IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, August 2022
Impact Factor: 6.252

A Survey on: E-Commerce Data Analysis and Security Platform in the Era of Bigdata

Kalpana Gampa[1] and Dr.Vemuri Madhukar[2]
Ph.D Scholar, Department of Computer Science [1]
Research Supervisor, Department of Computer Science [2]
Chaitanya Deemed to be University, Hanamkonda, Telangana, India

Abstract: *With the rapid development of e-commerce and mobile communication, e-commerce platform has been widely used in various industries. How much e-commerce stores, whether it can guarantee the security of transaction information, whether it can analyze and study structured and unstructured data, and whether it can guarantee the security of stored data are all the key factors we need to consider. In this paper, the data and e-commerce security together, and to analyze the security system of e-commerce and discuss the prevention of hidden security policy. When the emergence of e-commerce big data technology can effectively solve the problems existing in e-commerce security, the Hadoop structure is introduced through Apache Hadoop and the Hadoop product Yarn is analyzed with emphasis. From the perspective of electronic security data, the hidden dangers of e-commerce can be effectively analyzed, and the security system of e-commerce can be effectively improved. This article starts with the analysis of the existing electronic commerce system, summarizes its characteristics, and analyzes and solves its existing problems. Firstly, the characteristics of the relational database My Structured Query Language (MySQL) and the distributed database HBase are analyzed, their respective advantages and disadvantages are summarized, and the advantages and disadvantages of each are taken into account when storing data. My SQL is used to store structured business data in the system, while HBase is used to store unstructured data such as pictures. These two storage mechanisms together constitute a data storage subsystem. Secondly, considering the large amount of data in the e-commerce system and the complex calculation of the data mining algorithm, this paper uses Map Reduce to realize the parallelization of the data mining algorithm and builds a Hadoop-based commodity recommendation subsystem on this basis. We use JavaEE technology to design a full-featured web mall system. Finally, based on the impact of cloud computing, mobile e-commerce is analyzed, including relevant theories, service mode, architecture, core technology, and the application in e-commerce, which can realize e-commerce precision marketing, find the optimal path of logistics, and take effective security measures to avoid transaction risks. This method can avoid the disadvantages of the traditional e-commerce, where large-scale data cannot be processed in a timely manner, realize the value of mining data behind, and realize the precision marketing of e-commerce enterprises.*

Keywords: Data Analysis.

I. INTRODUCTION

Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*. The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Thus a fourth concept, *veracity,* refers to the quality or insightfulness of the data. Without sufficient investment in expertise for big data veracity, then the volume and variety of data can produce costs and risks that exceed an organization's capacity to create and capture *value* from big data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, August 2022
Impact Factor: 6.252

Current usage of the term *big data* tends to refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem." Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on". Scientists, business executives, medical practitioners, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet searches, fintech, healthcare analytics, geographic information systems, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, biology, and environmental research.

The size and number of available data sets have grown rapidly as data is collected by devices such as mobile devices, cheap and numerous information-sensing Internet of things devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×2^{60} bytes) of data are generated. Based on an IDC report prediction, the global data volume was predicted to grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of data. One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.

Relational database management systems and desktop statistical software packages used to visualize data often have difficulty processing and analyzing big data. The processing and analysis of big data may require "massively parallel software running on tens, hundreds, or even thousands of servers".What qualifies as "big data" varies depending on the capabilities of those analyzing it and their tools. Furthermore, expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

II. MOTIVATION

Big Data market is constantly increasing each year. In March 2012, The White House announced a national “Big Data Initiative” that consisted of six Federal departments and agencies committing more than $200 million to big data research projects [6]. Global Pulse which is an innovative lab that is based on the big data mining is also using the Big data to improve the life in developing countries. In today’s competitive & complex business world the various aspects of business are intermingled. Change in one aspect has direct or indirect effect on the other aspect. Within an organization, this complexity makes it difficult for business leaders to rely solely on experience (or intuition) to make decisions.

They need to rely on data - structured, unstructured or semi-structured - to back up their decisions. Existing tools don’t lend themselves to sophisticated data analysis at the scale the user requires. Tools like SAS, R, and Matlab support the decisive analysis but they are not designed for the massive datasets & neither DBMS nor Map Reduce can handle the data that are arrived at high rates. To bridge this gap the “Big Data” came into the scene. Big Data has given the organization a new way to analyze and visualize their data effectively. For example: Business: Customer Feedback, trends etc.

Health: Health care organizations are leveraging big data technology to capture all the information about a patient to get more complete view for insight into care coordination, health management & outcome. Use of big data helps to build a sustainable healthcare system & increase the access to healthcare. Energy & utility: Big data can also be the key to actually deploying condition based maintenance program and improve forecasting and scheduling of assets.

III. STATEMENT OF THE PROBLEM

Developing, implementing and enforcing data security best practices is made easier if organizations fully understand the privacy and compliance mandates to which they must adhere.The California Consumer Privacy Act (CCPA) went into effect January of this year. It enforces consumers' rights to control their personal information. Many experts believe a version of the CCPA will likely become federal law. CCPA itself is a take on the European Union's General Data Protection Regulation, which also protects consumers' personal data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, August 2022
Impact Factor: 6.252

While companies worry that the cost to comply with government mandates could be prohibitive, many are still going forward in their efforts to ensure data is able to be discovered, reported on and erased. That way, when consumers request to see their data and then delete it, businesses will be ready.

To follow the multiple compliance mandates, organizations can create a data inventory, establish processes to get consumers their information under deadline and make updates to the organization's privacy statement.

3.1 The Future of Data Security

AI and machine learning are going to be key in compliance efforts going forward. Companies are looking to automate some regulatory compliance processes, including data location and extraction. Inventories, as security expert Michael Cobb noted, become outdated unless automated scanning tools are deployed to sustain data discovery capture by recording regular snapshots of all applications and repositories where personal information resides. Automation, in his opinion, is the only way large organizations can remain compliant with a large volume of data that is structured and unstructured and stored in data centers and in the cloud.

Next-generation technology could also help companies fall in line with other compliance mandates, such as PCI DSS. For companies that have lagged behind on compliance, some security experts suggest considering a zero-trust model as a security strategy. With zero trust, companies would look at the full lifecycle of data management and broaden their focus beyond just payment card data to other forms of personal data, including financial data, intellectual property and customer data. They would make no assumptions on where data is expected to be found or how it is being used -- only that the risk must be mitigated.

Data security will remain a significant challenge well into the future, but creative applications of AI and machine learning and zero-trust models will help IT and info sec teams protect data and ensure consumer privacy.

IV. METHODOLOGY

4.1 Safeguard Distributed Programming Frameworks

To begin, create trust by using methods like Kerberos Authentication and ensuring that predefined security policies are followed. The data is then "de-identified" by decoupling all personally identifiable information (PII) from it, ensuring that personal privacy is not jeopardized.

Then, using mandatory access control (MAC), such as the Sentry tool in Apache HBase, you allow access to files based on a predefined security policy and ensure that untrusted code does not leak information through device resources. After that, the hard part is done; all left is to maintain the system to prevent data leakage. In a cloud or virtual environment, the IT department should be scanning worker nodes and mappers for bogus nodes and altered duplicates of results.

4.2 Secure Non-Relational Data

Non-relational databases, such as NoSQL, are common, but they're vulnerable to NoSQL injection attacks. By encrypting or hashing passwords and maintaining end-to-end encryption by using algorithms such as advanced encryption standard (AES), RSA, or Safe Hash Algorithm 2.

With the advent of Big Data, the structured approach fails miserably to cater to the needs of the humongous information processing that tends to be unstructured in nature. SQL vs NoSQL vs NewSQL: The Full Comparison

4.3 Secure Data Storage and Transaction Logs

Storage control is a critical component of Big Data reliability. By using signed message digests to have a cryptographic identifier for each digital file or record and to use a technique known as a secure untrusted data repository (SUNDR) to detect unauthorized file modifications by malicious server agents

4.4 Endpoint Filtering and Validation

Using a mobile device management solution, you can use trusted credentials, perform resource verification, and link only trusted devices to the network. Using statistical similarity detection and outlier detection strategies, you can process malicious inputs while defending against Sybil attacks (one person posing as several identities) and ID-spoofing attacks.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, August 2022
Impact Factor: 6.252

4.5 Real-Time Compliance and Security Monitoring

Organizations can use techniques like Kerberos, safe shell, and internet protocol protection to get a grip on real-time data by using Big Data analytics. It's then simple to monitoring logs, set up front-end security mechanisms like routers and server-level firewalls, and start putting security controls in place at the cloud, network, and application levels.

Graph Databases uses graph architecture for semantic inquiry with nodes, edges, and properties to represent and store data. Role of Graph Databases in Big Data Analytics

4.5 Preserve Data Privacy

Employee awareness training centers on new privacy laws and ensures that information technology is kept up to date by using authorization processes. In addition, data leakage from different databases can be regulated by analyzing and tracking the infrastructure that connects the databases.

4.6 Big Data Cryptography

Mathematical cryptography has improved significantly. Enterprises can run Boolean queries on encrypted data by creating a method to scan and filter encrypted data, such as the searchable symmetric encryption (SSE) protocol.

V. RELATED WORK

5.1 Distributed Data

Most big data frameworks distribute data processing tasks throughout many systems for faster analysis. Hadoop, for example, is a popular open-source framework for distributed data processing and storage. Hadoop was originally designed without any security in mind.

Cyber criminals can force the Map Reduce mapper to show incorrect lists of values or key pairs, making the Map Reduce process worthless. Distributed processing may reduce the workload on a system, but eventually more systems mean more security issues.

A. Non-Relational Databases

Traditional relational databases use tabular schema of rows and columns. As a result, they cannot handle big data because it is highly scalable and diverse in structure. Non-relational databases, also known as NoSQL databases, are designed to overcome the limitations of relational databases.

Non-relational databases do not use the tabular schema of rows and columns. Instead, NoSQL databases optimize storage models according to data type. As a result, NoSQL databases are more flexible and scalable than their relational alternatives.

NoSQL databases favor performance and flexibility over security. Organizations that adopt NoSQL databases have to set up the database in a trusted environment with additional security measures.

5.2 Endpoint Vulnerabilities

Cybercriminals can manipulate data on endpoint devices and transmit the false data to data lakes. Security solutions that analyze logs from endpoints need to validate the authenticity of those endpoints.

For example, hackers can access manufacturing systems that use sensors to detect malfunctions in the processes. After gaining access, hackers make the sensors show fake results. Challenges like that are usually solved with fraud detection technologies.

5.3 Data Mining Solutions

Data mining is the heart of many big data environments. Data mining tools find patterns in unstructured data. The problem is that data often contains personal and financial information. For that reason, companies need to add extra security layers to protect against external and internal threats.

5.6. Access Controls

Companies sometimes prefer to restrict access to sensitive data like medical records that include personal information. But people that do not have access permission, such as medical researchers, still need to use this data. The solution in

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, August 2022
Impact Factor: 6.252

many organizations is to grant granular access. This means that individuals can access and see only the information they need to see. Big data technologies are not designed for granular access. A solution is to copy required data to a separate big data warehouse. For example, only the medical information is copied for medical research without patient names and addresses.

5.7 Addressing Big Data Security Threats

Security tools for big data are not new. They simply have more scalability and the ability to secure many data types. The list below explains common security techniques for big data.

A. Encryption

Big data encryption tools need to secure data-at-rest and in-transit across large data volumes. Companies also need to encrypt both user and machine-generated data. As a result, encryption tools have to operate on multiple big data storage formats like NoSQL databases and distributed file systems like Hadoop.

B. User Access Control

User access control is a basic network security tool. The lack of proper access control measures can be disastrous for big data systems. A robust user control policy has to be based on automated role-based settings and policies. Policy-driven access control protects big data platforms against insider threats by automatically managing complex user control levels, like multiple administrator settings.

C. Intrusion Detection and Prevention

The distributed architecture of big data is a plus for intrusion attempts. An Intrusion Prevention System (IPS) enables security teams to protect big data platforms from vulnerability exploits by examining network traffic. The IPS often sits directly behind the firewall and isolates the intrusion before it does actual damage.

D. Centralized Key Management

Key management is the process of protecting cryptographic keys from loss or misuse. Centralized key management offers more efficiency as opposed to distributed or application-specific management. Centralized management systems use a single point to secure keys and access audit logs and policies. A reliable key management system is essential for companies handling sensitive information.

REFERENCES

[1]. Lv Z, Song H, Basanta-Val P, et al. Next-generation big data analytics: State of the art, challenges, and future research topics. IEEE Transactions on Industrial Informatics, 2017, 13(4): 1891-1899.

[2]. Song M L, Fisher R, Wang J L, et al. Environmental performance evaluation with big data: Theories and methods. Annals of Operations Research, 2018, 270(1-2): 459-472.

[3]. Chen C L P, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 2014, 275: 314-347.

[4]. Diamantoulakis P D, Kapinas V M, Karagiannidis G K. Big data analytics for dynamic energy management in smart grids. Big Data Research, 2015, 2(3): 94-101.

[5]. De Hert P, Papakonstantinou V. The new General Data Protection Regulation: Still a sound system for the protection of individuals?. Computer Law & Security Review, 2016, 32(2): 179-194.

[6]. Li J Q, Yu F R, Deng G, et al. Industrial internet: A survey on the enabling technologies, applications, and challenges. IEEE Communications Surveys & Tutorials, 2017, 19(3): 1504-1526.

[7]. Ting Yuan. Research on issues related to information Security Management of B2C E-commerce Platform [D]. Heilongjiang University,2016.

[8]. Yanyan Lu. Research on Key Technologies of Big data Storage based on Hadoop [D]. North China Electric Power University,2016.

[9]. Wenmin Lin. Research on Big Data Service and its Key Technologies in cloud Environment [D]. Nanjing University,2015.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, August 2022
Impact Factor: 6.252

[10]. Jiaqi Fan. Data Mining Engine based on Big Data [D]. Beijing University of Posts and Telecommunications,2015.

[11]. Xiu Wang. Research on e-commerce Security Risk Assessment Model in cloud Computing Environment [D]. Anhui University of Finance and Economics,2015.

[12]. Hua Wang. Research on Computer E-commerce Security in the Era of Big Data [J]. Information Communications,2019(09):166-167.