International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Facing Security Problems in Big Data and Hadoop

P. S. Vijayalakshmi[1], T. Arunambika[2]

[1, 2]Assistant Professor, Department of Computer Science, Rathinam College of arts and science, Coimbatore-21

Abstract: ***One of the biggest concerns in our present age revolves around the security and protection of sensitive information. In our current era of Big Data, our organizations are collecting, analyzing, and making decisions based on analysis of massive amounts of data sets from various sources, and security in this process is becoming increasingly more important. Big data that resides within a Hadoop environment can contain sensitive financial data in the form of credit card and bank account numbers. It may contain proprietary corporate information and personally identifiable information (PII) such as the names, addresses and social security numbers of clients, customers and employees.***

Keywords: big data, Hadoop, authentication, security, Kerberos

1. Introduction

Smartphones with megapixel cameras, handheld computers, wireless sensor networks, ubiquitous social media, earth-orbiting satellites, space-bound telescopes—all generating more data than ever before: it is no exaggeration to say that over 90 percent of the world's data was produced in just the past two years[1].

NASA's Solar Dynamics Observatory uses four telescopes that gather eight images of the Sun every 12 seconds. In January 2015, the SDO captured its 100 millionth image of the Sun—just one example of the ways in which astronomers are collecting more and more data. Currently one petabyte of this data is publicly accessible online, and this volume grows at a rate of 0.5 petabytes per year. In CERN's Large Hadron Collider, 150 million sensors are capturing data about nearly 600 million collisions per second. On a similar scale, the work that won the 2013 Nobel Prize in chemistry involved measuring and visualizing the behavior of 50,000 or more atoms in a reaction over the course of a fraction of a millisecond[2].

In the social media domain, Facebook users add 300 million new photos a day; over 300 million Instagram users share 60 million photos every day; and more than 100 hours of video are uploaded to YouTube every minute[3].

In 2013, only 22 percent of data was considered useful, and less than 5 percent of that amount was actually analyzed. By 2020, more than 35 percent of all data could be considered useful due to increased production from sensors and IoT devices, and because it is increasingly engineered to meet specific goals, such as scientific discovery or process optimization. For example, IoT data from giant gas turbines that generate electricity has tremendous value since this can optimize power generation and assist with maintenance and repair. Likewise, the Square Kilometre Array (SKA) radio telescope project, expected to be operational in 2020, will produce 2.8 Gbytes of astronomy data per second that will help create the biggest map of the universe ever made[4].

- Volume—data measurement is in terabytes (240) or even petabytes (250), and is rapidly heading toward exabytes (260)
- Velocity—data production occurs at very high rates, and, because of this sheer volume, some applications require real-time data processing to determine whether to store a piece of data;
- Variety—data is heterogeneous and can be highly structured, semi-structured, or totally unstructured;
- Value—through predictive models that answer what-if queries, analysis of this data can yield counterintuitive insights and actionable intelligence.

For example, many natural language– and speech-related problems are ill suited for mathematically precise algorithmic solutions[5]. For example, for the part-of-speech tagging problem, training data consists of several sentences and part-of-speech annotations for each word in the sentences. now the accurate selection of a mathematical model loses its importance because there is big enough data to compensate[6].

2. Problems

For some problems, precise solutions are intractable, and may require faster and approximated algorithms that run the risk of decreasing the quality of the solution[7].

For many organizations Hadoop has evolved into an enterprise data platform. That poses new security challenges as data that was once siloed is brought together in a vast data lake and made accessible to a variety of users across the organization. Among these challenges are:

- Ensuring the proper authentication of users who access Hadoop.
- Ensuring that authorized Hadoop users can only access the data that they are entitled to access.
- Ensuring that data access histories for all users are recorded in accordance with compliance regulations and for other important purposes.
- Ensuring the protection of data—both at rest and in transit—through enterprise-grade encryption.

Of course, a bit of caution is imperative when dealing with a trendy topic.Such data has the potential to create confusion and misinformation rather than provide actionable insights. As described in “Big Data or Right Data?,” we need to ask the right kind of questions:

- How do we process, filter, and sample the source data to obtain the right data?

- How do we determine the trustworthiness of such data?
- How much noise is there?
- How do we distinguish between valid data and spam? Is the data distribution valid, or is there a hidden bias that needs to be corrected? How do we correct bias?
- How do we determine and eliminate duplicates?

Common security questions are:

- How do you enforce authentication for users and applications on all types of clients (e.g. web consoles and processes)?
- How do you make sure that rogue services aren't impersonating real services .How do you enforce access control to the data, based on existing access control policies and user credentials?
- How can Attribute-Based Access Control (ABAC) or Role-Based Access Control (RBAC) be implemented?
- How can Hadoop integrate with existing enterprise security services?
- How do you control who is authorized to access, modify and stop MapReduce jobs?
- How can you encrypt data in transit?
- How do you encrypt data at rest?
- How can you keep track of, and audit events & keep track of data provenance?
- What are the best network approaches for protecting my Hadoop cluster on the network?

One of the biggest concerns in our present age revolves around the security and protection of sensitive information. In our current era of Big Data, our organizations are collecting, analyzing, and making decisions based on analysis of massive amounts of data sets from various sources, and security in this process is becoming increasingly more important. Network security breaches from internal and external attackers are on the rise, often taking months to be detected, and those affected are paying the price.

3. A (Brief) History of Hadoop Security

Because impersonation was prevalent and done by most users, the security controls that did exist were not really effective.

Well-intended users can make mistakes (e.g. deleting massive amounts of data within seconds with a distributed delete). Anyone could submit a job to a JobTracker and it could be arbitrarily executed.

Today's Hadoop Security Challenges

There are number of security challenges for organizations securing Hadoop.

- **Authentication**. How can you ensure all users who access the Hadoop system are who they say they are and are allowed to access it?
- **Access control**. How can you ensure users who access Hadoop can only access the data they are entitled to access, with the same policies applied consistently however they access the Hadoop system.
- **Auditing**. How can you ensure that all users' data access histories are recorded for compliance and other purposes -- such as forensics, if the worst should happen.
- **Data protection**. Essentially this comes down to enterprise-grade encryption for data at rest and in motion.

Common security questions are:

How do you enforce authentication for users and applications on all types of clients (e.g. web consoles and processes)?

How do you make sure that rogue services aren't impersonating real services (e.g. rogue TaskTrackers and Tasks, unauthorized processes presenting block IDs to DataNodes to get access to data blocks, etc?) [8].

How do you enforce access control to the data, based on existing access control policies and user credentials?

How can Attribute-Based Access Control (ABAC) or Role-Based Access Control (RBAC) be implemented?

How can Hadoop integrate with existing enterprise security services?

How do you control who is authorized to access, modify and stop MapReduce jobs?

How can you encrypt data in transmit?

How do you encrypt data at rest?

How can you keep track of, and audit events & keep track of data provenance?

What are the best network approaches for protecting my Hadoop cluster on the network?

Hadoop security-complementing tools that we see in the industry.

1) No "Data at Rest" Encryption. Currently, data is not encrypted at rest on HDFS.
2) A Kerberos-Centric Approach – Hadoop security relies on Kerberos for authentication.
3) Limited Authorization Capabilities – Although Hadoop can be configured to perform authorization based on user and group permissions and Access Control Lists (ACLs), this may not be enough for every organization. Many organizations use flexible and dynamic access control policies based on XACML and Attribute-Based Access Control.
4) Complexity of the Security Model and Configuration. For network encryption, there are also three encryption mechanisms that must be configured – Quality of Protection for SASL mechanisms, and SSL for web consoles, HDFS Data Transfer Encryption. All of these settings need to be separately configured – and it is easy to make mistakes.

When it comes to auditing, Cloudera's solution is Cloudera Navigator. This records Hadoop activity details including:

- A timestamp ,the object that was accessed

- Details of the operation performed on an object,the user
- The ip address of that user
- The service instance through which the data was accessed

References

[1] V. Gudivada, D. Rao, and V. Raghavan, "Big Data–Driven Natural Language–Processing Research and Applications," Big Data Analytics, V. Govindaraju, V. Raghavan, and C.R. Rao, eds., Elsevier, 2015 (in press).
[2] V. Gudivada, D. Rao, and V. Raghavan, "Renaissance in Data Management Systems: SQL, NoSQL, and NewSQL," Computer (in press).
[3] R. Baeza-Yates. "Big Data or Right Data?" Proc. 7th Alberto Mendelzon Int'l Workshop on Foundations of Data Management (AMW 13), 2013, vol. 1087, paper 14; http://ceur-ws.org/Vol-1087/paper14.pdf.
[4] Ponemon Institute, "2013 Cost of Data Breach Study: Global Analysis", May 2013,
[5] Business Insider, "Playstation Network Crisis May Cost Sony Billions",
[6] For more information see "CNN/Money – 5 Data Breaches – From Embarrassing to Deadly", and Wikipedia's page on the AOL search data leak on anonymized records
[7] https://www.esecurityplanet.com/network-security/hadoop-security-still-evolving.html