INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

BIG DATA HOLDS BIG PROMISE FOR SECURITY

NEHA S. PAWAR, PROF. S. P. AKARTE

Computer Science and Engineering, Prof. Ram Meghe Institute of Technology and Research, Badnera.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Big data are a collection of data sets so large and complex that it become difficult to process using on-hand database management tools or traditional data processing applications. Big Data extends the advanced security capabilities of Databases such as redaction, privilege controls, and virtual private database to limit privileged user access to Hadoop and NoSQL data**.** While big data analytics tools for security were often custom built in the past this year leading security organizations will deploy commercial, of the shelf big data solution in their SOCs. We predict big data analytics will have distruptive impact on many categories in the information security purpose. This paper is about big data security, as big data contains large. Voluminous unstructured and structured data therefore security is one of most important factor, we have mentioned security issues, approaches and their solution.

Keywords: Big data, Apache Hadoop, NoSQL, Encryption, Authentication, Authorization.

Corresponding Author: MS. NEHA S. PAWAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Neha S. Pawar, IJPRET, 2015; Volume 3 (9): 1289-1295

PAPER-QR CODE

INTRODUCTION

Big Data is about quickly deriving business value from a range of New and emerging data sources, including social media data, location data Generated by smart phones and other roaming devices, public information Available online and data from sensors embedded in cars, buildings and other objects and much more besides.

Big Data has also been defined by the four "V"s: Volume, Velocity, Variety, and Volume.

Volume:- While volume indicates more data, it is the granular nature of the data that is unique. Big Data requires processing high volumes of low-density data, that is, data .of unknown value. It is the task of Big Data to convert low-density data into high-density data, that is, data that has value.

Velocity: - A fast rate that data is received and perhaps acted upon. Reflects the sheer speed at which this data is generated and

Changes.

Variety:- Big Data can come from many different sources, in various formats and structures. such as text, audio, and video require additional processing to both derive meaning and the supporting metadata.

Value:- Data has intrinsic value—but it must be discovered. There are a range of quantitative and investigative techniques to derive value from data.

In addition, Big Data has popularized two foundational storage and processing technologies: Apache Hadoop and the NoSQL database.

Big data is transforming the global business landscape. Organizations are alaysing huge amount of diverse, fast changing data and new technologies that help them to grow their business better. For all of this security is one of the big factor. In this paper we have introduced few security approaches, security issues and security solutions releated to the big data.

2. Storage and Processing technologies of big data.

Big Data has two storage and processing technologies: Apache Hadoop and the NoSQL database

2.1 NoSQL database:-A NoSQL database is capable of doing much more than some think. The "No" part of the NoSQL label can be thought of as "not only SQL," which communicates the fact that a NoSQL database doesn't completely discard all features/functions that define a relational

database. NoSQL databases is that they don't conform to the standard Codd-Date relational model 2, where data is normalized to a third logical form. Such data structures often require resource-intensive join operations to satisfy end user requests. Instead, data in a NoSQL database is greatly denormalized and resides in structures organized in a variety of formats (e.g., columnar, document, key/value, and graph).

2.2 Apache Hadoop:-Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation inspired by google's Map Reduce.

Hadoop makes it possible to run applications on systems with thousands of nodes involving thousands of terabytes. Its distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating uninterrupted in case of a node failure.

3. Big Data Security:-

As Big Data is used to describe massive volumes of structured and unstructured data therefore the Big Data ecosystem must be secure. Data security approach ensures the right people, internal or external, get access to the appropriate data and information at right time and place, within the right channel. There are several security approaches, issues and solutions which are mentioned below.

3.1:- Big data security approaches:-

Th**ere** are four security approaches which are mentioned below.

1. Security management:- Security management driven by big data analysis creates a unified view of multiple data source and centralizes threat research capabilities, instead of forcing security analyst to deal with disparate tools that disrupt and potentially derail their workflows. The convergence of SIEM and network monitoring capabilities creates a unified security management system to assimilate all information that could possibly inform security. It investigates threat intelligence and also offers the flexibility to integrate security data from existing technologies.

2. Identity and access management: Next approach enable risk-based, adaptive identity controls that continuously evaluate and adjust the level of protection and access based on asset critically and risk. It provide continuous risk assessment of user activity, especially when accessing sensitive resources, even after initial authentication. Profiles are based on historical

behavior; a deep, complex user profile; a rich view of identities and a data-driven perspective of what normal behavior looks like.

3. Fraud prevention: In this approach, financial fraud, transaction fraud or the fraudulent use of corporate recourses advanced security technologies analyze massive amount of behavioral data and other diverse indicator which distinguishes between malicious and legitimate business activities. It can predict that session intelligence and behavioral and click string analysis will combine to stop business logic abuse in which attacker find of law in functioning of an IT based system and exploit it for illicit gain.

4. Governance, Risk, and Compliance (GRC): GRC platform enables us to understand business risk and to priorities security activities in business. They analyze large volumes of data to facilate better, smarter decision about the level. They also inform about valuable assets that are at high levels of risk and help priorities steps that an organization should take to solve risks.

3.2 SECURITY ISSUES

Hadoop present some unique set of security issues for data centre managers and security professionals. The security issues are depicted below:

1. Fragmented Data: Big Data clusters contain data that portray the quality of fluidity, allowing multiple copies moving to-and-fro various nodes ensuring redundancy and resiliency. As a result, more complexity is added as a result of the fragmentation which poses a security issue due to the absence of a security model.

2. Availability:- The availability of Resources leads to virtual processing of data at any instant or instance where it is available, these progresses to large levels of parallel computation. As a result, complicated environments are created that are at high risks of attacks than their counterparts of repositories that are centrally managed and monolithic, which enables easier security implications.

3. Controlling Data Access: Commissioned data environments provision access at the schema level, devoid of finer granularity in addressing proposed users in terms of roles and access related scenarios. Many of the available database security schemas provide role based access.

4. Node-to-node communication: A concern with Hadoop and a variety of players available in this field is that, they don't implement secure communication; they bring into use the RPC (Remote Procedure Call) over TCP/IP.

5. Client Interaction: Communication of client takes place with resource manager, data nodes. However, there is a catch. Even though efficient communication is facilitated by this model, it makes cumbersome to shield nodes from clients and vice-versa and also name servers from nodes. To either service.

3.3 Security solutions:

1. Authentication:-

Authentication is verifying user or system identity accessing the system. Hadoop provides Kerberos as a primary authentication. Initially SASL/GSSAPI was used to implement Kerberos and mutually authenticate users, their applications, and Hadoop services over the RPC connections. Hadoop also supports "Pluggable" Authentication for HTTP Web Consoles meaning that implementers of web applications and web consoles could implement their own authentication mechanism for HTTP connections. This includes but was not limited to HTTP SPNEGO authentication. The Hadoop components support SASL Framework i.e. the RPC layer can be changed to support the SASL based mutual authentication viz. SASL Digest-MD5 authentication or SASL GSSAPI/Kerberos authentication.

2. Authorization and ACLs:-

Authorization is a process of specifying access control privileges for user or system. In Hadoop, access controls is implemented by using file-based permissions that follow the UNIX permissions model. Access control to files in HDFS could be enforced by the Name Node based on file permissions and ACLs of users and groups. Map Reduce provides ACLs for job queues; that define which users or groups can submit jobs to a queue and change queue properties. Hadoop offers fine-grained authorization using file permissions in HDFS and resource level access control using ACLs for Map Reduce and coarser grained access control at a service level. HBase offers user

Authorization on tables, column families. The user authorization is implemented using coprocessors. Coprocessors are like database triggers in HBase .They intercept any request to the table before and after, now we can use the Project Rhino to extend HBase support for cell level ACLs. In Hive, authorization is implemented using Apache Sentry .Pig provides authorization using ACLs for job queues; Zookeeper also offers authorization using node ACLs. Hue provides access control via file system permission; it also offers ACLs for job queue.

3. Encryption:-

Encryption ensures confidentiality and privacy of user information, and it secures the sensitive data in Hadoop. Hadoop is a distributed system running on distinct machines, which means that data must be transmitted over the network on a regular basis, there is an increasing need of demand to move sensitive information into the Hadoop ecosystem to generate valuable perceptions. Sensitive data within the cluster needs special kind of protection and should be secured both at rest and in motion. This data needs to be protected during the transfer to and from the Hadoop system. The simple authentication and security layer (SASL) authentication framework is used for encrypting the data in motion in hadoop ecosystem. SASL security gives guarantee of the data being exchanged between client and servers and make sure that, the data is not readable by a "man-in-middle". SASL supports various authentication mechanisms, for example, DIGEST-MD5, CRAM-MD5, etc. The data at rest can be protected in two ways: First, when file is stored in Hadoop, the complete file can be encrypted first and then stored in Hadoop. In this approach, the data blocks in each Data Node can't be decrypted until we put all the blocks back and create the entire encrypted file. Second, to applying encryption to data blocks once they are loaded in Hadoop system.

4. CONCLUSION:-

In "Big Data" where large number of voluminous structured and unstructured data accumulated, stored and process security is a major concern. As hadoop is free, open source, Java-based programming framework that supports the processing of large data sets in a distributed computing environment therefore it become popular day by day it gaining larger acceptance within the industry, a natural

Concern over the security has spread. In this paper we have tried to cover all security approaches, issues and solutions.

REFERENCES:-

1. Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.

2. Bamford, J. (2013). Five myths about the National Security Agency. The Washington Post. http://articles.washingtonpost.com/2013-06-21/opinions/40114085_1_national-security-agency-foreign-intelligence-surveillancecourt- guardian. [Accessed June 25, 2013] Bamford, J. (2012).

3. Cloud Security Alliance “Top Ten big Data Security and Privacy Challenges”.

4. Kevin T. Smith “Big Data Security: The Evolution of Hadoop’s Security Model”.