WEBINAR SUMMARY

December 9, 2021

The Unbelievable Insecurity of the Big Data Stack: An Offensive Approach to Analyzing Huge and Complex Big Data Infrastructures

Sheila A. Berta, Head of Research, Dreamlab Technologies
Alex Heid, Chief Research Officer, SecurityScorecard

KEY TAKEAWAYS

- Understanding the security issues of a big data stack starts with understanding the stack.
- It is important to have a holistic view of big data architectures.
- A unique methodology to analyze the security risks involves looking at each individual layer.
- Recommendations to improve the security of big data stacks focus on reducing the attack surface.
- Enterprises have shifted from a fortress to an ecosystem model. SecurityScorecard provides an outside-in view of the cybersecurity posture of any organizations' digital footprint.

in partnership with

SecurityScorecard

OVERVIEW

The sheer magnitude of big data stacks makes them difficult to protect against cyberattacks. Big data infrastructures have many complex components involved in transporting, storing, processing, accessing, and managing data. There are multiple applications, services, and ports, resulting in a large attack surface. To date, there has not been a good methodology for assessing the security risks of these infrastructures. A new methodology involves analyzing the attack vectors for each layer of the stack.

Along with understanding the attack vectors for each layer, it is important to take steps to improve security and reduce risk. Steps include reducing the attack surface, setting up a firewall, securing credentials, implementing authentication, managing authorization, and having secure communications. In addition, SecurityScorecard provides an organization's "digital footprint," which gives an overview of risk factors and attack surfaces so organizations can take steps to reduce their risks.

CONTEXT

Sheila Berta summarized the layers of the big data stack, provided an example of big data architecture, discussed attack vectors for each layer, and offered recommendations to secure the big data stack. Alex Heid described the evolution to an ecosystem model, touched on third-party risks, and shared how SecurityScorecard helps organizations understand their risks through a "digital footprint."

KEY TAKEAWAYS

Understanding the security issues of a big data stack starts with understanding the stack.

To protect the big data infrastructure you must first understand and visualize them. There are four main layers of the stack, along with a cluster management layer.

Figure 1: Layers of the big data stack

The first layer is data ingestion, where data—often in large volume—is brought in and transported to a central storage location, which is the second layer. Because most data is raw, it must be processed, which is the third layer. The fourth layer is data access, which is how users access and consume data. Another layer that is not part of the stack, but which is present in all big data infrastructures, is the cluster management layer.

For each layer multiple technologies can be implemented. For example, Flume and Kafka are data ingestion technologies, while Hadoop HDFS and Hive are widely used data storage applications. For data processing, Hadoop YARN and Spark are frequently used, as is Presto for data access. Apache ZooKeeper is a common cluster management tool, along with Apache Ambari.

> When we analyze an entire big data infrastructure, we find many different and complex technologies interacting with each other, and they meet different functions according to the layer of the stack where they are located.
>
> *Sheila A. Berta, Dreamlab Technologies*

It is important to have a holistic view of big data architectures.

An example of a big data architecture is shown below.

Figure 2: Example of a big data architecture

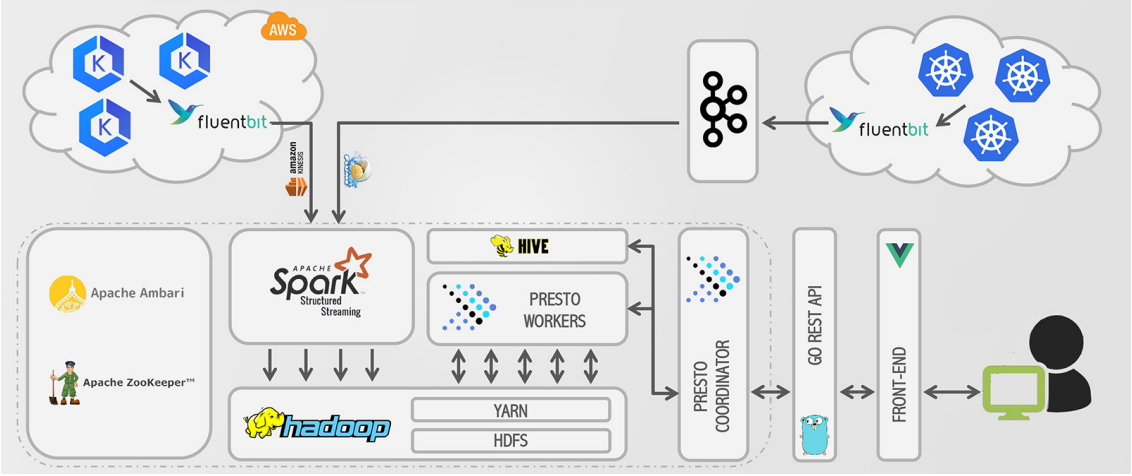


Figure 2 shows how data is ingested in two clouds and transported (using fluentbit) for storage (in Hive and other applications). Data is processed in Hadoop, with YARN and HDFS, and can be accessed with Presto, where additional software can be used to visualize the data. It is common to find Apache ZooKeeper as part of cluster management to centralize the configuration of these components, and an administration tool like Apache Ambari for cluster monitoring.

A unique methodology to analyze the security risks involves looking at each individual layer.

Currently, there is no accepted methodology for assessing the security of big data stacks or big data infrastructures, and there are few technical resources available to analyze various attack vectors. In addition, what are considered vulnerabilities in conventional infrastructures, or even in the cloud, are not necessarily vulnerabilities in the big data stack.

However, protecting all layers is necessary to keep applications and data safe. Understanding how data, layers, and applications interact helps organizations recognize potential vulnerabilities. This in turn helps organizations protect data and applications. The big question is, how can we analyze these complex infrastructures? Berta offered a new idea.

> I would like to propose a methodology where the analysis is based on the different layers of the big data stack.
>
> *Sheila A. Berta, Dreamlab Technologies*

Berta's methodology involves dissecting and analyzing each layer. She conducted research by attacking different technologies involved in each layer to identify vulnerabilities.

1. Management layer. Exploiting the cluster management layer is a quick way for hackers to steal data and compromise components. Berta exposed potential vulnerabilities in two management platforms, ZooKeeper and Ambari, that are commonly used to configure components in the cluster. She was able to remotely attack the centralized cluster configuration managed by ZooKeeper.

   Zookeeper:

   – Zookeeper runs a service on all nodes, which allows cluster administrators to connect to nodes and update configurations.

   – The official Zookeeper client allows potential hackers to run specific commands that connect to specific nodes and execute commands.

   – Utilizing LS and Get commands, users can browse the hierarchical structure and find information about the configuration of the components that make up the cluster.

   – This information can be used for attacks and to create new configurations, modify existing ones, and delete configurations which can affect component performance.

   Ambari

   – Ambari has a dashboard where users control everything, as well as a database.

   – In the installation process, users are permitted to change the credentials for the dashboard, but not the database.

   – Hackers can use default credentials to access the database and get username and authorization keys by updating the key for the admin user.

   – Bad actors are then able to log into the dashboard with admin credentials.

The main takeaway of this research is to analyze the security of administration and monitoring tools.

2. Storage layer. Hadoop, an open-source software framework for storing data and running applications on clusters of commodity hardware, was the next target. Berta was able to craft packets to enable remote communication with the Hadoop RPC/IPC to compromise the Hadoop Distributed File System.

   Hadoop

   – Hadoop's Distributed File System (HDFS) has two main components: one node that saves metadata of the file stored in the cluster and another that stores the clusters' actual data.

   – If hackers can connect to the master nodes via open ports, users can execute Hadoop commands and access the stored data.

   – This can be accomplished by manually crafting four configuration files that Hadoop uses to perform operations over the Hadoop file system. The information for the configuration files can be found by accessing visible dashboard data and potentially data from ZooKeeper, if hackers cannot access the dashboard data.

- Users can then impersonate HDFS users, patch the configuration files to Hadoop, and execute commands remotely on the target cluster.
- Hackers can then remove, delete, replace, move, and copy any files and directories.

> All these dashboards are exposed by default and don't require any authentication . . . we have seen that it's easy to get all this information in a remote way.
>
> *Sheila A. Berta, Dreamlab Technologies*

3. **Processing layer.** Berta developed a malicious YARN application to achieve remote code execution. The attack was performed utilizing a similar methodology as the storage layer which gave users access to the application master file and client file. Users are able to change the application using various commands. Complicated commands, such as a reverse shell on the cluster nodes, are possible utilizing this methodology. Spark, a popular technology for processing data, is often installed on top of Hadoop. Berta was able to easily break into this technology.
4. **Ingestion layer.** Attackers will look to interfere with ingestion channels. Sending information from one database to another or a database to a data lake opens up an opportunity for hackers to exploit communication channels. Unsecure channels are welcome invitations to hackers to interfere with data transmission and infect databases with malicious data. Berta was able to interfere with data ingestion channels. This was apparent in the data transfer tool, Sqoop.

   Organizations need to check that the interfaces that are waiting for the data cannot be reached by an attacker
5. **Data access layer.** Data access technologies are often hierarchical data format-based storage technologies, which provide interfaces to access information. This is a potential issue as hackers could access the data and compromise components through dashboards and interfaces. Berta was able to abuse the drivers of HDFS-based storage technologies, such as Hive/HBase.

Recommendations to improve the security of big data stacks focus on reducing the attack surface.

Berta offered the following recommendations.

Recommendation	Description
1. Reduce the attack surface	Many attacks are based on exposed interfaces. To improve security, reduce the attack surface by removing dashboards and interfaces that are not used, or block access to them. Make sure that dashboards or even the IPC ports are not exposed to the internet or to an insecure network.
2. Set up a firewall	Block necessary ports and secure the perimeter.
3. Secure credentials	Change all default credentials in the technologies implemented.
4. Implement authentication	Most technologies support advanced authentication mechanisms.
5. Manage authorization	Apply the principle of least privilege.
6. Secure communications	Secure the communication channels between the different technologies.

> Remember that in a big data infrastructure there are many, many different technologies communicating with each other. Make sure those communications are happening in a secure way.
>
> *Sheila A. Berta, Dreamlab Technologies*

Enterprises have shifted from a fortress to an ecosystem model. SecurityScorecard provides an outside-in view of the cybersecurity posture of any organizations' digital footprint.

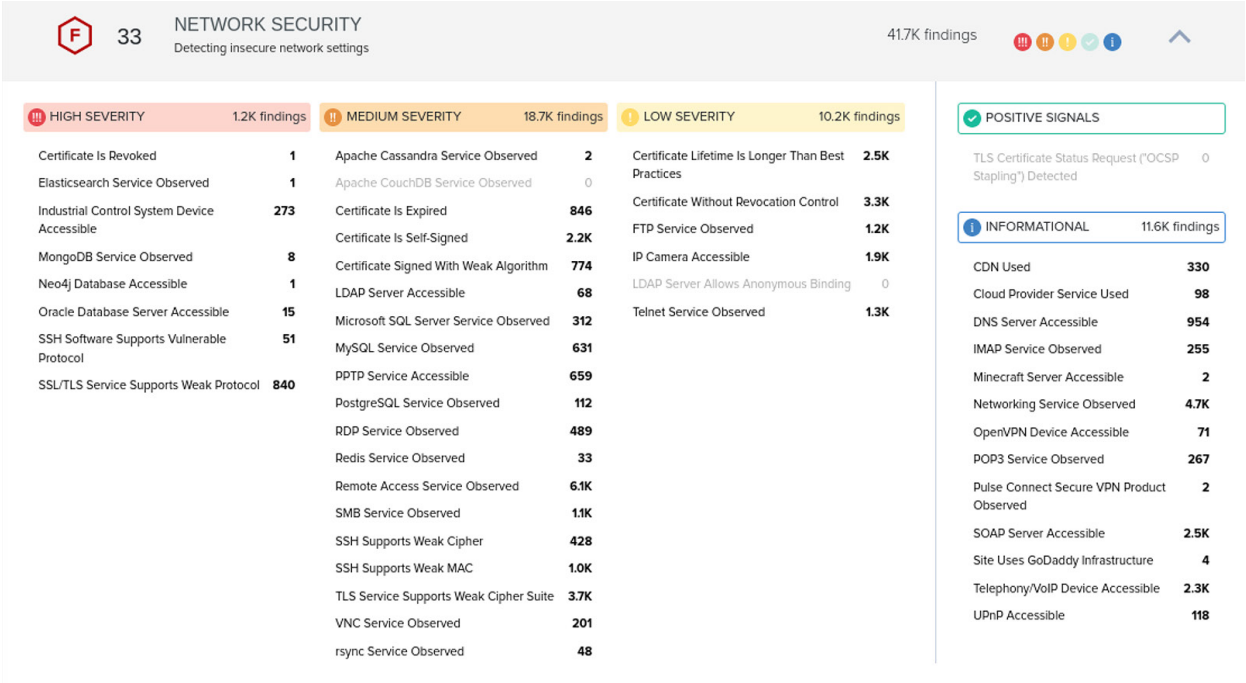
In the fortress model of the past, a security team knew what was within their firewalls and what data was moving in and out. But in today's big data era, where organizations use numerous third-party vendors and cloud services, the prevailing model is an ecosystem model. As shown in Figure 3, at the center of the ecosystem model is You.

Figure 3: Ecosystem Model

In this ecosystem model, the first ring is a company's immediate suppliers (third parties), followed by the suppliers to those suppliers (fourth parties), and so on (nth party).

SecurityScorecard's Platform continuously and non-intrusively collects data across the internet, maps an organization's digital footprint, and enables organizations to understand and visualize their security risks. It categorizes every digital asset made available or discoverable by an enterprise. In one example, an enterprise viewing its digital footprint could see 215,000 company IP addresses, which included multiple services, applications, and ports. SecurityScorecard can provide an enterprise a broad view of its ecosystem and an overview of its risk factors. SecurityScorecard's easy-to-understand A to F grades enable any organization to easily understand its own or its suppliers' security posture.

Figure 4: Overview of Risk Factors: Network Security, from SecurityScorecard



> The way to prevent [security issues] is to understand the technologies you're deploying, disable the stuff you're not using, block stuff from the public internet that shouldn't be connected, and because you can't protect what you can't see, have a continuous monitoring service or solution . . . that will go a long way to knowing your attack surface.
>
> *Alex Heid, SecurityScorecard*

ADDITIONAL INFORMATION

Every organization has access to its own Scorecard. Sign up for a free SecurityScorecard account and see your security posture at https://securityscorecard.com/free-account.

BIOGRAPHIES

Sheila A. Berta

Head of Research, Dreamlab Technologies

Sheila A. Berta is an offensive security specialist who started at 12 years old by learning on her own. At the age of 15, she wrote her first book about web hacking, published in several countries. Over the years, Berta has discovered vulnerabilities in popular web applications and software, as well as given courses at universities and private institutes in Argentina. She specializes in offensive techniques, reverse engineering, and exploit writing and is also a developer in ASM (MCU and MPU x86/x64), C/C++, Python, and Go. In recent years she focused on cloud native and big data security. As an international speaker, she has spoken at important security conferences such as Black Hat Briefings, DEF CON, HITB, Ekoparty, IEEE ArgenCon, and others. Berta currently works as head of research at Dreamlab Technologies.

Alex Heid

Chief Research Officer, SecurityScorecard

Alexander Heid serves as chief research & development officer at SecurityScorecard. Heid joined the company in June 2014 and has been instrumental in developing the company's threat reconnaissance capabilities and building its security-centric platform. A recognized expert in the field, he frequently presents at industry conferences and is sought out by the media and analysts to discuss cybersecurity issues. Prior to joining the company, Heid held senior security roles within the financial industry, and was a senior analyst at Prolexic Technologies during the #OpAbabil DDoS campaigns. In addition, he is cofounder and president/CEO of HackMiami and served as chapter chair for South Florida OWASP.