Intl J Engg Sci Adv Research 2018 March; 4(1): 83-89 ISSN NO: 2395-0730

Challenges to Big Data Security Analytics

Kuldeep Kumar Katiyar
Department of Computer Science & Engineering
Rama University, Uttar Pradesh, Kanpur, India
kuldeepsrms@gmail.com

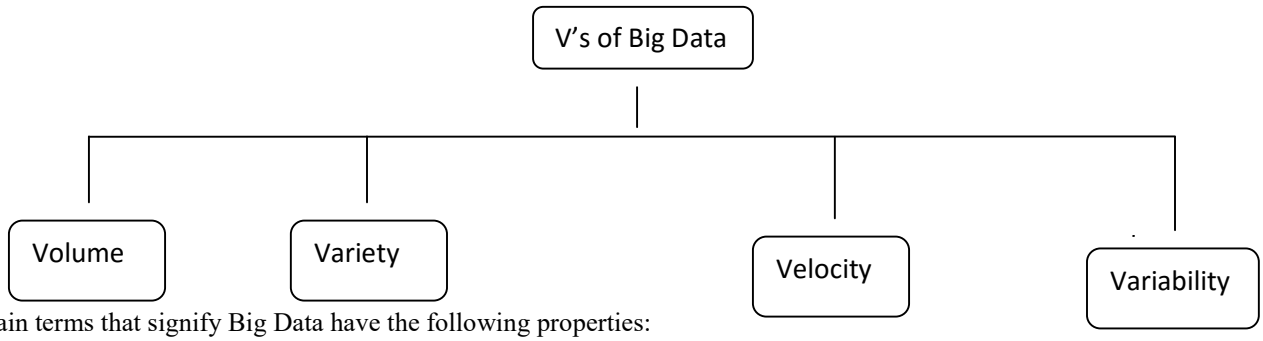
Intl J Engg Sci Adv Research 2018 March; 4(1): 83-89 ISSN NO: 2395-0730

Abstract: Data with high volume, variety, velocity, variability and veracity termed as Big Data. Big data is usually processed in distributed environment with a number of connected machines supporting applications typically termed as Big Data Analytics. However, the amount of sensitive data typically processed in typical Big Data Analytics has made Big Data Analytics applications an eye catch to anomalous users. Processing big volume of data in distributed environment also makes it an attractive prey. The vulnerabilities involve in that as well as situations in which these vulnerabilities arises. We focused on the security aspects which arises in the practical environment in industries and organizations. Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations, fraud detection, network forensics, data privacy issues and data provenance problems.

Keywords:*Big Data Analytics*, *HADOOP,HDFS,MapReduce,SecurityIssues,Tools.*

INTRODUCTION

With the growth of technology,there is a huge expansion in the data generation and its exchange over the Internet.Big data is a buzzword, or catch-phrase, utilizes to describe a massive volume of both structured and unstructured data that is so huge that it's complicated to process using traditional database and software techniques. In most enterprise scenarios the data is too large or it moves too fast or it exceeds current processing capacity. Big data has the potential to help organizations to improve operations and make faster, more intelligent decisions. The term “Big Data is believed to be originated from the Web search companies who had to query loosely structured very large distributed data.Evolution of big data brings many security issues with it.Traditional security mechanisms are designed for securing small and static data –sets.Big data is efficiently processed in distributed environments instead of single machine.Algorithms like *Map-Reduce* and *SCOPE* became the base of big data processing in distributed environment.



The three main terms that signify Big Data have the following properties:

a) *Volume*: Many factors contribute towards increasing Volume- storing transaction data, live streaming data and data collected from sensors etc.,

b) *Variety*: Today data comes in all types of formats - from traditional databases, text documents, emails, video, audio, transactions etc.,

c) *Velocity*: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.

The other two dimensions that need to consider with respect to Big Data are Variability and Complexity.

d) *Variability*: Along with the Velocity, the data flows can be highly consistent with periodic peaks.
e) *Complexity*: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

The challenges include analysis, capture, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (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 Exabyte (2.5×10^{18}) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization. One of the efficient well-known technologies that deal with the Big Data is Hadoop.

HADOOP

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Hadoop is open-source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers.
Hadoop is:

[1] *Reliable*: The software is fault tolerant, it expects and handles hardware and software failures.
[2] *Scalable*: Designed for massive scale ofprocessors, memory, and local attached storage.
[3] *Distributed*: Handles replication. Offers massively parallel programming model, Map Reduce.

Hadoop has two main sub projects – Map Reduce and Hadoop Distributed File System (**HDFS**)

MAP REDUCE

Hadoop Map Reduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework.

HADOOP DISTRIBUTED FILE SYSTEM

HDFS is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures.

The big data application refers to the large scale distributed applications which usually work with large data sets. Data exploration and analysis turned into a difficult problem in many sectors in the span of big data. With large and complex data, computation becomes difficult to be handled bythe traditional data processing applications which triggers the development of big data applications. Google's map reduce framework and apache Hadoop are the defect software systems for big data applications, in which these applications generates a huge amount of intermediate data. Manufacturing and Bioinformatics are the two major areas of big data applications. Big data provide an infrastructure for transparency in manufacturing industry, which has the ability to unravel uncertainties such as consistent component performance and availability. In these big data applications, a conceptual framework of predictive manufacturing begins with data acquisition where there is a possibility to acquire different types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data. The combination of sensory data and historical data constructs the big data in manufacturing. This generated big data from the above combination acts as the input into predictive tools and preventive strategies such as prognostics and health management. Another important application for Hadoop is Bioinformatics which covers the next generation sequencing and other biological domains. Bioinformatics which requires a large scale data analysis, uses Hadoop. Cloud computing gets the parallel distributed computing framework together with computer clusters and web interfaces.

SECURITY ISSUES

Big data originates from multiple sources including 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. Thanks to cloud computing and the socialization of the Internet, petabytes of unstructured data are created daily online and much of this information has an intrinsic business value if it can be captured and analyzed.

For example, mobile communications companies collect data from cell towers; oil and gas companies collect data from refinery sensors and seismic exploration; electric power utilities collect data from power plants and distribution systems. Businesses collect large amounts of user-generated data from prospects and customers including credit card numbers, social security numbers, data on buying habits and patterns of usage.

The influx of big data and the need to move this information throughout an organization has created a massive new target for hackers and other cybercriminals. This data, which was previously unusable by organizations is now highly valuable, is subject to privacy laws and compliance regulations, and must be protected. Following are the security issues related to Big Data and Hadoop Technology:

[17] Hadoop is Not a Secure Technology:

Hadoop, like many open source technologies such as UNIX and TCP/IP, was not created with security in mind. Hadoop evolved from other open-source Apache projects, directed at building open source web search engines. Hadoop was a spin off sub-project of Apache Lucene and Nutch projects, which used a Map Reduce facility and a distributed file system with no built-in security. Hadoop is also the open-source version of the Google Map Reduce framework, and no security was designed into the software as the data being stored (publicURLs) was not subject to privacy regulation

The open source Hadoop community supports some security features through the current implementation of Kerberos, the use of firewalls, and basic HDFS permissions. Kerberos is not a mandatory requirement for a Hadoop cluster, making it possible to run entire clusters without deploying any security. Kerberos is also difficult to install and configure on the cluster, and to integrate with Active Directory (AD) and Lightweight Directory Access Protocol, (LDAP) services. This makes security problematic to deploy, and thus constrains the adoption of even the most basic security functions for users of Hadoop.

Enterprise organizations have been subjected to the risks associated with data security breaches for decades now, and expect that any new technology that is adopted by IT and installed in the datacenter will meet a minimum set of security requirements. Enterprises want the same security capabilities for big data as that in place for "non-big data" information systems, including solutions that address user authentication and access control, policy enforcement and management, and data masking and encryption. To date, the open source community has not addressed these security gaps, and remains focused on creating improved Hadoop technologies such as Map Reduce 2.0.

The characteristics of Hadoop's distributed computing architecture present a unique set of challenges for datacenter managers and security professionals.

• *Distributed computing* - Data is processed anywhere resources are available, enabling massively parallel computation. This creates complicated environmentsthat are highly vulnerable to attack, as opposed to the centralized repositories that are monolithic and easier to secure.

• *Fragmented data* - Data within big data clusters is fluid, with multiple copies moving to and from different nodes to ensure redundancy and resiliency. Data can become sliced into fragments that are shared across multiple servers. This fragmentationAdds more complexity to the security challenge.

• *Access to data* - Role-Based Access Control (RBAC) is central to most database security frameworks, but most big data environments only offer access control at the schema level, with no finer granularity to address users by role and related access.

• *Node-to-node communication* - Hadoop and the vast majority of distributions don't communicate securely; they use RPC over TCP/IP.

• Virtually no security - Big data stacks build in almost no security. Aside from service-level authorization and web proxy capabilities from YARN, no facilities are available to protect data stores, applications, or core Hadoop features. All bigData installations are built on the web services model, with few or no facilities for countering common web threats.

[18] Secure Computations in Distributed Programming frameworks:

Distributed programming frameworks utilize parallelism in computation and storage to process massive amounts of data. A popular example is the MapReduce framework, which splits an input file into multiple chunks. In the first phase of MapReduce, a Mapper for each chunk reads the data, performs some computation and outputs a list of key/value pairs. In the next phase, a Reducer combines the values belonging to each distinct key and outputs the result. There are two major attack measures: securing the mappers and securing the data in the presence of an untrusted mapper. Untrusted mappers could return wrong results, which will in turn generate incorrect aggregate results, with large data sets, it is impossible to identify, resulting in significant damage, especially for scientific and financial computations.

[19] Real time Security:

Real time security monitoring has always been a challenge, given the number of alerts by devices. These alerts lead to many false positives, which are mostly ignored as humans cannot cope with the shear amount. This problem might increase with big data given the volume and velocity of data streams. Big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data. Which provide real time anomaly detection based on scalable analytics.

[20] Data Security:

Data security is essential for many corporate applications. Data warehouse users are accustomed not only to carefully defined metrics and dimensions and attributes, but also to a reliable set of administration policies and security controls. These rigorous processes are often lacking with unstructured data sources and open source analysis tools. Pay attention to the security and data governance requirements of each analysis project and make sure that the tools you are using can accommodate those requirements.

TOOLS FOR ANALYZING BIG DATA

There are five key approaches to analyzing big data and generating insight:

• *Discovery tools* are useful throughout the information lifecycle for rapid, intuitive exploration and analysis of information from any combination of structured and unstructured sources. These tools permit analysis alongside traditional BI source systems. Because there is no need for up-front modeling, users can draw new insights, come to meaningful conclusions, and make informed decisions quickly.

• *BI tools* are important for reporting, analysis and performance management, primarily with transactional data from data warehouses and production information systems. BI Tools provide comprehensive capabilities for business intelligence and performance management, including enterprise reporting, dashboards, ad-hoc analysis, scorecards, and what-if scenario analysis on an integrated, enterprise scale platform.

• *In-Database Analytics* include a variety of techniques for finding patterns and relationships in your data. Because these techniques are applied directly within the database, you eliminate data movement to and from other analytical servers, which accelerates information cycle times and reduces total cost of ownership.

• *Hadoop* is useful for pre-processing data to identity macro trends or fined nuggets of information, such as out of-range values. It enables businesses to unlock potential value from new data using inexpensive commodity servers. Organizations primarily use Hadoop as a precursor to advanced forms of analytics.

• *Decision Management* includes predictive modeling, business rules, and self-learning to take informed action based on the current context. This type of analysis enables individual recommendations across multiple channels, maximizing the value of every customer interaction. Oracle Advanced Analytics scores can be integrated to operationalize complex predictive analytic models and create real-time decision processes.

CONCLUSION

Using big data tools to analyze the massive amount of threat data received daily, and correlating the different components of an attack, allows a security vendor to continuously update their global threat intelligence and equates to improved threat knowledge

and insight. Customers benefit through improved, faster, and broader threat protection. By reducing risk, they avoid potential recovery costs, adverse brand impacts, and legal implications.

REFERENCES

[1] Alvaro A. Cárdenas, Pratyusa K. Manadhata, Sreeranga P. Rajan, "Big Data Analytics for Security" in IEEE Security & Privacy, IEEE, pp. 74-76, 2013.

[2] An Oracle White Paper March 2013, Big Data Analytics - Advanced Analytics in Oracle Database.

[3] A Trend Micro White Paper, September 2012, Addressing Big Data Security Challenges: The Right Tools for Smart Protection.

[4] Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri, security issues associated with big data in cloud computing, International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.

[5] M. Chithik Raja, Munir Ahmed Rabbani, Big Data Analytics Security Issues in Data Driven Information System, International Journal of Innovative Research in Computer and Communication Engineering *(An ISO 3297: 2007 Certified Organization)* Vol. 2, Issue 10, October 2014.

[6] Katal, M. Wazid, R. Goudar, "Big data Issues challenges tools and good practices", *the sixth international conference on contemporary computing*, pp. 404-409, Aug. 2013.