A Complete Research on Techniques & Technologies of Big Web Data Preparation to Web User Usage Behaviour

N. Silpa, V. V. R. Maheswara Rao

***Abstract*— The rapid advancements in data digitization, the most powerful inventions of learning methodologies in data collection and reduced cost of data storage further enabled the World Wide Web with immense amount of data at significant rate in all the key domains. The generated web data is non-scalable, high dimensional, widely distributed, heterogeneous, dynamic in nature and having useful insights, and thus, it evolved as big data. This situation creates inevitably increasing opportunities in extracting structured solutions from unstructured weblog data for the present big data researchers. Moreover, to provide value addition to any key domain and derive actionable knowledge for various applications, such as, web usage analysis for improvements in fraud detection, product analysis and customer segmentation, got the focus in big data era by the web analysts. To improve operational performance and to discover hidden insights accurately, a comprehensive process is required to investigate the web user usage behavior by analyzing big web data.**

Towards this, the authors concentrate on reviewing the techniques and technologies of web data collection and preparation for investigating web user usage behavior effectively. In the present paper, the researchers initially pay an attention to explore web log data preparation methods in the traditional approach. Later, the review emphasizes on Hadoop approach for big data preparation and processing. This approach able to concentrate comprehensively on both the stages: distributed data storage and parallel processing of weblog data and to leverage the strengths of techniques and technologies of individual stages. Moreover, the authors deliberately review the possible potential research paths that results in an improved methodologies for data storage and optimized processing speed in the era of big web data.

***Keywords*— Big Data Analytics, Web Data, Hadoop, HDFS, MapReduce, Web Analytics, Web User Behavior, Web Data Preparation.**

I. INTRODUCTION

With the increased usage of World Wide Web, the WWW becomes a vast repository which is unstructured, unlabeled, noisy, redundant and less reliable data on numerous web servers [7, 22, 23, 39] over wide geographical regions. In addition, web services and web enabled systems [33, 40] are exponentially growing. The user usage data reached to astronomical proportions and it is heterogeneous, non-scalable, distributed and incremental nature of the weblog.

Web Mining is evolved by inspiring the techniques of data mining [40, 41, 43] in order to analyze the large volume of digital data generated by web applications and retrieves valuable hidden insights. The task of web mining can be performed for analyzing the content of webpage, structure of website and click stream data recorded in the weblog [7, 40, 41, 45], that is depicted in the Figure 1.

The techniques of web mining are applied on the semi-structured content of webpage including text, HTML tags, XML documents, forms, tables and images to retrieve the more relevant data against to the web user query. Web content mining empowers the ability of search engines to meet the expectations of web user.

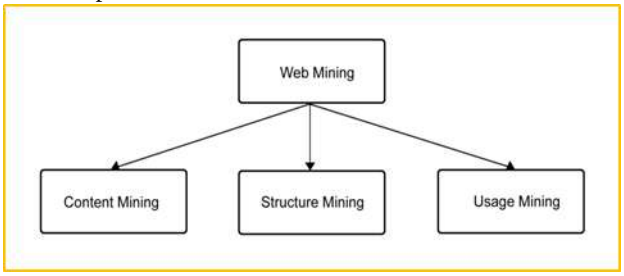

Figure 1. Web Mining Types

In order to find-out the valuable association among the web pages it is necessary to employ the techniques of web mining. This knowledge is useful to the website administrator in order to develop the effective design of website.

Another important application of web mining is to find the interesting patterns in the web user usage data. The hidden knowledge of browsing patterns models the behavior of web user that can be helpful for all key domains to meet the demands of web user.

On the other hand, in the technology perspective, the WWW leverages new technologies in the Industries and also evolves the corporate with infinite paths to attract the attention of the stakeholders. To attain this, Big data is evolved [1, 2, 5, 8, 21] with a capacity of high data storage in terms of Volume, the capability to process high speed generation of digital data in terms of Velocity and deal with unstructured, semi structured, quasi structured and structured data in terms of Variety. And characteristic of Veracity concentrates the importance of addressing the uncertainty in digital data. The inherent Value is an essential parameter

Revised Version Manuscript Received on 10 September, 2019.

N. Silpa, Research Scholar, Dept. of CSE, Centurion University of Technology and Management, Odissa, India. Assistant Professor, Department of CSE, Shri Vishnu Engineering College for Women(Autonomous), Bhimavaram, Andhra Pradesh, India.
(Email: nrusimhadri.silpa@gmail.com)
Dr. V. V. R. Maheswara Rao, Professor, Dept. of CSE, Shri Vishnu Engineering College for Women(Autonomous), Bhimavaram, Andhra Pradesh, India.
(Email: mahesh_vvr@yahoo.com)

hidden in the typical characteristics of data that creates an additional value to the Industry [21, 22, 23, 25, 26, 29].

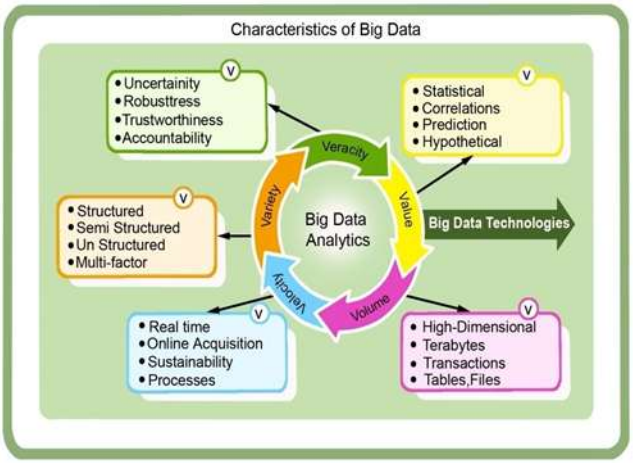

Figure 2. Big Data Characteristics

Along with the characteristics, that is 5Vs, the advanced technologies with its innovative architecture able to define the big data. In addition, these architectures also identify the value from the hidden insights of large volume of data [6, 7, 9, 16]. The definition with its five dimensions as depicted in Figure 2 helps to understand the nature of big web data.

Volume: The size of the web usage data resided in weblog is represented as the first characteristic of big data [9, 16, 22]. The volume of big data evolved into its present stage as megabytes to giga bytes, giga bytes to terra bytes, terra bytes to peta bytes, peta bytes to exa bytes. The present-day web data logs are huge in volume and could not able to process effectively by the techniques and technologies of traditional systems. In future, this will continue to expand exponentially at an unprecedented rate, is a prime motivation to create revolutionary data management mechanisms.

Velocity: The rate of weblog data generation is the key parameter to define the second characteristic of big web data along with proactive response based on the analysis [9, 16, 23, 26]. The advancements in the digital devices create large volume of data in a rapid manner but the computational speed of conventional approaches failed to process at user expectation time. This situation demands a new approach which is agile and responds quickly is another motivation for the current web data researchers.

Variety: Variety shows various kinds of web data as a third characteristic of big data [9, 16, 22]. Variety in big web data is a measure of heterogeneity of data representation. With high usage of gadgets, sensors, social networks, ecommerce websites makes the weblog with various kinds of data. However, the emergence of new data resources enables the researchers towards new storage and processing technologies, which enable to leverage web data in an innovative aspect.

Veracity: The fourth characteristic of big data is veracity that denotes level of certainty and reliability in the web usage data recorded in the weblog [23, 26]. The need to acknowledge and plan for this dimension of uncertainty of big web data is still a major quality concern in processing of weblog data. This triggered the present researchers towards the robust parallel processing and distributed storage techniques to manage certainty.

Value: The fifth characteristic derives the value from the insights of the data patterns and creates added advantages for the respective web applications in all the key domains.

Thus, handling and retrieving useful insights from such typical weblog data triggers many typical issues for big data analytics [9, 16, 22, 23, 26]. The success of big web data analytics necessitates the employment of state-of-the-art techniques and technologies that are scalable, adoptive and robust in nature. They able to filter irrelevant data intelligently, create a solid platform to systematically organize, understand & access the weblog data efficiently and computationally feasible. Towards this, the present authors systematically present all possible potential future research challenges and opportunities to the present researchers of big web data techniques and technologies extensively.

Data Preparation: Many authors [35, 38, 39, 40] in the literature have expressed the significance of weblog data preprocessing stage in the overall process of web usage mining. However, most of the authors pay an attention only on the traditional approaches to pre-process the web usage data. To perform the analysis on web usage data efficiently, the preprocessing stage is an important task due to the typical nature of weblog as it is non-scalable, impractical, incremental, rapidly growing and so on.

Traditional Systems: At present, the storage of large data with the existing systems and processing the same by the conventional techniques is an open challenge for all the organizations. The session identification and user identification are carried out using sequential processing in the traditional systems. These traditional approaches load the weblog data record by record to perform analysis. To overcome the problems in the sequential processing, an active parallel processing is required in the era of high volume data [1, 2, 3, 4, 6]. To accomplish the same, the Hadoop ecosystem is a scalable proven model that works on the well distributed storage and efficient parallel processing.

Hadoop: A Hadoop ecosystem [4, 6, 9, 26] uses HDFS as data storage engine and MapReduce as parallel execution engine. On one hand, the Big Data problem determines how the Big Data platform should be designed, for example, which modules or subsystems should be integrated into the platform and so on. On the other hand, the architectural design of the platform can determine complexity and scalability of the Hadoop eco-system.

The technique of parallel computing and distributed data processing are changed the era of processing and analyzing the huge volume of data in all the key domains. These techniques reduce the cost incurred in the hardware and minimize the execution time. The innovative framework of Hadoop architecture is well suited for the present day data engineering problems.

Hadoop framework access a large semi structure data in a parallel computation model [4, 6, 10, 13]. Log files usually generated from the web server comprise of large volume of data that cannot be handled by a traditional database or other

programming languages for computation. The Hadoop approach aims on preprocessing the log file and keeps track on sessions accessed by the user and the system architecture is shown in Figure 3. The work is divided into phases, where the storage and processing is made in HDFS.

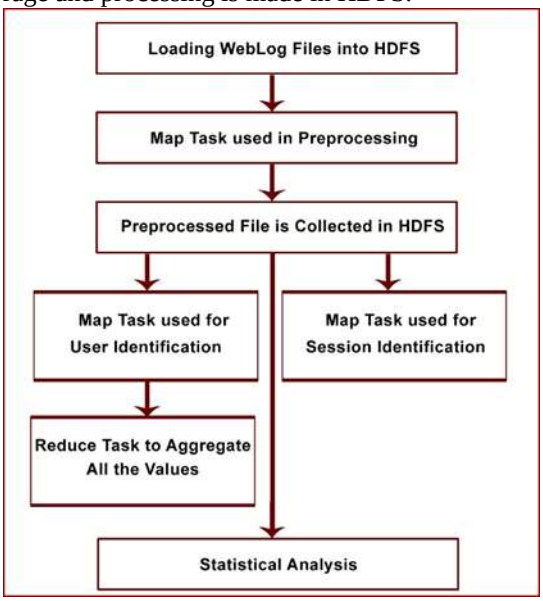

Figure 3. Hadoop Architecture

The remaining sections of the present paper are planned as follows: the contributions of the prominent researchers are earmarked as related work and it is given in section II. After that, the systematic literature review on the methodologies involved in the preparation of big web data through both traditional and Hadoop approach is presented in section II. At last, conclusion remarks and future research paths are recorded in section IV.

II. RELATED WORK

The rapid development of the WWW creates an easy environment to humankind for storing, sharing and retrieving the desired data through internet. Due to this, the web user usage data recorded in the weblogs is becoming so massive. As well as the organizations pay an attention to serve the high aspirant web customers to offer optimal services. To find the various solutions in this path, the authors conduct an exhaustive literature survey from existing traditional approaches to cutting edge technologies for weblog analytics over a last decade. In this section, the authors solely concentrate on the approaches for weblog data collection and Data Preparation as they are the crucial and critical stages and play a significant role in weblog analytics.

The related works of traditional weblog analytics stated in the literature [7, 33, 35, 36, 37] provided a detailed taxonomy, various methodologies to generate web user usage patterns by make use of weblog files. Certain number of authors [35, 37, 45] concentrated on the series of operations involved in the important and computationally intensive stage of weblog data preparation such as various data collection from various sources, and data preprocessing including cleaning of irreverent or noise web data, selection of relevant features, identification of unique user, performing sessionization, completion of incomplete paths, etc., on the original click stream data available in weblog. In the same period, the researchers [35, 37, 42, 44] put forward the works done so far in traditional weblog analytics. And they explored the major bottleneck issues and unsatisfactory results while processing such a huge amount of weblog data.

The era of big data has arrived to offer a best solution for this problem, stated by many prominent researchers [1, 2, 3, 5] in recent past. According to their research contributions, it is observed that the parallel processing technique and distributed technology of big data analytics are not only capable of storing such a huge amount of weblog data and also process the typical characteristics of the data.

In the latest literature, several research works [5, 6, 9, 12] have been carried out in weblog mining using big data analytics. The authors presented the history, concept, models, analytics, tools, applications, challenges, trends and advantages of big data tools. However, in the present paper, the authors pay an attention only on remarkable findings of techniques and technologies involved in preparing the big web data suitable to analytics. In this direction, some of the authors [10, 11, 12, 14, 27] described various methodologies discovered by both research and industry community to pre-process the weblog data efficiently in Big Data environment. According to the research works made by the authors [10, 20, 27, 31] are endorsed that big data storage, big data cleansing, unique user identification, session identification and so on are important and crucial tasks in the big data preprocessing model.

In addition to that, the researchers [3, 5, 6, 21, 23] also revealed the requirement of high-performance and extensible distributed data storage systems to analyze such a big repository of weblog having complex characteristics in data like velocity, veracity, variety etc. With this aim, a few number of authors [4, 5, 6, 10, 13, 17] identified that the complete framework of apache Hadoop well suitable for collecting the weblog data, storing high volume of weblog data over a distributed network, processing the hidden usage patterns and investigating web user usage behavior.

The research papers [4, 11, 16, 17, 18, 28] analyze the challenging issues of Hadoop echo-system in dealing with different kinds of data. The authors find-out the necessity of paying an attention by the present researchers for the development of more innovative techniques by makes use of machine learning paradigm. Many other research paths welcomed the researcher to offer optimal results in serving the high aspirant web users by analyzing their browsing behavior effectively.

III. PROPOSED RESEARCH REVIEW

Over the past decade, analysis of web usage data has emerged and spanned across diverse business verticals and organizations. The web usage data analysis aim to retrieve, extract and evaluate actionable insights to meet the expectations of web user. The traditional web usage data techniques and tools primarily used structure data collected from various weblogs. In addition, the data storage models are unable to store the high velocity data. Moreover, the computational algorithms used in the traditional systems are

unable to process and analyze the data efficiently as the data is generated rapidly.

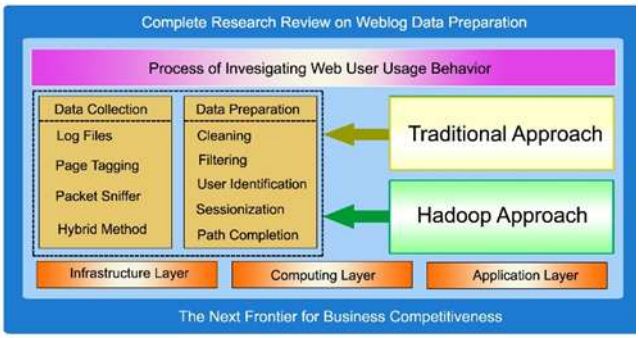

Figure 4. Systematic Framework of the Proposed Study

The value of the web usage mining using traditional methodologies decrease over time, many applications in the present scenario require real time analysis of the transaction data that leads to unbiased conclusions. This situation demands the employment of techniques and technologies of big data analytics for analyzing web usage data effectively [2, 5, 8, 22, 23].

To explore the research paths in the adaption of big data analytics for the web usage data, and retrieve real time knowledge, towards this, the authors in the present paper carefully conduct the survey with Traditional Approach and Hadoop Approach as shown in figure 4.

Initially, the review concentrates on techniques and technologies of data collection and data preparation with respect to Traditional Approach. Subsequently, the authors continue the journey to emphasize the technological growth in weblog storage and computational efficiency for future research in weblog data analysis with the era of big data Hadoop eco-system.

A. Traditional Approach

1. Web Usage Data Collection:

Data collection is a very important and primary step in web user usage mining [7, 35, 36] since the data is unstructured, incremental and diversified in nature. This stage is typical and significant to capture the data from various available web data sources as shown in Figure 5, since the performance, capability, comprehensiveness, and effectiveness of any web data analytics value based on the techniques used for capturing web user usage data. Web miner has to pay an attention to collect the data from various web sources, there are many ways in the literature [33, 35, 36, 45] to collect web usage data as follows.

Figure 5. Systematic Framework of the Proposed Study

a. Log Files:

From the perspective of web usage analytics, weblogs are the original and easily accessible source of data collection. [33, 36, 37, 45]. When the web user browses the website, the click stream data is stored in the log file. The different types of information are maintained differently in log files by different web servers, however, the entries available in the log file are: user_name, time_stamp, visiting_path, request_type, user_agent, URL, success_rate, page_last_visited etc.

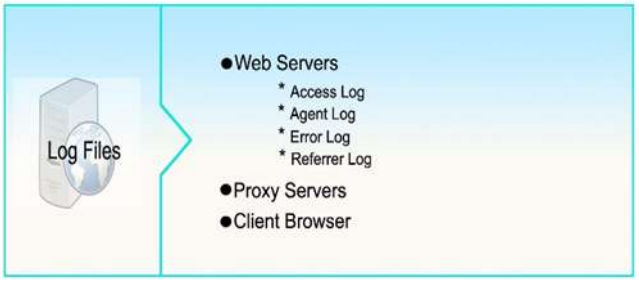

Figure 6. Log Files

The log files are available in 3 possible locations: Web_servers, Proxy_servers and Client_browsers as given in Figure 6.

Web Servers: The data collected using the web servers is richest, usually provide the most complete and accurate web user usage data. There are 4 types of web server log files: access_logs, agent_logs, error_logs and referrer_logs.

Access_log is a major weblog server that records the each click performed by the web server on the website. It generally captures the information about user in the attributes of Client IP, Client name, Date, Time, Server site name, Server IP and soon.

Agent_logs maintain the information of browser used by the web user, version of the browser and operating system. Such kind of data is helpful for well designing of the website and made changes accordingly by website designers.

Error logs store the data belong to error found on website, whenever the user requests a particular website and browser does not display the page and send the error code to the user.

Referrer logs have the information regarding a user came from the particular website by using the user's page link.

According to literature [33, 35, 36, 45] it is noticed that Web servers provide more complete and accurate usage of data, as they explicitly saves the user click stream data but the could not record cache files, practically, it is very difficult to have web servers' data.

Proxy Servers: The collection of web usage data using proxy servers is more reliable [37, 45] as they are placed at client location and act as intermediary between the client browsers and real web servers. They capture all requests made by the clients to the original server and store automatically in weblogs and further they improve navigation speed through caching. There are many ways [35, 36] to collect the web user usage data through proxy servers as follows:

- ✓ Browsing of Single web user on a single website
- ✓ Browsing of Single web user over multiple websites
- ✓ Browsing of Multiple web users visiting a specific website and
- ✓ Browsing of Multiple web users visiting multiple websites

Among all, the way of collecting the web usage data using browsing of "multiple users visiting multiple websites" is a right choice for investigating the web user usage behavior [33, 45].

The sample weblog format collected by proxy server is typically displayed as ASCII text. Each log entry contains fields identifying date and time of the request, the IP address of client system, the resource requested, possible parameters used in identifying the web user usage behavior, status of the request, http method used, the user agent, the referring web resources etc.

Such records are helpful to generate the reports related to the percentage of users browsing the sites for a specific category, the no. of unique users accessing the website, while not a measure of unique users, the quantity of hits the web server receives in a specific hour and day, the average length of a user's session, specific location, duration, average download times, and the user navigation through the site etc. These reports are certainly useful for web miner to understand the user expectation on the web [35, 36].

The log file collected at proxy server is unstructured since the information contains different types of entries. These entries do not have definite number of attributes, identifiable structure and defined relation. This resulted in ambiguities and it was difficult to understand using computer programs. The other characteristic of weblog data is heterogeneous, it means that the data is collected from many number of sources, largely unknown and unlimited, and varying formats. Moreover, the nature of web usage data is distributed in manner. Similarly, the other characteristics of weblog data are: non scalable, voluminous, time dimension, incremental, exponentially growing that decreases the operational performance of web mining techniques. However, the proxy server data encounters with a problem of misinterpreting the IP address and also issue of caching.

Client Browser: One way collecting the weblog files from clients' browser is to make use of remote agent. The other way is with the web user acceptance, change the source code of an user browser by using java scripts and applets [33, 35, 36, 45]. However, the log files are resided in the clients' browser, the entries in log files are recorded by the Web server only.

The data collection from the Client-side is good when compared with server side data collection as it is located at the source of the user behaviors, providing relief from caching and session identification problems. The literature is clearly showing the disadvantage of collecting weblog files through client browser in terms of loading of applet, low performance of applet, unable to capture all user clicks etc. In addition, this method captures only browsing data of single web user on single website.

b. Page Tagging:

The other approach to record the visitor activity when a page was successfully loaded also solves accuracy problems in the log file analysis is page tagging. This method is evolved from script-based data collection which allots a cookie to individual user, and then processes the data remotely. The popularity of page tagging stores the more visitors' information requires no further request to the web server as presented in Figure 7.

Unlike a weblog files, the web data received through this method is tokenized and allows for generating real time reports. The authors [36, 45] also reported that the page tagging method consumes more bandwidth and also fails to record unloaded or failure pages.

Figure 7. Page Tagging

c. Packet Sniffer:

Another web data collection method is the use of Packet sniffer [45] to collect the data through server logs has two methods: (a) deployment of software (b) deployment of hardware device. Either of these methods is able to extract the data by monitoring network traffic. Mainly, this method of data collection suits for recommending web personalization.

When a packet sniffer is placed directly outside a Web server, the segment of usage information available to the packet sniffer is same to the information available to the Web server as depicted in Figure 8. Packet sniffers are also capable of storing information in a log file based on the actual content of the data packets in addition to the HTTP headers. The sniffers able to capture the information about the time stamp which is not available in the log files. In this method, the packet is decoded as per the defined configuration of the software or hardware tool. As a result, this method led to different formats for the collected web data, and also needs additional process. Packet sniffers also are placed farther out in a network away from a single web server in order to collect web usage data for multiple Web servers.

The main advantage of collecting the data through this method is able to store more network-level data which is generally not available in the weblog file. This detailed data includes status of request, user request, response time and so on. In addition to that, all the details related to requested webpage by the user is also maintained.

The authors in the literature notified the disadvantages of this method relating with weblog data. The establishment of the network is may be stateless or state oriented, thus, the order of transmission of the packets may be interrupted. In addition, the network may be failed in some practical environments, so, the details stored by this method may lead to contradiction. Moreover, for the system security, present state sniffer data is a noteworthy threat, thus, sniffers data is not advisable for data collection [45] endorsed by the authors in the literature.

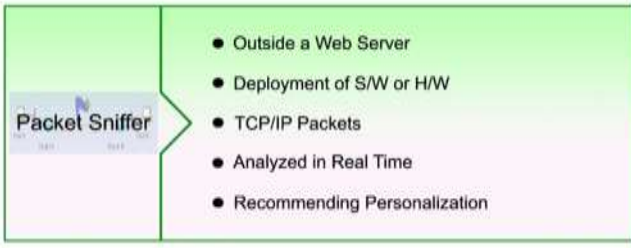

Figure 8. Packet Sniffer

d. Hybrid Method:

In many of the practical applications, hybrid method suits to collect the data which allows combination of multiple data collection methods. Many of the researchers in the literature [33, 35, 36, 41, 45] predominantly recommended that hybrid method gives high accuracy of tracking sessions across multiple domains, eliminating the caching problems, and tracking detailed web data metrics as shown in Figure 9.

Figure 9. Hybrid Method

The collected weblog entries through the above stated possible methods are semi or quasi structured data that are generated by different sources. The high volume of web user usage clicks leverages next generation techniques to develop hybrid modern tools suitable to big data which help in further improvement in manipulating and managing large data along with the storage requirements.

2. Web Usage Data Preparation

Among all the stages of web analytics, web data preparation is the most crucial stage [35, 37, 38, 39, 40, 44], as appropriate data is required to find-out valuable insights from weblog data.

In the process of investigating the web user usage behavior, the Weblog data prompted many issues as it stores all the browsing details. To retrieve and analyze the behavior of web user efficiently, the collected Weblog data need to be Pre-Processed as it contains irrelevant data also.

Thus, the authors in this paper provide in-depth review on web log data preparation that comprises from traditional techniques to recent technologies.

At first, the review concentrates on various data preparation methods like web data cleansing, unique web user identification, web sessionization, and path completion as presented in Figure 10. Subsequently, it focusses on latest technologies and architectures using Hadoop Environment with its sub systems in order to carry out efficient weblog data preparation to improve accuracy and data quality for the applications related to web usage mining like fraud detection, recommender system, forecasting etc.

Figure 10. Stages of Web User Usage Data Preparation

a. Data Cleaning and Feature Selection :

After collecting the data, the web analyst initially focus on removing the data errors, inconsistencies, outliers or irrelevant data in the process of data cleaning also called as data cleanup. This process enables the web analyst to filter out unnecessary data which significantly reduces weblog size and use less storage space.

Data cleaning as depicted in Figure 11 is usually [34, 35, 36, 37, 39] site-specific, and involves the task of eliminating local and global noise. In addition to that, this process also discards the records that have the image file extensions that are presented in the URI field weblog entry. The mentioned extension files in the webpages are not actually interested by the users rather it is just the files embedded in the webpage. The cleaning process further removes the records having status code over 299 or under 200 of failed HTTP requests. Moreover, the web analyst also eliminates the log entries generated by web robots [37, 40, 42] as they are out of the analysis scope.

Feature selection is also an important phase to identify and select the necessary fields from many existing attributes in the weblog, and rest of the attributes are dropped in the process. The authors in the literature prominently addressed the importance of data cleaning and feature selection process for accurate investigation of web user interested usage pattern.

Figure 11. Data Cleaning

b. Determination and Identification of the User:

The aim of the stage is to find-out who accesses webpage [33, 34, 35, 37, 40, 42], that is determined by make use of information related to (a) client type, (b) website topology and (c) cookies. Based on client type, the web miner inspects the agent field in log file to find-out the differences in operating system or web browser. The difference in any parameter indicates the different users for the records even though the records are having same IP address. When user intentionally does like this, it leads to misconception of web miner.

The other method of identifying the unique user is based on the website topology, in this method the process is performed with the help of unvisited webpage along with the similar IP.

Another well-known method of performing user identification is the usage of cookies. Cookie is a variable that saves a parameter value in the client browser. The web server creates a cookie and is passed to the client machine. The created cookie information is sent for further actions made by the same web user with the same browser. However, information of cookies is not logged in web server and there is a possibility to destroy the cookies after some period of time automatically. In addition to that, storing of cookies can be turned on and off by the web user.

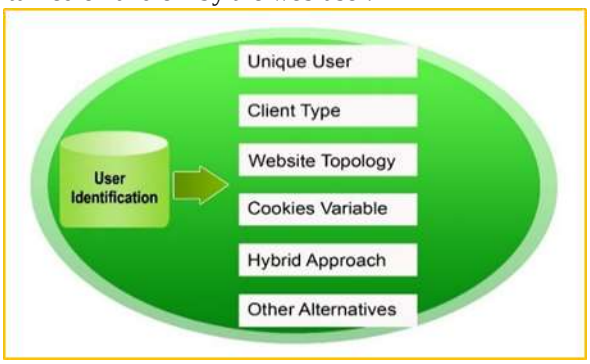

Figure 12. User Identification

The authors in the literature [37, 40, 42] are recommended that the combination of one or more above mentioned approaches as presented in Figure 12 leverages the efficiency in the process of identifying the unique users.

c. Identification of Session:

The task of sessionization as shown in Figure 13 is to find-out the consecutive web pages of a website browsed by a web user in a given time slot. In the literature, the researchers [34, 35, 37, 44, 46] proposed various methodologies to construct and create sessions from weblog by applying heuristics. One among them is to find the session using time-oriented heuristics. The time threshold is defined depends on the type of application, website topology, web users' interest and many other parameters.

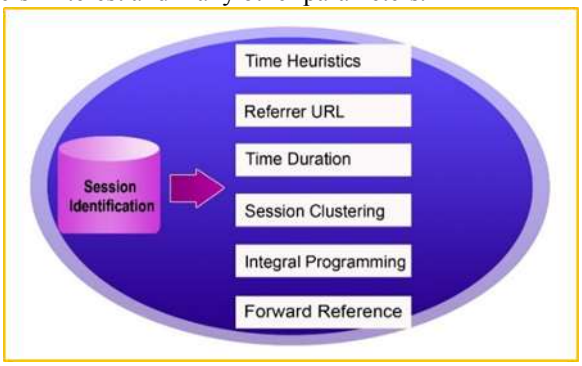

Figure 13. Session Identification

The other method of performing sessionization is depends on the duration of spent time on observing webpages. In this method, the visited web pages are categorized into two groups as informative and navigational web pages. The web user spent more attention to observe the content of Informative web pages rather than the navigational web pages as informative web pages are the ultimate destination of the any web user. The website topology is also considered in addition to the observing time on the web page for defining the session.

The value in the referrer URL is also used for session identification. If the value of refferer URL is absent for a webpage, it is treated as new session otherwise, it treats as existing session. If two or more continuous web page requests are forwarded then they fall within the same session.

In addition to the above methods, present researchers reveal other alternative approaches for web sessionization such as session clustering, integral programming etc.

d. Path Completion:

After sessionization, it is necessary to perform the process of Path completion [34, 40, 45] as presented in Figure 14. The concept of adding the details of page accesses that is actually occurred but not recorded in the weblog is known as path completion. In caching, either client side or proxy server side generally requires the path completion. In addition, in many web applications, the actual occurred page accesses are missed when user clicks on the back or previous type of buttons. Hence, these situations lead to incomplete paths of visited web pages and thus, there is a requirement to detect such missing page sequences in the process of path completion.

In the literature [40, 45] certain number of authors performed the path completion by applying knowledge of website topology and referrer logs. The identified missing page accesses are inserted to the session along with their time duration estimates. These estimates help in classifying web page into navigational or content.

Figure 14. Path Completion

In turn, in the recent past the web usage data analysis got attention of many researchers [22, 23, 39], yet, the data preparation in investigating the web user usage behavior has received less attention than it deserves. Many researchers [35, 38, 39, 40, 44] are doing research in web data preparation with its all possible stages. They endorsed that web usage data preparation is essential and critical phase in the overall web usage mining process. This stage is strengthened by adopting more intelligent algorithms. This helps any organization to track the behavior of web user to meet the desires of specific users and cross-marketing strategies effectively.

B. Hadoop Approach

In the last decade, the web log data preparation stage that includes Data Collection and Data Pre-processing taken a shift and gained its importance. It is not a new problem, however, there is a lot of change taken place in terms of five characteristics of big weblog data. Thus, the value reside in the hidden patterns has to be discovered appropriately to make sense to any business applications.

The existing methods of Web Log data preparation are not suitable for the fast-growing generation of digital data as they unable to collect and conduct analysis of web log data at the same time [5, 6, 9]. Due to the rise in amount of data in logs extremely high, so, processing speed of the existing techniques is limited. Moreover, the percentage of inadequate data in the weblog poses many more challenges in extracting of required knowledge. In addition, in the existing computing platform managing & processing got many issues when compared with conventional weblog data preparation.

The computing technology has also changed to centralized computing platform, need the revolution of business models, create unlimited research opportunities to the present researchers. To improve the decision making for any business applications, it is necessary to adapt the Hadoop echo system [12, 17, 19], that leverages the techniques and technologies used in many applications.

Towards this, many authors [1, 2, 3, 17, 28] in the literature adapted Apache Hadoop platform that enables distributed storage and parallel processing for developing optimal solutions for many applications. The environment of Hadoop-based technology and architecture carry-out distant parallel study and able to overcomes the challenges of traditional techniques with its sub systems. Further, it improves the depository volume and reduces the congestion efficiency of web log analysis. And also, optimizes the processing time effectively.

The Hadoop ecosystem is an open accessible platform, utilized with a goal of providing better distributed storage and processing high volumes of data in parallel. The eco system performs operations on large datasets. The response time is optimized in the Hadoop ecosystem as the computation is moved to the data unlike the traditional systems. Further, this environment proportionally increased from one server to many client systems.

The prime ability of the Hadoop ecosystem is to develop distributed applications which enables for better storage and parallel processing at high speed. The popularity of Hadoop ecosystem with its tools and technologies create a basis of well-structured Hadoop framework. This structure supported by indispensable technologies [5, 13, 19] like HDFS, MapReduce, Hive, Hbase, Pig, Zookeeper, Chukwa and soon.

The authors in the present paper concentrates on exploring the methodology for preparing the weblog data using Hadoop Ecosystem with its core computing sub systems Hadoop Distributed File System (HDFS) and MapReduce as follows.

1. Architecture of Hadoop for processing of Web Logs:

The work explored by the author [4, 6, 10, 13, 17, 18] in the recent past witnessed that the weblog pre-processing is carried out with the well-defined Hadoop ecosystem architecture with the core subsystems HDFS & MapReduce as shown in figure 15. Some more researchers revealed that this customized architecture improves the scalability than the traditional approaches. This architecture is proven in the literature [6, 13, 17] and is simple to implement, optimizes processing time as the collection and analysis are performed in parallel. Additionally, the chosen architecture on preparation of weblog data monitors the sessions accessed by the web user. The superior methodology of Hadoop initially, puts entire data over the HDFS. Subsequently, the framework executes the queries resulting with large amount of customized parallel solutions efficiently shown in figure efficiently.

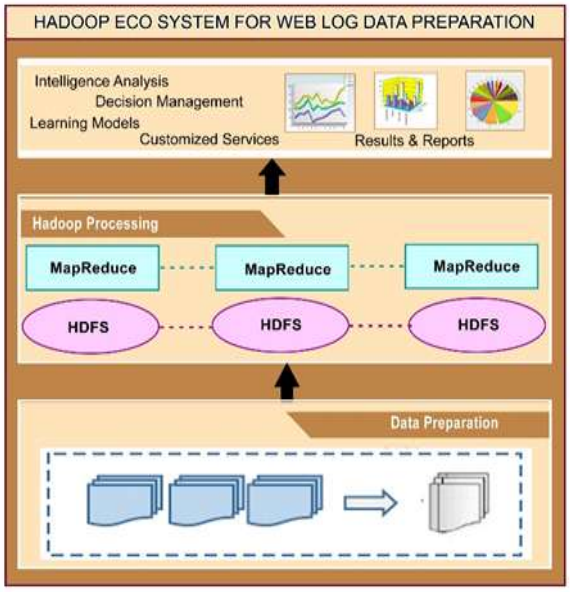

Figure 15. Architecture of Hadoop Eco-System

a. Data Cleaning:

The Data cleaning phase is important in the customized Hadoop architecture. Initially, all log files collected from various web servers are loaded into a single weblog file. Generally, this weblog contains more number of irrelevant records that are created by automatic requests made by web robots, spiders and web crawlers. This misleading and incomplete data need to be filtered-out from the weblog. In addition, the entries in the weblog having the status is error or failure have been removed. The authors in the literature explored that "identification of status code is an important task in the data training". Only the records having 200 as the status code is treated as relevant data. Many authors in the literature [10, 11, 12, 14, 20, 27] are also prominently expressed that removal of failure requests, irrelevant file requests, inappropriate access method requests, web robots requests, internal dummy connection requests and irrelevant log fields to extract more relevant data for the analysis of weblog.

b. HDFS - Storage:

Hadoop Distributed File System (HDFS) is the primary and important sub system of Hadoop ecosystem. It is a distributed file system that is more suitable for executing on the simple hardware systems. In addition, it achieves high

throughput and provides easy access to the high volume data. The HDFS is framed suitable to read and write the streaming large volume and variety of data unlike traditional approaches [5, 13, 19, 26, 28]. HDFS is designed as client / server architecture as shown in Figure 16.

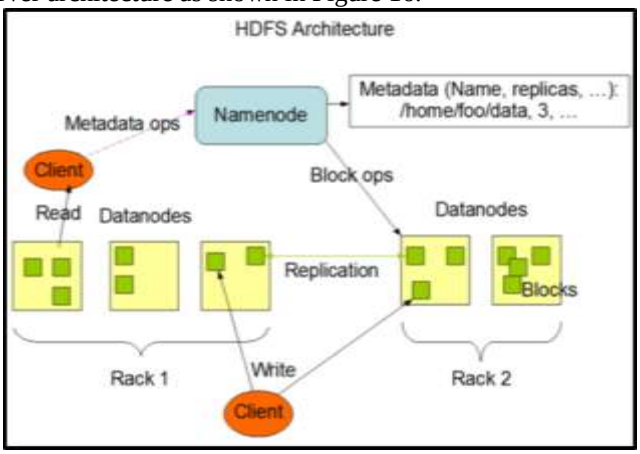

Figure 16. Architecture of HDFS

In the HDFS architecture, the name node is treated as a server node, it is used to store the file system image consisting of the relation of the file and location of the data nodes. Similarly, the Data node is treated as client node, which is used to store files as data blocks.

To provide fault tolerance, the HDFS stores the data blocks with redundancy backup in the Data nodes. Datanode communicates to the name node about the status of storage lists. So, the data node is accessable to the user in a direct way. The reading and writing data in HDFS is carried-out with the support of API. The reading process in HDFS is similar to programming logic, calling a method and performing the execution. The Name node is incharge for handling less data functions, data nodes are responsible for accessing huge size data functions. The process of data writing is little bit complex compared to reading process as any one Data node which creates errors that leads to failure of writing. The authority management of HDFS is similar to Posix, the authority permission for files or directories are based on owner, group or others.

The cleaned relevant web log data is loaded into the HDFS subsystem of Hadoop [19]. This subsystem splits the weblog data into blocks according to the defined block size. Further, the weblog data is distributed among the data nodes of Hadoop cluster with the help of name node and maintains the data as per the value of replication factor. It also re-replicates the data of failed blocks automatically. The Job Tracker of MapReduce takes the weblog data which is stored in HDFS in a specified input file format like TextInputFormat, NlineInputFormat, KeyValueInputFormat and so on.

c. MapReduce - Processing:

The MapReduce is introduced by Google and became another important core subsystem in the Hadoop ecosystem. The literature is proved the usage of MapReduce in searching data, sorting the data and weblog analysis. The literature is also evident that the MapReduce able to provide a parallel processing while analysing the Big Data and opens many research paths for the current day scenario of web log data analysis.

The work explored by the authors [4, 5, 6] witnessed that the MapReduce is a standard practical programming model with two folds of processing. The first fold is Map function and the second fold is Reduce function. The MapReduce got the focus of many present day researchers [12, 15, 17, 26] as it is allows to implement on normal hardware and able to offers commendable expansibility with effective implementation.

The users in this framework develop client jobs and able to submit to a cluster machine. These jobs contain map function code and reduce function code along with job configuration file. After that, the job tracker creates the task trackers and schedules them optimally on the cluster of machines. A job tracker takes the input from the user in terms one or more files on a distributed filesystem and passed to the task trackers and vice versa in case of generating output.

Further, the future open challenges mentioned in the recent survey carried out by [13, 18, 19, 24, 28, 30 , 31] confirmed that the MapReduce and HDFS work together in Hadoop ecosystem as core sub systems.

The architecture of MapReduce includes three basic components: Client, JobTracker and TaskTracker as shown in Figure 17.

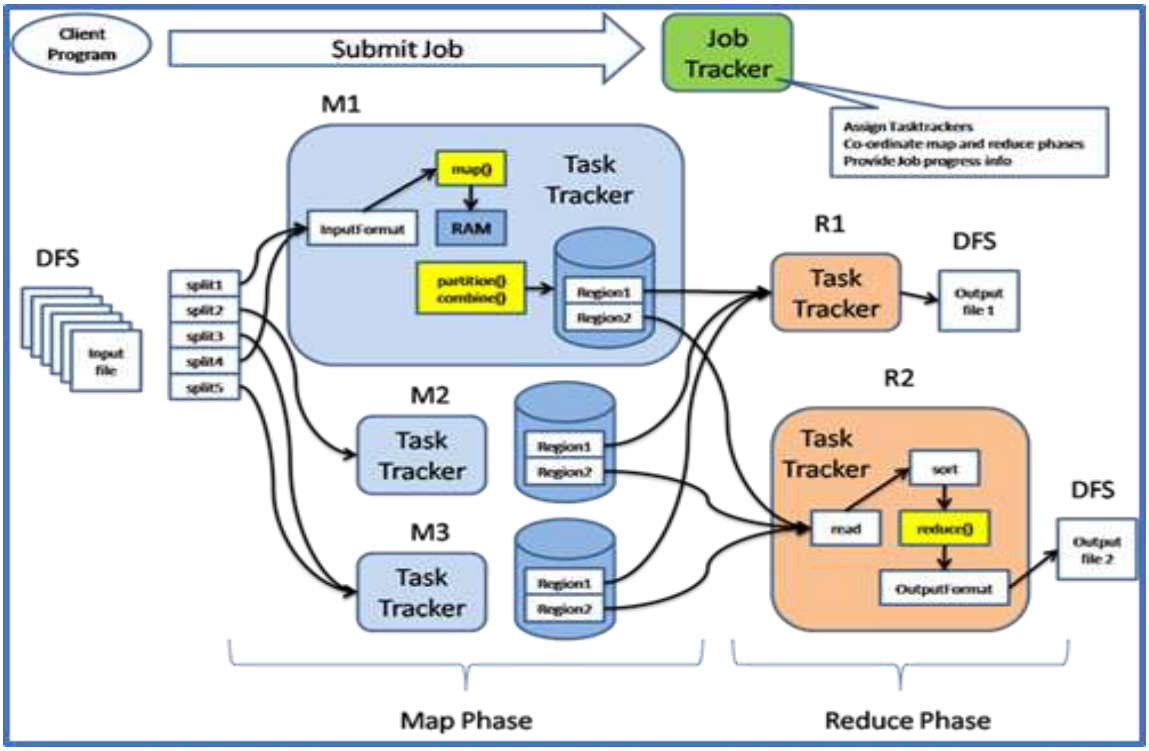

Figure 17. Process of MapReduce

Client: All the jobs of the client are formed in a .jar file and it is given to HDFS, and the path is given to the Job Tracker.

Job Tracker: The responsibility of coordinating all the jobs is taken care by Job Tracker. The Job Tracker monitor all the jobs presently executing on MapReduce. The Job Tracker also redistributes the failed tasks.

Task Tracker: The Task Tracker is given the responsibility of executing the assigned jobs by the job tracker. A well-established active communication is developed between job tracker and task tracker.

The MapReduce function is capable of learning in nature and more typical in executing. The MapReduce and HDFS nodes have good communication and able to work together. The scheduling of tasks performed quickly and the cluster system works efficiently. In brief, the process of MapReduce can be done as shown in Figure 17.

The efforts in the recent past [24, 28, 31] focused on MapReduce capability of processing of high scaled data. They also concentrated on the issues of automatic work distribution with enhanced parallel programming in the context of weblog pre-processing.

Map function executes key and value pair consists of IP Address along with UserAgent and identifies unique sessions [10, 31]. The mapper contains session id (key) and other entries in the weblog file (values). Further, the mapper extracts the attributes like IP address, time stamp, requested URL, user agent. Later, the combination of IP, time stamp along with an attribute of user agent are made the composite key and it sent to the reducer phase.

Then, the derived values are shuffled and sorted according to the time stamp for all entries. The sorted values grouped according to the IP. The reducer works on each key and value pair and then discovers the association among the values. Finally, this Hadoop ecosystem enables the web log analysts to transform the web log data into statistical data and able to generate reports accordingly, thus creating the value addition.

IV. CONCLUSION AND FUTURE SCOPE & RESULTS

The literature [23, 29, 32] is clearly evident that, web usage characterization is a promising and attractive task of web data mining. This extensive research study noticed and emphasized that the usage characterization consists of mainly two interdependent stages: web log pre-processing & storage models and web log processing & analysis. The authors in the present paper also observed the necessity of comprehensive approach in the analysis of weblog data processing and which has been triggered as the formal basis for the future. The review further proven that the analysis of web log data offer thoughts on improving the information about the problems of the different users and can provide customized solutions.

As a future work, Hadoop approach able to concentrates comprehensively on both the stages: distributed data storage & parallel processing of weblog data and to leverage the strengths of techniques and technologies of individual stages [11, 16, 21]. In addition, the comprehensive approach is planned to test with different weblogs that cover a large spectrum of various applications, such as, web usage analysis for improvements in fraud detection, product analysis and customer segmentation. Further, future efforts towards application of Hadoop model in the analysis of web content and structure mining, creates future research paths to leverage in deriving real time knowledge from web user usage data.

V. ACKNOWLEDGMENT

The authors record their gratitude to the authorities of Centurion University of Technology and Management (CUTM), Odissa, India for providing the opportunity to carry out this research. The authors also recorded their acknowledgements to the authorities of Shri Vishnu Engineering College for Women (Automonous), Bhimavaram, A.P., India for their constant support and cooperation.

REFERENCES

1. AC. Priya Ranjani, M. Sridhar, "Distributed Web Usage Mining Based Recommender System in Big Data Analytics using Hybrid Firefly Algorithm", Vol.8, Issue.4, pp.386-393, 2019.
2. AlexNeilsona, Indratmoa, BenDanielb, StevanusTjandrac, "Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications", Big Data Research, Elsevier, Vol.254, pp.1-10, 2019.
3. ShadiKhalifaa, PatrickMartina, RebeccaYoung, "Label-Aware Distributed Ensemble Learning: A Simplified Distributed Classifier Training Model for Big Data", Big Data Research, Elsevier, Vol.15, pp.1-11, 2019.
4. Yassine Azizi, Mostafa Azizi, Mohamed Elbokhari, "Log files Analysis Using MapReduce to Improve Security", Second International Conference on Intelligent Computing in Data Sciences (ICDS 2018), Science Direct, Elsevier, pp.37-44, 2019.
5. Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, "Big Data technologies: A survey", Journal of King Saud University - Computer and Information Sciences, Vol.30, pp.431-448, 2018.
6. Dr. Venkatesh Naganathan, "Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends", International Research Journal of Engineering and Technology (IRJET), Vol.5, Issue.5, pp.1948-1964, 2018.
7. Jayanti Mehra, Dr. R S Thakur, "An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining", International Journal of Applied Engineering Research , Vol.13, Issue.2, pp. 1227-1232, 2018.
8. Jens Baum, Christoph Laroque, Benjamin Oeser, Anders Skoogh, Mukund Subramaniyan, "Applications of Big Data analytics and Related Technologies in Maintenance-Literature-Based Research", Machines, MDPI, pp.1-12, 2018.
9. Pritee Chunarkar-Patil, Akshanda Bhosale, "Big Data Analytics", MedCrave, Open Access Journal of Science, Vol.2, Issue.5, pp.326-335, 2018.
10. Mitali Srivastava, Rakhi Garg, P.K. Mishra, "A MapReduce-Based User Identification Algorithm in Web Usage Mining", International Journal of Information Technology and Web Engineering, Vol.13, Issue.2, pp.11-23, 2018.
11. Yuji Roh, Geon Heo, Steven Euijong Whang, "A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective", ArXiv, pp.1-19, 2018.
12. Abhiruchi Shinde, Neha Vautre, Prajakta Yadav, Sapna Kumari, "Log Analysis Engine with Integration of Hadoop and Spark", International Research Journal of Engineering and Technology (IRJET), Vol.4, Issue.3, pp.1671-1676, 2017.
13. Imam Fahrur Rozi1, Ridwan Rismanto, Siti Romlah, "Implementation Of Big Data FrameWork In Web Access Log Analysis", International Journal of Advanced Engineering and Management Research, Vol.2, Issue.5, pp.1692-1701, 2017.
14. Jayaram Hariharakrishnan, Srividya, Mohanavalli.S, Sundhara Kumar K.B, "Survey of Pre-processing Techniques for Mining Big Data", International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, pp.1-6, 2017.
15. Mitali Srivastava, Rakhi Garg, P. K. Mishra, "A MapReduce-Based Parallel Data Cleaning Algorithm In Web Usage Mining", International Journal of Computer Science and Applications,Vol.14, Issue.2, pp.115-129, 2017.
16. Dr.S.Vijayarani, Ms. S.Sharmila, "Research In Big Data - An Overview", Informatics Engineering, an International Journal (IEIJ), Vol.4, Issue.3,pp.1-20, 2016.
17. Lavanya KS, Srinivasa R, "Customer behavior analysis of web server logs using Hive in Hadoop Framework", 1st International Conference on Innovations in Computing & Networking (ICICN16), CSE, RRCE, pp.408-412, 2016.
18. Pooja D. Savant, Debnath Bhattacharyya, "A Hadoop-based Retail E-Commerce Weblog Analysis System", IOSR Journal of Computer Engineering (IOSR-JCE), Vol.18, Issue.3, pp.4-12, 2016.
19. Pooja D. Savant, Debnath Bhattacharyya, Tai-hoon Kim, "Hadoop based Weblog Analysis: A Review", International Journal of Software Engineering and Its Applications, Vol.10, Issue.6, pp.13-30, 2016.
20. Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, Francisco Herrera, "Big data preprocessing: methods and Prospects", BioMed Central, pp.1-22, 2016.
21. Amir Gandomi,Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics", Elsevier, pp.1-8, 2015.
22. Dr. V.V.R. Maheswara Rao, Dr. V. Valli Kumari, N Silpa "An Extensive Study on Leading Research Paths on Big Data Techniques & Technologies", International Journal of Computer Engineering & Technology (IJCET), Vol. 6, Issue.12, pp. 20-34, 2015.
23. Dr. V.V.R. Maheswara Rao, Dr. V. Valli Kumari, N Silpa "A Comprehensive Study on Potential Research Opportunities of Big Data Analytics to Leverage the Transformation in Various Key Domains" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 5, Issue.5, pp.1-18, 2015.
24. Hemant Hingave, Rasika Ingle, "An approach for MapReduce based Log analysis using Hadoop", Nagpur, India,pp.1-5, 2015.
25. Lisbeth Rodríguez-Mazahua, Cristian-Aarón Rodríguez-Enríquez1,José Luis Sánchez-Cervantes1,Jair Cervantes,JorgeLuis García-Alcaraz, Giner Alor-Hernández, "A general perspective of Big Data: applications, tools,challenges and trends", J Supercomput-Springer, pp.1-41, 2015.
26. Marlina Abdul Latib, Saiful Adli Ismail, Haslina Md Sarkan, Rasimah Che Mohd Yusoff, "Analyzing Log In Big Data Environment: A Review", ARPN Journal of Engineering and Applied Sciences, Vol.10, Issue.23,pp. 17777-17784, 2015.
27. Zuhair Khayyat, Alekh Jindal, Ihab F.Llyas, Samuel Madden,Mourad Ouzzani,Paolo Papotti, et.al., "BigDansing: A System for Big Data Cleansing", pp.1-7, 2015.
28. Chen-Hau Wang, Ching-Tsorng Tsai, Chai-Chen Fan, Shyan-Ming Yuan, "A Hadoop Based Weblog Analysis System", 7th International Conference on Ubi-Media Computing and Workshops, IEEE, pp.72-77, 2014.

29. Han Hu, Yonggang Wen, Xuelong Li., "Toward Scalable Systems for Big Data Analytics:A Technology Tutorial", IEEE. Translations and content mining, Vol.2, pp.652-687, 2014.
30. Meijing Li, Xiuming Yu, Keun Ho Ryu, "MapReduce-based web mining for prediction of web-user navigation", Information Science, Vol.40, Issue.5, pp.557-568, 2014
31. Savitha K, Vijaya MS, "An Efficient Analysis of Web Server Log Files for Session Identification using Hadoop Mapreduce", Proc. of Int. Conf. on Advances in Communication, Network, and Computing, CNC, Elsevier, pp.241-246, 2014.
32. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Minninvg with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol.26, Issue.1, pp.97-107, 2014.
33. Chintan R.Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod, "Web Usage Mining: A Review on Process, Methods and Techniques", International Conference on Information Communication and Embedded Systems (ICICES), pp. 1-7, 2013.
34. V.Chitraa, Dr.Antony Selvadoss Thanamani, "Web Log Data Cleaning For Enhancing Mining Process", IJCCTS Transaction, Vol.01, Issue.03, pp.49-55, 2012.
35. Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", Singapore, pp.1-5, 2012.
36. L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, pp.99-110, 2011.
37. Ma Shu-yue, Liu Wen-cai, Wang Shuo, "The Study on the Preprocessing in Web Log Mining", Fourth International Symposium on Knowledge Acquisition and Modeling, Sanya, China, pp.1-3, 2011.
38. Michal Munk, Martin Drlík, "Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System", International Conference on Computational Science (ICCS 2011), Published by Elsevier Ltd, pp. 1640-1649, 2011.
39. V.V.R.Maheswara Rao, Dr. V. Valli Kumari, "An Enhanced Pre-Processing Research Framework for Web Log Data using A Learning Algorithm" published in an International Journal of Computer Science And Information Technology - IJCSIT, Vol.1, pp.01-15, 2011.
40. V.V.R.Maheswara Rao, Dr. V. Valli Kumari, Dr. KVSVN Raju "An Intelligent System for Web Usage Data Preprocessing" The First International Conference on Computer Science and Information Technology - CCSIT-2011, Bangalore, Vol.131, Issue.1, pp.481-490, 2011.
41. Brijendra Singh, Hemant Kumar Singh, "Web Data Mining Research: A Survey", IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, pp.1-10, 2010.
42. Michal Munka, Jozef Kapustaa, Peter Sveca, "Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor", International Conference on Computational Science (ICCS 2010), Published by Elsevier Ltd, pp.2273-2280, 2010.
43. V.V.R.Maheswara Rao, Dr. V. Valli Kumari, Dr. KVSVN Raju "An Effective Intelligent Pre Processing System of Web Log data adaptable to Incremental Mining" International Conference on Advances in Communication, Network and Computing (CNC 2010), Calicut, IEEE Xplore, pp. 53-59, 2010.
44. Olfa Nasraoui, Maha Soliman, Esin Saka, Antonio Badia and Richard Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge And Data Engineering, Vol.20, Issue.2, pp.1-13, 2008.
45. Jaideep Srivastava, Robert Cooley , Mukund Deshpande, Pang-Ning Tan, "Web Usagae Mining: Discovery and Applications of Usage Patterns from web Data", SIGKDD Explorations, Vol.1, Issue.2, pp.12-23, 2000.

AUTHORS PROFILE

N Silpa completed her M. Tech. degree in Computer Science and Engineering at Jawaharlal Nehru Technological University, Kakinada, India. She is working as Asst. Professor in the Dept of Computer Science and Engineering Department at Shri Vishnu Engineering College for Women (Autonomous), Bhimavaram, AP, India. She is currently pursuing her Ph.D. in Computer Science Engineering Centurion University of Technology and Management (CUTM), Odissa, India. Her Research interests include Data Mining, Web Mining, Big Data Analytics, Tex Mining, Artificial Intelligence and Machine Learning. She has 10 years of teaching experience and 5 years of Research Experience.

Dr. V.V.R. Maheswara Rao holds Ph.D. degree in Computer Science & Engineering from Acharya Nagarjuna University, Guntur, India.. He is working as Professor in the Dept of Computer Science and Engineering at Shri Vishnu Engineering College for Women (Autonomous), Bhimavaram, AP, India. His Research interests include Web Mining, Big Data Analytics, Artificial Intelligence and Machine Learning. He has completed two DST funded Research projects. Presently he is working on one more DST Research project. He has 21 years of experience that include 6 years of Industry experience, 15 years of Teaching experience and 11 years of Research experience.