Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 9, June-2019, ISSN 2230-7540

Study on the Challenges of Big Data Computing and the Security Issues Associated with Big Data in Hadoop

Ruchi Sawhney[1]* Prof. (Dr.) K. P. Yadav[2]

[1] Research Scholar, Department of Computer Science, Himalayan University, Itanagar, Arunachal Pradesh

Abstract – As the world is digitizing the speed at which the quantity of records exceeds due to exclusive assets in specific format, it is not always feasible for the conventional machine to calculate and analyze this kind of massive facts for which massive recording tool like Hadoop is used, which is an open supply software. In a dispensed environment, it shops and computes records. Big Data Systems have become increasingly critical in the past few years. In reality, most organizations depend on data from massive amounts of information. Modern data methodology, however, indicates a reduced overall efficiency, reliability, incremental sensitivity, and lack of scalability. Masses of work have been done to clear up the confusing Big Data headache. As a result, different technology styles have been advanced. As the arena is being digitized the velocity in which the amount of information is overdue from different assets in different layout, it is not feasible for the traditional system to measure and analyze this type of large records for which massive data tool such as Hadoop is used, which is an open source software program.

----------X----------

INTRODUCTION

Big Data

Data is one of the maximum crucial and important components of different activities in latest global. Therefore, in each and every 2d, a tremendous amount of facts are generated. A rapid improvement in modern-day information in one-of - a-kind domain names requires a high-brow information analysis tool that can be useful in meeting the need to analyze a large amount of facts.

Big data usually refers datasets which have grown too massive for and end up too tough to paintings with by conventional gear and database management structures. It also implies datasets which have a extremely good deal of variety and velocity, producing a need to increase possible answers to extract fee and expertise from extensive-ranging, fast-transferring datasets (Elgendy, N. And Elragal, A., 2014). According to the Oxford English Dictionary, "Big records" as a term is defined as "extraordinarily massive information sets that may be analysed computationally to expose patterns, tendencies, and associations, mainly referring to human behaviour and interactions". Arunachalam et al. (2018) argued that this definition does now not give the entire photo of big records, but, as big information must be differentiated from statistics as being hard to address the use of traditional information analyses. Big records therefore inherently requires greater sophisticated techniques for managing complexity, as this is exponentially accelerated. By 2011, the time period large information had end up pretty considerable, however suggests the frequency distribution of the "large facts" within the ProQuest Research Library more absolutely (Gandomi and Haider, 2015).

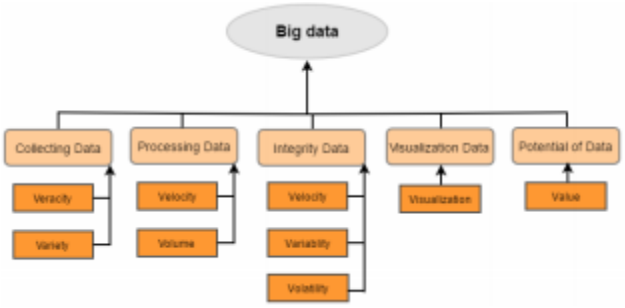

Fig.1.1. Big Data with 9V's Characteristics

Characteristics of Big Data

Big facts is data whose scale, circulation, assortment, or potentially practicality require the utilization of late specialized structures, examination, and apparatuses so as to empower bits of knowledge that free new resources of business esteem. Three chief highlights speak to enormous data: amount, range, and velocity, or the 3 V's. The volume of the insights is its size, and how gigantic it's miles. Velocity alludes back to the

accuse of which data is changing, or how habitually it's far made. At last, assortment comprises of the particular arrangements and styles of realities, just as the various sorts of employments and strategies for breaking down the insights. Data amount is the essential quality of gigantic insights. Big data might be measured through length in TBs or PBs, notwithstanding even the amount of insights, exchanges, tables, or documents.

Also, one of the issues that make big certainties without a doubt big is that it's originating from a more extensive assortment of assets than any time in recent memory, for example, logs, clickstreams, and online life. Utilizing those advantages for examination strategy that not bizarre ward records is currently joined through unstructured data, together with literary substance and human language, and semi-built up data, which incorporate extensible Markup Language (XML) or Rich Site Summary (RSS) channels. There's also measurements, that is hard to order since it originates from sound, video, and different contraptions. Besides, multi-dimensional records can be drawn from an insights distribution center to add old setting to enormous data. Along these lines, with tremendous data, assortment is essentially as enormous as degree. In addition, big certainties can be portrayed by its velocity or speed. This is to a great extent the recurrence of records time or the recurrence of data conveyance. The main edge of enormous actualities is gushing data that is gathered continuously from the sites. A few scientists and organizations have referenced the expansion of a fourth V, or veracity. Veracity makes a speciality of the great of the statistics. This characterizes huge records satisfactory as excellent, horrific, or undefined because of records inconsistency, incompleteness, ambiguity, latency, deception, and approximations.

Hadoop

Hadoop is a free, Java-based programming frame work that guides in the preparing of enormous arrangements of data in an appropriated processing condition. It is a piece of the Apache venture supported by the Apache Software Foundation. Hadoop group utilizes a Master/Slave structure. Utilizing Hadoop, huge data sets can be prepared over a bunch of servers and applications can be run on frameworks with a large number of hubs including a great many terabytes. Circulated document framework in Hadoop helps in quick data move rates and permits the framework to proceed with its typical activity even on account of some hub disappointments. This methodology diminishes the danger of a whole framework disappointment, even on account of countless hub disappointments. Hadoop empowers a registering arrangement that is adaptable, financially savvy, flaw open minded and adaptable. Hadoop Framework is utilized by mainstream organizations like Google, Yahoo, Amazon and IBM and so forth. To help their applications including colossal measures of data. Hadoop has two principle sub extends to be specific Map Reduce and Hadoop Distributed File System (HDFS) .

ISSUES AND CHALLENGES IN BIG DATA

Big Data Issues and Challenges Related to Characteristics of Big Data

Volume- Over time, structures and people being are constantly growing new information, resulting in an amazing extent of statistics. For example, extra than 5 billion individuals are the use of diverse mobiles gadgets for this reason, the quantity every yr is growing more and more. Volume of records saved in corporation repositories have grown from gigabytes to petabytes. Many elements make contributions to the boom in records volume like transaction based information saved through the years, unstructured information streaming in from social media and many others. Huge amounts of sensor and system-to-device information being accumulated. In the past days, immoderate facts volume turned into a storage trouble. But with decreasing storage prices, other troubles emerge, including the way to determine relevance within massive records volumes and how to use analytics to create price from applicable facts. Volume referred as amount of data.

Velocity-Velocity method the timeliness of Big Data, ought to be hastily and well-timed conducted, consequently information is flowing at high pace and have to be treated a practical way. Data is streaming in at awesome pace and must be dealt with in a well-timed way. RFID sensors and smart metering are using the want to cope with fast-moving of records in near-real time. It is a challenge for maximum companies to reacting fast enough to address records speed. Velocity alluded the velocity of data handling. For time-unstable methodologies, for example, getting misrepresentation, big insights must be utilized. It streams into your undertaking to have the option to amplify its charge. Velocity alludes to the speed at which new realities is created and the speed at which certainties activities round. Report submit with the guide of unquestionably show that the velocity of Internet have tons higher than the creating us of a Numerous practical potential outcomes lie in these impalpable things.

Variety. Organized and unorganized facts are producing a selection of data kinds. It is really worth citing that there exist various sorts of information, which consist of semi-established and unstructured facts such as audio, video, internet page, and text, in addition to traditional based data. Today information comes in unique sorts of formats. Structured and numeric statistics in conventional databases. Information constituted of line-of-enterprise programs. Unstructured textual content documents, e mail, video, audio and

financial transactions. Managing, merging and governing specific forms of statistics are something many organizations nonetheless conflict with. Different kinds and resources of facts are there. Data range exploded from established and legacy data saved in corporation storages to unstructured, semi dependent, audio, video and so on. Assortment is a proportion of the wealth of the realities portrayal content, pics, video, sound, and sensor actualities. In truth, seventy five percent of overall data are unstructured. It might be unstructured and it can incorporate such a significant number of excellent sorts of realities from XML to video to SMS. The chief undertaking to arranged the futile realities into important actualities is enormous task for the examination.

Value- The foremost goal of Big Data is to get utility fee in terms of reading or discovering new features from the unique statistics. As it implies, Big Data's value thing defines as associated with a substantial length. Statistical size plays a very critical role in figuring out statistical costs. Value is the maximum important dimension of big records — the end game. Big information in itself may be of no precise fee, the amount of facts extracted from large statistics, the analysis executed in this information and the conclusions derived and the measures positioned into impact based totally on those conclusions make the cost of huge facts its most important dimension. The value of huge information is in how agencies will positioned this information to apply to make their merchandise greater powerful, far-reaching and ubiquitous. For example, Google is the maximum used amongst search engines like google which resulted inside the word “google” making an professional access in the English dictionary in 2006. On the other hand, Facebook remains the maximum critical social media website since its release in 2004 and despite facing stiff opposition from Twitter, Instagram and Snapchat. Despite Google’s being a frontrunner inside the massive statistics and analytics subject, it still wasn’t able to usurp Facebook’s role as the most famous social media website with Google+, Google’s social networking internet site. There are numerous reasons at the back of this interesting prevalence however one in all them is that Facebook is extra adept at using huge information to healthy its social media platform than Google has been with Google+.

REVIEW OF LITERATURE

According to Dr. Saravanakumar N. M. et. Al “Big Data technology are beneficial in health care industry. By the usage of it we can exchange the complete fitness care cost chain from drug evaluation to sufferers worrying first-class. But the unstructured nature of Big Data of fitness industry is observed. So, it's far important to shape the Big Data of health industry” [. 49 Srivathsan, M and Yogesh, Arjun K, also describe technological advancement in healthcare sector. This advancement is feasible simplest through the implementation of Prognotive computing consists of the Big Data analytics. In this procedure structured and unstructured biomedical information can be acquired from a wide range of experiments and surveys accumulated via hospitals, laboratories, pharmaceutical businesses or maybe social media.

Ping Jiang and Jonathan Winkleyet Al studied and analysed the records produced through the wearable sensors. They presented a “Big Data healthcare system for aged people”. Such a Big Data device can provide wealthy facts to healthcare companies about individuals’ fitness situations and their dwelling environment. Thus indicated the need of the Big Data era in gathering and managing the records produced. Ruchie Bhardwaj et. Al has mentioned the Big Data technology in health care industry. According to the researcher, the five cost pathways are consisted of right living, proper care, right carriers, right fee and right innovation. They 50 outline the framework of the brand new enterprise. These tactics cause a greater a success remedy for sufferers. The future is brilliant for the newest intersection between era and healthcare.

Kevin Hamlenand et. al. proposed that documents could be saved rather than some literary content in a fragmented server. The advantage of putting away scrambled data is that, despite the fact that interloper can enter the database; the actual data cannot be obtained by the individual in question. The downside, however, is that encryption requires different overheads. Instead of preparing the plain content, in cryptographic structure, the vast majority of the activity will occur. Subsequently, the fit handling protocol transmitted to the safety layer more noteworthy.

Xiaochun Yun suggested execution of FastRAQ-Big Data inquiry in the fulfillment of all technical questions. In the first place, a balanced stage calculation is used to hole goliath measurements into free bundles, which is a factor close to the approximation of each component. FastRAQ resulted roughly by method of close integration from all segments by estimate. The stage of Linux is profitable for completing FastRAQ and evaluating execution on trillions of insights. As with the designers, FastRAQ can provide impressive starting rates of gigantic continuous actualities. This clarifies the 1:n base problem of full investigation, but m: n formal issue by and by open air. Look in for Big Data, i.e. Odd country dataflow structure is delineated in Vasiliki Kalavri as an extensible and vernacular independent framework m2r2. This execution of adaptation is practiced on the Pig dataflow system and the results are therefore dealt with in finding, usual sub-question coordinating patching despite garbage collection. The assessment is carried out using the TPC-H pig benchmark and the study decreases in the execution of the inquiry with a guidance of 65%.

Portrayal of variable test coordination is carried out through the ludicrous placing of the near-to-descriptor line diagram.

Hadoop starts from an open source web crawler, Apache Nutch. The initiators composed an open source execution based on Google's appropriated file system **(Ghemawat, Gobiof, and Leung, 2003), alluded to as Nutch Distributed Filesystem (NDFS)**, in the wake of realizing that present designs would never again scale up to the billions of pages on the web. In 2004, Google released a paper presenting MapReduce, a parallel programming model and a related use for the preparation, consideration and production of large record units through a bunch of item machines (Dean and Ghemawat, 2008), to people in general. All Nutch calculations were carried nearly a year later to use MapReduce and NDFS. In 2006, Nutch have become a different subproject underneath the name Hadoop and after two years it turned into a zenith degree adventure at Apache, affirming its accomplishment. In that one year, numerous associations worldwide, including Last. Fm and Facebook, use Hadoop. For some, Hadoop is an equivalent word for huge data due to its ability to keep and monitor enormous amounts of (unstructured) realities in a monetarily mindful manner within a shorter period of time (Kuil, 2012).

OBJECTIVES OF THE STUDY

1. To study the optimal fetching techniques of Hadoop.

2. To study the Hive fetching technique of Hadoop.

3. To study the big data computing and its association with Hadoop.

HYPOTHESIS

H1: An optimal fetching strategy in analytics and huge facts impacts its overall performance.

RESEARCH METHODOLOGY

Experimental Evaluation

We have consistently picked the present day for our inquiries-at the hour of the work-stable version of the system envisaged for the exams. Furthermore, we selected freely accessible data sets, attainable each time, and score our setup of arrangements in everything about articles, a perfect method to promote reproducibility. We made use of advanced machines in a cloud domain all of our analyzes. The use of virtual machines enabled us to create a smooth, remote environment in which the most important tools were installed. To pick delegate applications for our evaluation, we either picked applications that were depicted in each gadget's first examination papers or projects that fit the region in which our advancement process turned into focused applications.

METHODOLOGY OF BIG DATA ANALYTICS

This phase explains diverse levels of lifecycle of massive records analytics.

A. **Data identity and collection-** In this section, huge kind of statistics sources are diagnosed depending upon the severity of trouble. More facts sources mean extra chances of finding hidden correlations and patterns. Tools are had to seize keywords, information and facts from those heterogeneous data sources.

B. **Data garage**- The captured established and unstructured information need to be saved in databases/ data warehouse. NoSQL databases are had to accommodate Big Data. Various frameworks and databases were advanced by organizations like Apache, Oracle and so on. That allow analytics gear to fetch and method information from these repositories.

C. **Data filtering and noise removal-** This section is devoted to removal of replicated, corrupt, null and inappropriate facts gadgets from the accrued facts. However, filtered and eliminated records might be of some importance in some other context or evaluation. Hence, it is really useful to hold a copy of unique facts units in compressed shape to save storage space.

D. **Data classification and extraction-** This section is responsible for extracting incongruent statistics and changing it right into a commonplace facts format that the underlying analytics device can use for its purpose. This may contain extracting applicable fields or texts to lessen the extent of facts to be submitted to analytics engine.

E. **Data cleaning, validation and aggregation-** This degree applies validation regulations based totally on the enterprise case to verify the necessity and relevance of statistics extracted for analysis. Although it is able to be difficult from time to time to use validation constraints to the extracted statistics due to complexity. Aggregation facilitates to combine multiple information sets into fewer numbers based totally on commonplace fields. This simplifies in addition information processing.

F. **Data evaluation and processing-** This degree incorporates out actual statistics mining and evaluation to establish particular and hidden patterns for making commercial enterprise decisions. Data analytics technique may additionally vary relying upon the state of affairs. Exploratory, confirmatory, predictive, prescriptive, diagnostic or descriptive.

G. **Data visualization-** This phase involves representation of analysis effects into visual or graphical shape that makes it easier to understand for the target audience.

Research layout

This examine will cover name of the examine, importance of the take a look at, pursuits and targets of the take a look at, studies speculation and research layout. This studies has designed based upon descriptive take a look at as it goals to examine the fetching strategies in analytics and massive information and tricky its destiny enhancement with structural framework in Hadoop. The research design contains the subsequent steps:

Data collection

This investigation consolidates research systems that are both number one and auxiliary. In this way, it may be possible to accumulate and investigate the data on the basis of present examinations. DATA SET For our execution of the errand, we use Wikipedia actualities set which is openly to be obtained from them. A collection of documents includes approximately clients realities that may attempt to discover the records as well as actualities in the washroom. The site hit static document provides detailed insights into the log of innovations that were generated on the fly, or articles from Wikipedia or adventure that were dispatched through as much as the network that the hurls demand from an inward host. The channel out creates actualities such as the endeavor call, the solicitation's page size, and the name of the listed website page. There are various records containing up to 110 MB of data and each consolidates the actualities as a base 1 lake. From these we use certain archives where it is possible to record over 1.5 Lac and handle that we install Linux on this gadget and then HADOOP.

Tools and strategies

In addition to security features, we will donate large statistics processing strategies / mechanism that would improve the security of the computer environment. Since the Hadoop environment is a mixture of many exclusive technologies, we can suggest solutions that collectively make the environment safe as well as suggested responses that encourage the use of more than one technology / equipment to mitigate large-scale information processing and analytics. Security points are designed in such a way that they do not reduce cloud systems ' efficiency and scaling. In line with this, SPSS statistical package deal of facts evaluation will rent to analyze the quantitative records.

DATA ANALYSIS

Process Capability Enhancement

Application-Level Optimization: according to the Big Data Application-Level Optimization Transfers across Pipeline, Parallelism and Competition investigations, how capability enhancement has been developed to improve performance. Application-organizing transfer alteration parameters such as pipeline, parallelism and efficiency are crucial home gear for practical cloud packages to go beyond measurement transmission bottlenecks. Such parameters will increase the absolute best rate of transmission. Through these pointers and portrayals, the examinations propose optimization calculations that can offer the most throughputs between cloud and intra-cloud movements a typical development. Developmental optimization: the following comments on the accentuation of evolutionary optimization on maximizing the efficiency of the system. By using knowledgeable hereditary administrators, the proposed rendition presents a decent variety. The inquiries are also familiar with a measurement that adapts to the problem of the top dimensionality. The POPULATION EA set of rules on this uses a gander to monitor mostly the matter of advanced dimensional issues. Together with complex multi-model response field, the POPULATION EA prototype is able to handle high dimensional optimization.

Memory Management Enhancement

In-Memory Big Data Optimization: This study on InMemory Big Data Management and Memory Management Enhancement Processing outline. The exams provide a complete basic period investigation into memory control and related work investigation. This analysis dedicated to prepare estimates for the management and handling of inmemory data and realistic methods for organizing and implementing impressive and functional memory structures.

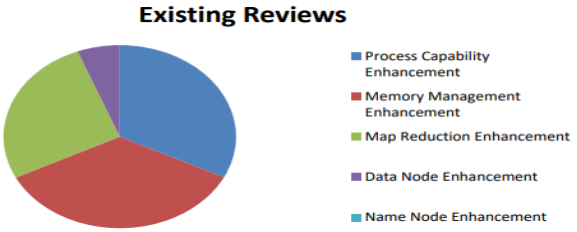

Fig.1.2. Statistics of Survey on Various Enhancement Techniques of Big Data.

A. Map Reduction Enhancement

Stage optimization in line with Big Data: for the most part, the decision to pick a specific stage for a particular application depends on a few components: record length, optimization of speed or output, and improvement of rendition. There are different big data stages blessing with different capacities and the choice of the definite stage requires in-force information about the capacities of such stages.

B. Data Node Enhancement

Knowledge discovery in multi-target optimization: Statistics mining strategies survey for information discovery in multi-target optimization means enhancement of statistics nodes. Maximum multi-goal optimizers start with a population of arbitrarily produced results or entities that are subject to iterative deviation and assortment until a specific number of generations or performance goals or characteristic estimates are affected.

C. Name Node Enhancement

Optimization of small files: Bo Dong et. Al carried out an investigation into putting away and getting a small report in which the author plans an increasingly favorable method for blowing up the capacity and adding small file skills to HDFS. Converging various small files and prefetching for fundamentally relevant small documents and gathering documentation and prefetching for legally related small archives on this work has been suggested.

Big Data Analysis

Big Data Analysis applies to observations that surpass the generic carport, storage, and recording limit for traditional databases using data analysis systems. Big Data Assessment could be adapted with significant statistical models. Big Data Analytics can be described as the use of prevailing scientific approaches on enormous assurances. Big Data Analysis involves different realities mining methodologies to locate the targets. In this chapter, we will summarize items by analyzing different data mining strategies that are used in Big Data Analysis over and over again.

Machine Learning: Machine analysis is a fully grown and happily analyzed field of workstation innovative skill, particularly concerned with the development of record types, designs, and various regularities. Computer collecting information on which to hold PC in order to break down diverse types and decide on highly dependent choices.

Cluster Analysis: Clustering is a method for grouping large data sets of correlative comparisons on a solo basis. For record reasons or occasions, there is no predefined style name. Bundling organizations data times into subsets in such a way as to assemble comparable occurrences by and large, whereas explicit occasions have a place with unmistakable organizations and organizations are called as groups. Grouping could be organized into bunching partitioning, Hierarchical grouping, Density-based grouping, Model-based grouping, Grid-based clustering.

Correlation Analysis: Correlation is a method for exploring the relationship among two quantitative, nonstop variables. The connection coefficient (r) of Pearson is a level of the relationship's power between the two factors. It decides, in different expressions, the low connection between factors.

Statistical Analysis: A development of digital or semi-automated processes in realities to run over once in the past obscure types, consisting of links that can be used to anticipate large amounts of customers. There are computational obstacles to massive data assessment, the realities can be too big to keep memory in PC frameworks; and some of the figuring undertakings may take too long to even think about waiting for the impacts. With recently created factual philosophies and/or computational strategies, these limits can be moved toward both.

Regression Analysis: Regression assessment is a state-of - the-art demonstration technique that explores the connection between an established (objective) and unprejudiced (s) variable (predictor). Regression assessment is a basic device for measurement display and contemplation. For Big Data analysis, i.e., seven relapse systems are used. Linear Regression, Logistic Regression, Polynomial Regression, Stepwise Regression, Lasso Regression, Elastic Net Regression. The most common methods of big data evaluation are linear analysis and polynomial evaluation.

Analysis: A look at the colossal basics of data with public support and investigation for customer ponders Big data is shown with the three measurements guide Volume, Velocity and range. Extra calculations are expense and precision. Using unstructured documents, we need to search for the type of investigation to find a social example for the client. Discovery of this genius data comes at this point 9, 8, 10, 1, 5. For handling purposes, Hadoop is used with Mapreduce calculations for programming structure. The current condition of Apache Hadoop incorporates the Hadoop kernel, Mapreduce, HDFS, and quantities of various added substances such as Apache Hive, Base, and Zookeeper. Actually specifically endorsing publicly here could be combined with enormous guarantees using Amazon web administration technology such as Elastic Map lessen and Mechanical Turk to remove top K customer information queries from unsure actualities. Publicly supporting assistants in collecting pose insights that may have dynamic musings or the ability to issue check numbers to

commit to customer stories. Publicly supporting feedback from your consumers can allow you to consider what you could change[about] your item and a solution to better serve them, "crowd sourcing can help you analyze their dissatisfactions and what parts of your item do not perform as suggested. The process can be partitioned into Discovering customer appreciate event through literary substance analysis or through the use of literary substance research. YouTube and important advances in solution. From different perspectives, these records should be changed and used using mining, cleaning and displaying. Additionally, it can be completed from this assumption assessment that can provide us with customer securities and steadfastness. Reconciliation across spots of huge record structures wants to be accomplished. In the phase of data interpretation, certainties are pictured and documents imaginable for customers are made in which facts are examined and findings explained to decision-makers to decode the observations to distinguish meaning and information eight.

CONCLUSION

To deal with huge data and to work with it and getting profits by it a part of science advanced called Data Science. Data Science is the part of science that manages finding information from immense arrangements of data, for the most part unstructured and semi organized, by ideals of data surmising and investigation. It's an unrest that is changing the world and discovers application across different ventures like money, retail, medicinal services, assembling, sports and correspondence. Internet searcher and advanced promoting organizations like Google, Yahoo and Bing, person to person communication organizations like Facebook, Twitter and fund and web based business organizations like Amazon and EBay are requiring and will require loads of data researchers. Most definitely the current advances are promising to develop as fresher vulnerabilities to large data emerge and the requirement for making sure about them increments.

REFERENCE

[1] Mareketing Cloud, "10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations." 2017.

[2] W. E. Forum, Personal Data: The Emergence of a New Asset Class. 2017.

[3] "Gartner IT Glossary (n.d.). Retrieved from." [Online]. Available: http://www.gartner.com/it-glossary/big-data/ (2018)

[4] M. Fuchs, W. Höpken, and M. Lexhagen, "Big data analytics for knowledge generation in tourism destinations - A case from Sweden," J. Destin. Mark. Manag., 2018

[5] S. Sinha and A. Bansal, "A Framework for Effective Data Analytics for Tourism Sector: Big Data Approach," Int. J. Grid High Perform. Comput., vol. 9, no. 4, 2017

[6] R. Priyadarshini, R. K. Barik, C. Panigrahi, H. Dubey, and B. K. Mishra (2018). "An Investigation Into the Efficacy of Deep Learning Tools for Big Data Analysis in Health Care," Int. J. Grid High Perform. Comput., vol. 10, no. 3, 2018.

[7] Hadoop, "Apache Hadoop." 2018

[8] Apache Hadoop, "HDFS Architecture Guide."

[9] Hbase, "Apache Hbase.".

[10] Hive, "Apache Hive." Apache Hive.

[11] ZooKeeper, "Apache ZooKeeper.".

[12] Oozie, "Apache Oozie."

[13] M. Merabet, S. Mohamed Benslimane, M. Barhamgi, C. B. Lyon, and C. Bonnet (2018). "A Predictive Map Task Scheduler for Optimizing Data Locality in MapReduce Clusters," Int. J. Grid High Perform. Comput., Vol. 10, No. 4, 2018

[14] Cano (2018). "A survey on graphic processing unit computing for large-scale data mining," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2018

[15] D. Puthal, S. Nepal, R. Ranjan, et al. (2019) A dynamic prime number based efficient security mechanism for big sensing data streams, J. Comput. Syst. Sci., 83 (2017), pp. 22–42.

[16] Y. Zhe, M. Philip and R. Michael, (2019) Anomaly Detection Using Proximity Graph and PageRank Algorithm, IEEE T. Inf. Foren. Sec., 7 (2012), 1288–1300.

[17] T. D. Huynh, M. Ebden, J. Fischer, et al. (2018) Provenance Network Analytics: An approach to data analytics using data provenance, Data Min. Knowl. Disc., 32 (2018), pp. 708–735.

[18] K. P. Kibiwott, Y. Zhao, J. Kogo, et al. (2018) Verifiable fully outsourced attribute-based signcryption system for IoT eHealth big data in

cloud computing, Mathematical Biosciences and Engineering, 16 (2019), pp. 3561–3594.

Corresponding Author

Ruchi Sawhney*

Research Scholar, Department of Computer Science, Himalayan University, Itanagar, Arunachal Pradesh