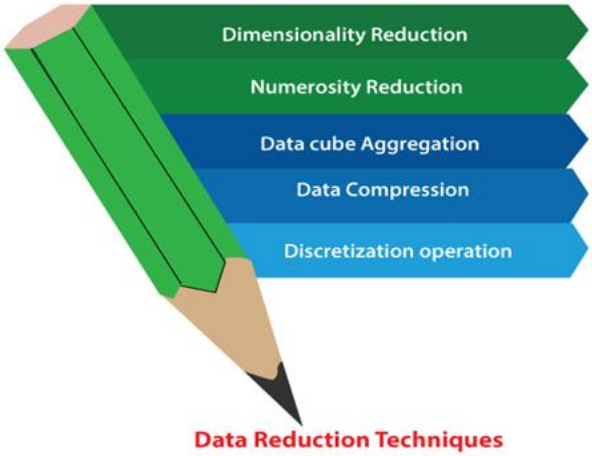


*Figure 5: Data Reduction Technique*

(Source: JavaPoint, 2020) [11]

The above figure shows the analysis of the data reduction approaches that serve to decrease the volume of primary data and offer its representation in a significantly lesser quantity. There are certain approaches commonly employed in data reduction as shown in the figure, can provide a smaller data set that is equivalent to the original in terms of quality.

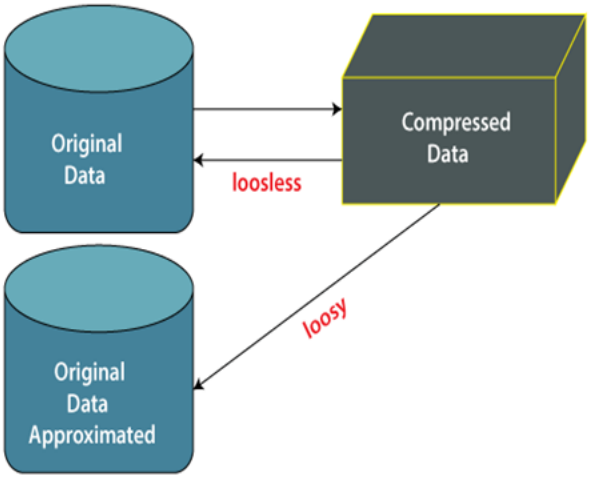


*Figure 6: Data Compression*

(Source: JavaPoint, 2020) [11]

The above figure shows that compression can be done by altering the structure of data in a way that requires less storage space such as modifying, encoding, or converting. Data compression involves arriving at an efficient summary of information after excluding the least significant details and encoding data in binary format. Data that can be restored with high precision after it has been compressed is referred to as Lossless compression. On the other hand, the method in which it is impossible to retrieve the uncompressed format from the compressed format is termed Lossy compression. Data compression is another use of dimensionality and numerosity reduction methods.

**5. Solution and Implementation**

To reduce unnecessary data and save costs, many researchers have implemented various techniques to get solutions in 2015, Qiao implemented rank of the factorization and provided Singular Value Decomposition (SVD) to obtain Non-Negative Matrix Factorization (NMF) algorithms [3]. Then, in 2015, due to the problem of a large amount of data from the ambulatory system, Kumar et al. introduced an ECG sign compression procedure. Also, they have employed SVD, and wavelet difference reductions as the major components of their algorithm, and the compression capability that the algorithm generated ranged from as low as 7% to as high as '21%" 4:1 of which is the quality of signal reconstruction was very high [4].

They also implemented their method on actual data such as Diabetic, Waveform, human movement recognition using smartphone, and thyroid data sets and obtained good results. In addition, the method of Hadamard-based random projection was proposed by Menon et al in 2016 with the help of the fast SVD algorithm called FSVD obtained results of the experiments and proved that their new algorithm was more effective than Gaussian-based FSVD for dimensionality reduction when used in hyperspectral arrangement [1,7].

In addition, the Compression of encrypted images based on Distinct Wavelet Alter, SVD, plus Huffman coding was explained by Kumar and Manoj in the year 2017 [8]. According to the reduction and concentration structure, Olive cast off conventional PCA to view the drop and concentration depending on linear uncorrelated groupings of the advanced variables in 2017 [5]. He applied PCA in reduction as well as analysis of log data. In the year 2018, Feng et al. introduced the tensor SVD used for performing reduction on cloud cyber data [6].

**6. Results**

The result was extracted by creating models with the help of the very efficient above-mentioned methods. Thus, when analyzing large datasets and orienting data with a huge sum of variables, it has become necessary to decrease the sum of variables and to construe data in terms of linear blends of the observations. For this, Principal Component Analysis (PCA) is a procedure that promotes a manner of changing most possibly related

variables into other variables that can be explained mathematically [3]. The new basis helps in filtering out unnecessary data provides the inherent structure of the reduced dataset from a large dataset and saves costs.
However, it is used in a variety of contexts as a consequence of a large assortment of data. PCA applied a mathematical operation in a vector space transform to the reduction of log data size to save costs. Moreover, these outcomes indicate that adopting log data compression techniques substantially improves the effectiveness and economy of log management. From the faced approaches such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) in this study, it is established that the important information can be preserved and with equal accuracy, the amount of data can be reduced to as low as 70-80%. In addition, methods like NodeMerge also help in decreasing storage-related demands because data is stored according to the leading access patterns. If these methods are applied in monitoring tools, then it is likely that data processing time may be made faster, the tool's ability to detect anomalies made even better, and the costs of storage may further be reduced. Furthermore, the study established how reduction techniques by reducing unnecessary data can increase the practicality of the logs in dependability assurance and other performance monitoring elements; therefore, making the management of big data log files in various sectors feasible through the application of the technique.

## 7. Conclusion

In this paper, the efforts aimed at cutting down on unnecessary data and saving costs by handling immense log data which are crucial when handling the increasing log datasets in science and engineering disciplines. When it comes to defining what information to use and what information is redundant in the context of log file volumes important aspects are conserved through Deep compression algorithms, Principal Component Analysis (PCA), and Singular Value Decomposition (SVD). Besides, it helped to reduce storage space and improved the efficiency of the system management since dependability is provided.
This has gone a long way in making it easier to analyze and interpret such large datasets as evident from the significant features of data manifested after the text log messages had been converted into structured events through the use of PCA for dimensionality reduction. Besides, the practical approach, like NodeMerge, is an excellent example of how the storage space could be minimized, and correspondingly, the storage requirements would be reduced while, at the same time, avoiding the decrease in the quality of the causality analysis. The above techniques have practical implications that enhance the effective and efficient utilization of log data in authentic business environments. The extension of the above procedures is that more research will help to refine these procedures and, therefore, design new methods to improve the results of log data reduction in the future.

## References

[1]. R. Houari, A. Bounceur, M.-T. Kechadi, A.-K. Tari, and R. Euler, "Dimensionality reduction in data mining: A Copula approach," *Expert Systems with Applications*, vol. 64, pp. 247-260, 2016.
[2]. N. Salem and S. Hussein, "Data dimensional reduction and principal components analysis," *Procedia Computer Science*, vol. 163, pp. 292-299, 2019.
[3]. H. Qiao, "New SVD-based initialization strategy for non-negative matrix factorization," *Pattern Recognition Letters*, vol. 63, pp. 71-77, 2015
[4]. R. Kumar, A. Kumar, and G. K. Singh, "Electrocardiogram signal compression based on singular value decomposition (SVD) and adaptive scanning wavelet difference reduction (ASWDR) technique," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 12, pp. 1810-1822, 2015.
[5]. D. J. Olive, "Principal component analysis," in *Robust multivariate analysis*: Springer, 2017, pp. 189-217.
[6]. J. Feng, L. T. Yang, G. Dai, W. Wang, and D. Zou, "A secure high-order Lanczos-based orthogonal tensor SVD for big data reduction in a cloud environment," *IEEE Transactions on Big Data*, vol. 5, no. 3, pp. 355-367, 2018.
[7]. Menon, Vineetha, Qian Du, and James E. Fowler. "Fast SVD with random Hadamard projection for hyperspectral dimensionality reduction." IEEE Geoscience and Remote Sensing Letters 13.9 (2016): 1275-1279.

[8]. Kumar, Manoj, and Ankita Vaish. "An efficient encryption-then-compression technique for encrypted images using SVD." Digital signal processing 60 (2017): 81-89

[9]. Xu, Zhang, et al. "High fidelity data reduction for big data security dependency analyses." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016.

[10]. He, Pinjia, et al. "Towards automated log parsing for large-scale log data analysis." IEEE Transactions on Dependable and Secure Computing 15.6 (2017): 931-944.

[11]. JavaPoint. Data Reduction in Data Mining 2020. Retrieved from: https://www.javatpoint.com/data-reduction-in-data-mining.

[12]. Tang, Yutao, et al. "Nodemerge: Template-based efficient data reduction for big-data causality analysis." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

[13]. Zhang, Xu, et al. "Robust log-based anomaly detection on unstable log data." Proceedings of the 2019 27th ACM joint meeting on European Software Engineering Conference and symposium on the foundations of software engineering. 2019.