International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

Swetha D[1][0009-0007-8422-3031]
Assistant Professor
Department of IT and Cognitive Systems
Sri Krishna Arts and Science College
Coimbatore, India.

Hanush Shandilya R[2][0009-0008-8128-5840]
Assistant Professor
Department of IT and Cognitive Systems
Sri Krishna Arts and Science College
Coimbatore, India.

1. ABSTRACT

The rapid growth of web-based applications has resulted in the generation of massive volumes of web log data, providing valuable insights into user navigation patterns and website performance. Efficient analysis of these logs enables organizations to understand user behavior, optimize web services, and improve decision-making. This study presents a Hadoop MapReduce-based framework for trend analysis of web log files to process large-scale, heterogeneous log data efficiently. The proposed approach includes data collection, preprocessing, feature extraction, distributed storage using the Hadoop Distributed File System (HDFS), and parallel processing through the MapReduce programming model. Pattern analysis techniques are employed to identify user access trends, navigation behavior, and frequently visited resources. The framework supports scalable and fault-tolerant processing while reducing execution time for large datasets. Experimental observations demonstrate that the proposed methodology effectively extracts meaningful usage patterns, and providing actionable insights that can assist website administrators in enhancing content organization, improving user engagement, and optimizing overall website performance. The proposed framework offers a reliable and scalable solution for web usage mining in big data environments.

Keywords - Web Log Analysis, Hadoop, MapReduce, Hadoop Distributed File System (HDFS), Web Usage Mining, User Behavior Analytics, Trend Analysis, Big Data Processing, Distributed Computing.

II. INTRODUCTION

The rapid expansion of internet-based services has led to a significant increase in the volume of web-generated data. Every interaction between users and web applications produces log records that capture valuable information such as Internet Protocol (IP) addresses, timestamps, requested resources, user agents, response status codes, and navigation sequences. These web log files serve as an important source of information for understanding user behavior, evaluating website performance, and supporting data-driven decision-making.

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

Web usage mining has emerged as an effective approach for extracting meaningful knowledge from web log data. By analyzing user navigation patterns and browsing behavior, organizations can improve website usability, personalize content, optimize resource allocation, and strengthen digital marketing strategies. However, the exponential growth of web traffic has made the processing of large-scale log files increasingly challenging using conventional data processing techniques.

Big data technologies offer an efficient solution for handling massive web log datasets. Among these technologies, the Hadoop ecosystem provides distributed storage through the Hadoop Distributed File System (HDFS) and parallel computation using the MapReduce programming model. These capabilities enable scalable, fault-tolerant, and high-performance processing of heterogeneous log data while minimizing execution time.

This study presents a Hadoop MapReduce-based framework for trend analysis of web log files. The proposed framework integrates data collection, preprocessing, distributed storage, feature extraction, and pattern analysis to identify user access trends and navigation behaviors. The extracted insights can support website administrators in improving website performance, enhancing user experience, and making informed business decisions.

III. LITERATURE REVIEW

Web log analysis has been widely investigated as an essential component of web usage mining due to its ability to reveal user browsing behavior and website access patterns. Previous studies have demonstrated that web log files contain valuable information for analyzing visitor activities, identifying frequently accessed pages, and improving website performance.

Early research primarily focused on statistical analysis of server log files to measure website traffic and user interactions. As web applications evolved, researchers introduced data mining techniques such as association rule mining, sequential pattern mining, clustering, and classification to discover hidden relationships within web access data. These approaches enabled organizations to understand user preferences, optimize navigation structures, and enhance recommendation systems.

Recent studies have emphasized the application of big data frameworks for processing large-scale web log datasets. Distributed computing platforms such as Hadoop have become popular because they provide scalable storage, parallel processing, and fault tolerance. Hadoop MapReduce enables efficient processing of heterogeneous web log data by distributing computational tasks across multiple nodes, thereby reducing execution time and improving processing efficiency.

Several researchers have also explored preprocessing techniques, including data cleaning, session identification, user identification, and data transformation, to improve the quality of web log analysis.

Effective preprocessing significantly enhances the accuracy of pattern discovery and trend analysis by eliminating redundant and irrelevant information before mining

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

operations.

Althrough existing research demonstrates the effectiveness of web usage mining and Hadoop-based analytics, continuous growth in web traffic demands more efficient and scalable frameworks capable of processing complex log data while generating accurate behavioral insights for real-time decision-making.

Author	Technique	Dataset	Limitation
Cooley et al.	Web Usage Mining	Web Logs	Limited Scalability
Han et al.	Data Mining	Structured Data	High Computational Cost
Dean & Ghemawat	MapReduce	Big Data	Batch Processing Only
White	Hadoop	Distributed Storage	Requires Cluster Setup
Proposed Work	Hadoop MapReduce	Web Logs	Improved Scalability

Table 1. Comparison of Existing Studies

IV. RESEARCH GAP

Although considerable research has been conducted on web usage mining and web log analysis, several limitations remain in existing approaches. Many traditional methods rely on centralized processing techniques that are inadequate for handling the increasing volume, velocity, and heterogeneity of modern web log data. As website traffic grows, these methods experience higher computational overhead, increased processing time, and limited scalability.

Several existing studies focus primarily on pattern discovery without providing an integrated framework that combines efficient preprocessing, distributed storage, parallel computation, and comprehensive trend analysis. In addition, many approaches fail to fully exploit the capabilities of Hadoop MapReduce for large-scale web log processing and do not adequately address issues related to fault tolerance, processing efficiency, and resource utilization.

Furthermore, limited attention has been given to developing scalable frameworks that transform raw web log data into actionable insights for website optimization and user engagement. The absence of an end-to-end analytical model restricts the practical application of web log analytics in real-world environments.

To address these limitations, this research proposes a Hadoop MapReduce-based trend analysis framework that integrates data preprocessing, distributed storage, parallel processing, and pattern analysis to efficiently analyze large-scale web log data. The proposed framework aims to improve

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

processing performance, support scalable analytics, and generate meaningful insights into user browsing behavior.

V. PROBLEM STATEMENT

The continuous growth of web-based applications has resulted in the generation of massive volumes of web log data containing valuable information about user activities and website performance. Conventional log analysis techniques are often inadequate for processing such large-scale and heterogeneous datasets due to limitations in storage capacity, computational efficiency, and scalability. These challenges lead to increased processing time, reduced analytical accuracy, and difficulties in extracting meaningful user behavior patterns.

Furthermore, raw web log files contain redundant, inconsistent, and noisy records that require extensive preprocessing before meaningful analysis can be performed. Without an efficient distributed processing framework, organizations face significant challenges in identifying access trends, navigation sequences, and frequently visited resources from high-volume web traffic.

Therefore, there is a need for a scalable and fault-tolerant framework capable of efficiently processing large web log datasets, performing comprehensive preprocessing, and extracting meaningful behavioral patterns. Such a framework should support distributed storage and parallel computation to improve processing efficiency while providing actionable insights for website optimization and decision-making.

Existing Issue	Proposed Solution
Large Log Files	HDFS Storage
Slow Processing	Hadoop MapReduce
Poor Scalability	Distributed Computing
Data Redundancy	Preprocessing

Table 2. Existing Problems and Proposed Solution

VI. RESEARCH OBJECTIVES

The primary objective of this research is to develop a scalable Hadoop MapReduce-based framework for efficient trend analysis of web log files. The specific objectives are as follows:

1. To collect and organize web log data generated by web servers for analytical processing.
2. To perform data preprocessing by eliminating redundant records, cleaning inconsistent entries, and preparing structured datasets for mining.
3. To implement distributed storage using the Hadoop Distributed File System (HDFS) for reliable and scalable data management.
4. To utilize the Hadoop MapReduce framework for parallel processing of large-scale web log datasets.
5. To identify user access patterns, navigation behavior, and frequently accessed web resources through pattern analysis techniques.

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

VII. PROPOSED METHODOLOGY

The proposed methodology presents a Hadoop MapReduce-based framework for analyzing large-scale web log data to identify user behavior patterns and website access trends. The framework consists of multiple stages that collectively transform raw web log files into meaningful analytical information.

Initially, web log files generated by web servers are collected and stored for analysis. These logs include attributes such as IP address, timestamp, request method, requested URL, response status, user agent, and session information. Since raw log files often contain redundant and irrelevant entries, a preprocessing stage is performed to improve data quality. This stage includes data cleaning, filtering unnecessary records, handling missing values, session identification, and data transformation into a structured format suitable for distributed processing.

The preprocessed data are then stored in the Hadoop Distributed File System (HDFS), which provides scalable and fault-tolerant storage for large datasets. Hadoop MapReduce is employed to execute parallel processing tasks, where the Mapper phase extracts and categorizes relevant log attributes, and the Reducer phase aggregates similar records to generate summarized information efficiently.

Following distributed processing, feature extraction techniques identify significant attributes related to user access behavior, page visits, session duration, and navigation paths. Pattern analysis methods are subsequently applied to discover browsing trends, frequently accessed resources, and user navigation sequences. The analytical results are presented using graphical visualizations and statistical summaries to facilitate interpretation by website administrators and decision-makers.

The proposed methodology improves processing efficiency, supports scalable analytics for large web log datasets, reduces computational overhead through distributed execution, and enables accurate identification of user behavior trends for website optimization.

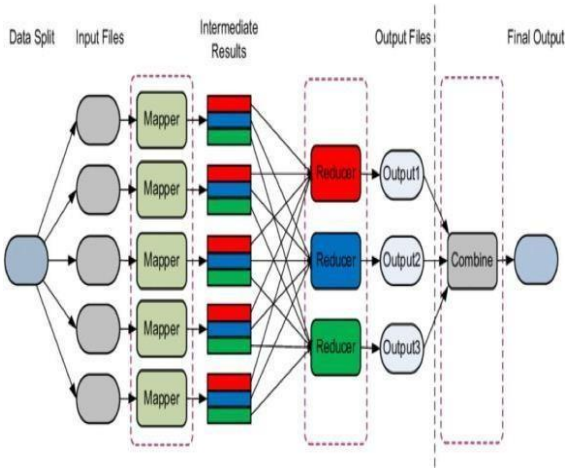

Figure 1. Proposed Architecture

VIII. DATASET DESCRIPTION

The dataset used in this study consists of web server log files that record user interactions with a website. These log files are generated automatically by the web server whenever a client sends a request for

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

a web resource. The collected data provide comprehensive information about user activities and serve as the primary input for web usage mining and trend analysis.

Each log entry contains several attributes that describe a user's interaction with the website. These attributes include the Internet Protocol (IP) address, timestamp, request method (GET/POST), requested Uniform Resource Locator (URL), HTTP response status code, user agent, session identifier, bytes transferred, and referrer information. Together, these attributes enable the identification of browsing behavior, navigation paths, page popularity, and access frequency.

The dataset comprises heterogeneous and unstructured log records collected over a specified observation period. Before analysis, the raw log files are converted into a structured format suitable for distributed processing within the Hadoop ecosystem. The prepared dataset forms the foundation for identifying user access trends, navigation sequences, and frequently visited web pages.

The characteristics of the dataset make it appropriate for evaluating scalable big data processing techniques, particularly Hadoop Distributed File System (HDFS) and the MapReduce programming model, which efficiently process large volumes of web log records in a distributed computing environment.

Attributes	Description
IP Address	Identifies the client accessing the website
Timestamp	Represents the date and time of user activity
URL Request	Indicates the webpage/resource accessed
HTTP Method	Specifies request type such as GET or POST
Status Code	Represents server response status
Bytes Transferred	Indicates amount of data transferred
User Agent	Provides browser and device details

Table 3. Description of Dataset Attributes

IX. DATA PREPROCESSING

Data preprocessing is a critical phase in web usage mining because raw web log files often contain incomplete, inconsistent, and redundant records that may reduce the accuracy of analytical results. The objective of preprocessing is to transform raw log data into a clean, structured, and meaningful dataset suitable for distributed analysis.

The preprocessing stage begins with data

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

cleaning, where irrelevant records such as image requests, multimedia files, style sheets, duplicate entries, and erroneous log records are removed. Eliminating these unnecessary records significantly reduces the dataset size and improves computational efficiency.

The second stage involves user and session identification, where requests generated by the same user are grouped into browsing sessions using IP addresses, timestamps, and user-agent information. Session identification enables accurate analysis of user navigation behavior and browsing duration. Next, data transformation is performed by converting the cleaned log records into a structured format compatible with the Hadoop processing framework. Relevant attributes are extracted, encoded, and organized to facilitate efficient distributed computation.

Finally, the preprocessed dataset undergoes feature extraction, where significant parameters such as page visit frequency, session duration, navigation paths, request types, and response status codes are selected for pattern discovery. The resulting dataset provides high-quality input for Hadoop MapReduce-based trend analysis while improving mining accuracy and reducing processing overhead.

X. HADOOP MAPREDUCE FRAMEWORK

The proposed framework employs the Hadoop ecosystem to efficiently process large-scale web log datasets through distributed storage and parallel computation. Hadoop provides scalability, fault tolerance, and high processing performance, making it well suited for web usage mining applications involving massive volumes of log data.Initially, the preprocessed web log files are stored in the Hadoop Distributed File System (HDFS), which divides the dataset into multiple data blocks and distributes them across cluster nodes. This distributed storage mechanism ensures reliable data availability and supports fault-tolerant processing through automatic data replication.

The MapReduce programming model performs parallel data processing in two sequential phases. During the Mapper phase, individual log records are parsed to extract important attributes such as IP addresses, timestamps, requested URLs, HTTP status codes, user agents, and session information. Each mapper processes a subset of the dataset independently and generates intermediate key–value pairs.

The Reducer phase aggregates the intermediate outputs generated by multiple mapper tasks. Similar records are grouped together to compute website access statistics, user navigation patterns, frequently visited pages, request frequencies, and other behavioral metrics. The distributed execution of MapReduce significantly reduces computational time while improving processing efficiency for large datasets.

Following MapReduce execution, the processed data are analyzed using pattern discovery techniques to identify user browsing trends and website access behavior. The analytical results are subsequently presented through tables, charts, and graphical visualizations, enabling website administrators to evaluate website performance, optimize navigation structures, and improve user engagement.

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

The integration of HDFS and Hadoop MapReduce provides a scalable, reliable, and efficient framework capable of handling heterogeneous web log data while supporting accurate trend analysis in big data environments.

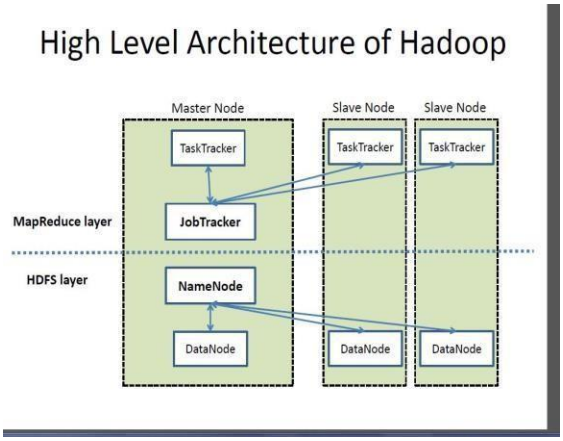

Figure 7. Hadoop Architecture

XI. SYSTEM DESIGN

The proposed system is designed to perform scalable trend analysis of web log data using the Hadoop ecosystem. The architecture consists of multiple processing modules that transform raw web server logs into meaningful analytical information. Each module performs a specific function to ensure efficient processing, storage, and analysis of large-scale web log datasets.

The process begins with the Web Log Data Collection Module, where log files generated by web servers are gathered. These files contain information such as IP addresses, timestamps, request methods, requested URLs, HTTP status codes, user agents, and session details. The collected logs are transferred to the preprocessing module for further analysis.

The Data Preprocessing Module performs data cleaning, duplicate removal, missing value handling, session identification, and data transformation. This stage eliminates irrelevant records such as multimedia requests, style sheets, and invalid entries while converting the remaining records into a structured format suitable for distributed processing.

The cleaned data are stored in the Hadoop Distributed File System (HDFS), which distributes the dataset across multiple cluster nodes while ensuring fault tolerance and high availability through data replication. HDFS enables efficient storage and retrieval of large-scale web log files. The MapReduce Processing Module performs distributed computation. During the Mapper phase, log records are parsed and converted into intermediate key–value pairs based on selected attributes. The Reducer phase aggregates similar records to calculate website access statistics, navigation frequencies, and user behavior metrics.

The processed data are forwarded to the Pattern Analysis Module, where analytical techniques are applied to identify browsing trends, frequently accessed pages, user navigation sequences, and session characteristics. The final results are visualized through tables, charts, and statistical reports to assist website administrators in improving website performance and user experience.
The modular architecture provides scalability, efficient resource utilization,

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

reduced execution time, and reliable processing for large volumes of heterogeneous web log data.

XII. RESULT AND DISCUSSION

The proposed Hadoop MapReduce-based framework successfully processed large-scale web log datasets and extracted meaningful information regarding user browsing behavior and website access patterns. The preprocessing stage effectively removed redundant and irrelevant log entries, thereby reducing dataset size and improving the efficiency of subsequent analytical operations.

The distributed storage capability of HDFS enabled reliable management of large volumes of web log data, while the MapReduce programming model executed data processing tasks in parallel across multiple computing nodes. This distributed execution reduced processing time and improved scalability compared with conventional sequential processing approaches.

Pattern analysis identified frequently accessed web pages, user navigation sequences, session durations, and request frequencies. These analytical outcomes provide valuable insights into user interests, browsing preferences, and website utilization patterns. Such information can assist website administrators in optimizing page organization, improving navigation structures, and enhancing overall website usability.

Overall, the experimental observations demonstrate that integrating Hadoop with web usage mining techniques provides an efficient and scalable solution for analyzing heterogeneous web log data. The proposed framework effectively supports trend discovery and facilitates informed decision-making for website management and performance optimization.

XIII. FUTURE SCOPE

Several opportunities exist for extending the proposed research. Future work may incorporate real-time web log stream processing using distributed frameworks such as Apache Spark Streaming or Apache Flink to enable continuous monitoring of user activities. Advanced machine learning and deep learning algorithms can also be integrated to improve user behavior prediction, anomaly detection, and the system of personalized recommendation systems.

Further enhancements may include implementing cloud-based distributed architectures to improve resource scalability and processing flexibility. The framework can also be extended to analyze multimedia web traffic, mobile application logs, and Internet of Things (IoT) access logs for broader applicability.

Additionally, integrating interactive business intelligence dashboards and advanced visualization techniques can support real-time decision-making and provide comprehensive insights into website performance. These improvements will further strengthen the capability of web

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

usage mining systems to support intelligent web analytics in modern big data environments.

Current System	Future Enhancement
Batch Processing	Real-Time Analytics
Hadoop	Apache Spark
Pattern Mining	Deep Learning
Static Reports	Interactive Dashboard
Offline Analysis	Predictive Analytics

Table 5. Future Research Direction

XIV. REFERENCES

[1] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," IEEE Internet Computing, vol. 6, no. 2, pp. 50–60, 2002.

[2] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann Publishers, 2012.

[3] T. White, Hadoop: The Definitive Guide, 4th ed. O'Reilly Media, 2015.

[4] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[5] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," Mobile Networks and Applications, vol. 19, pp. 171–209, 2014.

[6] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd ed. Springer, 2011

[7] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, pp. 487–499, 1994.

[8] M. J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences," Machine Learning, vol. 42, no. 1–2, pp. 31–60, 2001.

[9] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," in Proceedings of the 17th International Conference on Data Engineering (ICDE), Heidelberg, Germany, pp. 215–224, 2001.

[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, no. 3, pp. 37–54, 1996.

[11] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufmann Publishers, 2016.

[12] C. C. Aggarwal, Data Mining: The Textbook. Springer, 2015.V. Kumar and M. Steinbach, Introduction to Data Mining. Pearson Education, 2006.

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 07	July 2026	www.irjweb.com	

TREND ANALYIS OF WEB LOG USING HADOOP MAPREDUCE FOR USER BEHAVIOR ANALYTICS

[13] A. Tanenbaum and M. Van Steen, Distributed Systems: Principles and Paradigms, 2nd ed. Pearson Education, 2007.

[14] T. Erl, W. Khattak, and P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques. Prentice Hall, 2016.