World Wide Journal of Multidisciplinary Research and Development

WWJMRD 2017; 3(7): 206-209
www.wwjmrd.com
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

Deep Kumar
Software Engineer, Igniva
Solutions Private Limited,
Mohali, Punjab, India

A Review on Various Approaches Big Data Analysis and Security

Deep kumar

Abstract

Big Data is the emerging technology that has been used for various applications that contain huge memory data values. In this various organization has been adopting concept of big data that has been used under different applications so that maintenance, analysis and updating and deletion process is not easy through database tools. In this process various relations have been emerged between the instances and records available in the dataset. Due to availability of the large scale dataset security and confidentiality of the data are major concerns. In this paper various approaches and methods that can be used for security and availability concerns has been discussed. On the basis of these approaches optimum approach that can be used for classification process has been extracted that can be used in various applications of big data security.

Keywords: Hadoop, H-Base, Hive, Lucerne and Solar

Introduction

Big data

“Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.” (McKinsey Global Institute) “Big Data is the term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.” Every day, an organization create 2.5 quintillion bytes of data —so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. Such colossal amount of data that is being produced continuously is what can be coined as Big Data. Big Data decodes previously untouched data to derive new insight that gets integrated into business operations. However, as the amounts of data increases exponential, the current techniques are becoming obsolete. Dealing with Big Data requires comprehensive coding skills, domain knowledge and statistics. Web provides kind of opportunities for big data too. For example; social network analysis such as understanding user intelligence for more targeted advertising, marketing campaigns and capacity planning, customer behavior and buying patterns also sentiment analytics. According to these inferences firms optimization their content and recommendation engine some companies such as Google and Amazon publishing articles related to their work. Inspired by the writings published, developers are developing similar technologies as open source software such as Lucerne, Solar, Hadoop and H Base. Face book, Twitter and LinkedIn are going a step further thereby publishing open source projects for big data like Cassandra, Hive, Pig, Voldemort, Storm, Index Tank

Big Data processing

Big data technology is a new scientific trend. Driven by data analysis in high-dimension, big data technology works out data correlations to gain insight to the inherent mechanisms. Data-driven results only rely on an unrestrained selection of system raw data and a general statistical procedure (for data processing).On the other side, procedures for conventional model-based analysis, particularly decoupling practical interconnected system, are always based on assumptions and simplifications. Model-based results rely on identified causalities, specific parameters, sample selections, and training processes; imprecise or incomplete

Correspondence:
Deep Kumar
Software Engineer, Igniva
Solutions Private Limited,
Mohali, Punjab, India

lead to bad results. The results are often barely satisfied or even unsatisfied as the system size grows and complexity increases. Generally speaking, data-driven analysis tools, rather than model-based ones, are more suitable to complex large-scale interconnected systems with readily accessible data. Big data technologies are targeting to process high-volume, high-velocity, high-variety data assets to extract intended data value and ensure high-veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) new data models (supporting all data states and stages during the whole data life-cycle) and new infrastructure services and tools that allows also obtaining (and processing data) from a variety of sources and delivering data in a variety of forms to different data and information consumers and devices.

Application of big data

Big Data is slowly becoming ubiquitous. Every arena of business, health or general living standards now can implement big data analytics. To put simply, Big Data is a field which can be used in any zone whatsoever given that this large quantity of data can be harnessed to one‘s advantage. The major applications of Big Data have been listed below.

- The Third Eye-Data Visualization Organizations worldwide are slowly and perpetually recognizing the importance of big data analytics. From predicting customer purchasing behavior patterns to influencing them to make purchases to detecting fraud and misuse which until very recently used to be an incomprehensible task for most companies big data analytics is a one-stop solution. Business experts should have the opportunity to question and interpret data according to their business requirements irrespective of the complexity and volume of the data. In order to achieve this requirement, data scientists need to efficiently visualize and present this data in a comprehensible manner. Giants like Google, Face book, Twitter, EBay, Wal-Mart etc., adopted data visualization to ease complexity of handling data. Data visualization has shown immense positive outcomes in such business organizations. Implementing data analytics and data visualization, enterprises can finally begin to tap into the immense potential that big data possesses and ensure greater return on investments and business stability.
- Integration-An exigency of the 21st century integrating digital capabilities in decision -making of an organization is transforming enterprises. By transforming the processes, such companies are developing agility, flexibility and precision that enables new growth. Gartner described the confluence of mobile devices, social networks, cloud services and big data analytics as the as nexus of forces. Using social and mobile technologies to alter the way people connect and interact with the organizations and incorporating big data analytics in this process is proving to be a boon for organizations implementing it. Using this concept, enterprises are finding ways to leverage the data better either to increase revenues or to cut costs even if most of it is still focused on customer-centric outcomes. Such customer-centric objectives may still be the primary concern of most companies, a gradual shift to integrating big data technologies into the background operations and internal processes

Review of literature

PayamEzatpooret al [1] “Finding Top-k Dominance on Incomplete Big Data Using MapReduce Framework” This algorithm greatly improves the performance for incomplete data, but it is not designed to find top-k dominant values in incomplete big data. Several other algorithms have been proposed to find the TKD query, such as Sky-band Based and Upper Bound Based algorithms, but their performance is also questionable. Algorithms developed previously were among the first attempts to apply TKD query on incomplete data; however, these algorithms suffered from weak performance. This paper proposes Map Reduced Enhanced Bitmap Index Guided Algorithm (MRBIG) for dealing with the aforementioned issues. MRBIG uses the Map Reduce framework to enhance the performance of applying top-k dominance queries on large incomplete datasets. The proposed approach uses the MapReduce parallel computing approach involving multiple computing nodes. The framework separates the tasks between several computing nodes to independently and simultaneously work to find the result. This method has achieved up to two times faster processing time in finding the TKD query result when compared to previously proposed algorithms.

Leslie Mertz et al [2] “Machine Learning Takes on Health Care: Leonard D'Avolio'sCyft Employs Big Data to Benefit Patients and Providers” When Leonard D'Avolio was working on his Ph.D. degree in biomedical informatics, he saw the power of machine learning in transforming multiple industries; health care, however, was not among them. "The reason that Amazon, Netflix, and Google have transformed their industries is because they have embedded learning throughout every aspect of what they do. If we could prove that is possible in health care too, I thought we would have the potential to have a huge impact.

Suyan Long et al [3] “Information Service Research and Development of Digital Library in the Era of Big Data” A huge amount of data become an important strategic resource for enterprises and society, which can promote the various areas to begin a quantitative process. In the field of libraries, the rapid development of digital libraries has achieved cross-regional, object-oriented network query and dissemination, so that people at any time and place people can obtain the necessary information and make it into reality through the network. This has expanded the most important function of the traditional library, information service function. So in the era of large data, compared to traditional libraries, information service function of digital library are more prominent, which greatly promoted the sharing and utilization of resources.

Zhenyu Zhouet al [4] “Dependable Content Distribution in D2D-Based Cooperative Vehicular Networks: A Big Data-Integrated Coalition Game Approach”, the determination of content distribution groups with different lifetimes is formulated as a coalition formation game. We model the utility function based on the minimization of average network delay, which is transferable to the individual payoff of each coalition member according to its contribution. The merge and split process is implemented iteratively based on preference relations, and the final partition is proved to converge to a Nash-stable equilibrium. Finally, we evaluate the proposed algorithm based on real-world map and realistic vehicular traffic. Numerical results demonstrate that the proposed algorithm can achieve superior performance in terms of average network delay and content distribution efficiency compared with the other heuristic schemes.

Ivano Notarnicolaet al [5] “Distributed big-data optimization via block-iterative convexification and averaging” Author proposes a novel distributed solution method whereby at each iteration agents optimize and then communicate (in an uncoordinated fashion) only a subset of their decision variables. To deal with non-convexity of the cost function, the novel scheme hinges on Successive Convex Approximation (SCA) techniques coupled with i) a tracking mechanism instrumental to locally estimate gradient averages; and ii) a novel block-wise consensus-based protocol to perform local block-averaging operations and gradient tacking. Asymptotic convergence to stationary solutions of the no convex problem is established. Finally, numerical results show the effectiveness of the proposed algorithm and highlight how the block dimension impacts on the communication overhead and practical.

Approaches used

File Encryption

Since the data is present in the machines in a cluster, a hacker can steal all the critical information. Therefore, all the data stored should be encrypted. Different encryption keys should be used on different machines and the key information should be stored centrally behind strong firewalls. This way, even if a hacker is able to get the data, he cannot extract meaningful information from it and misuse it. User data will be stored securely in an encrypted manner.

Network Encryption

All the network communication should be encrypted as per industry standards. The RPC procedure calls which take place should happen over SSL so that even if a hacker can tap into network communication packets, he cannot extract useful information or manipulate packets.

Logging

The entire map reduce jobs which modify the data should be logged. Also, the information of users, which are responsible for those jobs should be logged. These logs should be audited regularly to find if any, malicious operations are performed or any malicious user is manipulating the data in the nodes.

Data privacy

Big Data analytics invades the personal privacy of individuals. The negative impact of the improper usage of the analysis that is made based on the Data may be Discrimination. The usage of predictive analytics to make determinations about user’s intelligence, habits, education, health, ability obtains a job, finance. The use of users associations in predictive analytics to make decisions that have a negative impact on individuals such opinions are automated, âAI and therefore more difficult to detect or prove, and may influence for example employment, promotions to fair housing and many more._ An embarrassment. Online shops and restaurants, government agencies, universities, online media corporations may be the reason for the personal information lockage resulting in revealing the personal information users, employees, especially very private information that people would like to keep separated from their business life (health problems, sexual orientation or an illnesses)

Technologies and strategies for Privacy Protection

Encryption algorithms, anonymization or data identification, deletion and non âAR retention methods helps to protect privacy. There are three basic types of cryptographic algorithms that were successfully adapted to B Dsystems: Cryptographic hash functions. Hash function produces a short (the length is always known) representation of any longer message. Hash function is a one-way function: it is easy to compute the8 Big Data security 169hash value from a particular input; but calculating the input from hash value is extremely computationally difficult or impossible. Hash function is also collision resistant: is also extremely difficult to find two particular inputs that produce the same hash value. Because of these feathers, hash functions used

Quantum cryptography and privacy with authentication for mobile data center

Quantum cryptography was proposed with GroverâAZs algorithm (GA), and Pair Hand authentication protocol, to asset secure communications between the mobile users and authentication servers. Proposed model includes several layers, and supports secure big data sending by mobile user to then earnest mobile data center. Data center front end Layer: verifications and identifications of the mobile user and big data using Quantum cryptography and authentication protocols. Data reading interface Layer: during each operation of the interface, provides the best performance to minimize the complexity_ Quantum key processing Layer: quantum key distribution (QKD) based on QC is taken into considerations, and the size of the big data and level of the security_ key management Layer: the size of the big data and traffic load, the security key generations is performed, protocols based on QC are applied

Conclusion

Big data is the extension of the data warehousing that has been used for management of small scale dataset. In today’s world due to advancements in the technology and new researches all the data has been stored in the digital manner that occupies large scale. Sensors have been designed that sense information after a particular interval of time. All the information that has been sensed through sensors has been stored over the large databases. Security is the major issue that has to resolve for better database management under big data. In this paper a review has been done on different approaches that have been used for data security. On the basis of review we can conclude that encryption based approaches provide better security to the big data rather than to password protection approaches.

References

1. Payam Ezatpoor“Finding Top-k Dominance on Incomplete Big Data Using MapReduce Framework”, IEEE Access, 23 January 2018

2. Leslie Mertz "Machine Learning Takes on Health Care: Leonard D'Avolio'sCyft Employs Big Data to Benefit Patients and Providers" IEEE Pulse, 23 January 2018, pp. 10-11
3. Suyan Long"Information Service Research and Development of Digital Library in the Era of Big Data", 2018, pp. 114-120
4. Zhenyu Zhou; Houjian Yu; Chen Xu; Yan Zhang; ShahidMumtaz; Jonathan Rodriguez "Dependable Content Distribution in D2D-Based Cooperative Vehicular Networks: A Big Data-Integrated Coalition Game Approach" IEEE Transactions on Intelligent Transportation Systems, 2018, pp. 1-12.
5. Ivano Notarnicola; Ying Sun; Gesualdo Scutari; Giuseppe Notarstefano, "Distributed big-data optimization via block-iterative convexification and averaging", Decision and Control (CDC), 2017
6. Mostafa Rahmani; George Atia "Robust and Scalable Column/Row Sampling from Corrupted Big Data" Computer Vision Workshop (ICCVW), 2017, pp. 112-125
7. Jeremy J. Harwood "Spectral ageing in the era of big data: integrated versus resolved models", Monthly Notices of the Royal Astronomical Society, pp. 2888 – 2894
8. Patrick Wils; Boris T. Gänsicke; Andrew J. Drake; John Southworth, "Data mining for dwarf novae in SDSS, GALEX and astrometric catalogues", Monthly Notices of the Royal Astronomical Society, 2010, pp. 436 – 446
9. R. J. Dodd Monthly Notices of the Royal Astronomical Society, 2016, pp. 959 – 972.
10. Vid Podpečan; Monika Zemenova; Nada Lavrač "Orange4WS Environment for Service-Oriented Data Mining" The Computer Journal, 2012, pp. 82 – 98