Voltage SecureData for Hadoop and IoT

Protecting Sensitive Data in and beyond the Data Lake

The Need to Secure Sensitive Data in Hadoop Ecosystems

Hadoop is a unique architecture designed to enable organizations to gain new analytic insights and operational efficiencies through the use of multiple standard, low-cost, high-speed, parallel processing nodes operating on very large sets of data. The resulting flexibility, performance, and scalability are unprecedented. But data security was not the primary design goal.

When used in an enterprise environment, the importance of security becomes paramount. Organizations must protect sensitive customer, partner and internal information and adhere to an ever-increasing set of compliance requirements. But by its nature, Hadoop poses many unique challenges to properly securing this environment, not least of which include automatic and complex replication of data across multiple nodes once entered into the HDFS data store. There are a number of traditional IT security controls that should be put in place as the basis for securing Hadoop, such as standard perimeter protection of the computing environment, and monitoring user and network activity with log management. But infrastructure protection by itself cannot prevent an organization from cyber-attacks and data breaches in even the most tightly controlled computing environments.

Hadoop is a much more vulnerable target—too open to be able to fully protect. Further exacerbating the risk is that the aggregation of data in Hadoop makes for an even more alluring target for hackers and data thieves. Hadoop presents brand new challenges to data risk management: the potential concentration of vast amounts of sensitive corporate and personal data in a low-trust environment. New methods of data protection at Petabyte scale are thus essential to mitigate these potentially huge Big Data exposures.

Data Protection Methodologies

There are several traditional data de-identification approaches that can be deployed to improve security in the Hadoop environment, such as storage level encryption, traditional field-level encryption and data masking. However, each of these approaches has limitations.

For example, with storage-level encryption the entire volume that the data set is stored in is encrypted at the disk volume level while “at rest” on the data store, which protects against unauthorized personnel who may have physically obtained the disk, from being able to read anything from it. This is a useful control in a Hadoop cluster or any large data store due to frequent disk repairs and swap-outs, but does nothing to

Benefits

- Voltage SecureData data-centric security platform provides unique benefits.
- The ability to protect data as close to its source as possible.
- Support for encryption, tokenization and data masking protection techniques.
- Supports the encryption and pseudonymization guidance in the new GDPR (General Data Protection Regulation) data privacy legislation for European Union resident data
- Data usable for many applications in its de-identified state.
- The ability to securely re-identify select data fields for live data access—only by authorized users and applications when required for business needs
- Protection techniques backed by security proofs and standards.
- The industry’s first Federal Information Processing Standard (FIPS) 140-2 validation of FPE, and the world’s first FIPS-validated AES-FF1 encryption configuration option to operate in strict FIPS mode.
- High performance, high scalability well matched with Hadoop speeds.
- Broad platform and application support—inside and outside Hadoop.
- Integrated with Apache NiFi processor ecosystem to enable data security in wide scale data flows such as the Internet of Things (IoT).

protect the data from any and all access when the disk is running—which is all the time.

Data masking is a useful technique for obfuscating sensitive data, most often used for creation of test and development data from live production information. However, masked data is intended to be irreversible, which limits its value for many analytic applications and post-processing requirements. Moreover, there is no guarantee that the specific masking transformation chosen for a specific sensitive data field fully obfuscates it from identification, particularly when correlated with other data in the Hadoop "data lake."

While all of these technologies potentially have a place in helping to secure data in Hadoop, none of them truly solves the problem nor meets the requirements of an end-to-end, data-centric solution.

Data-Centric Security

The obvious answer for true Hadoop security is to augment infrastructure controls with protecting the data itself. This data-centric security approach calls for de-identifying the data as close to its source as possible, transforming the sensitive data elements with usable, yet de-identified, equivalents that retain their format, behavior, and meaning. This protected form of the data can then be used in subsequent applications, analytic engines, data transfers and data stores, while being readily and securely re-identified for those specific applications and users that require it. For Hadoop, the best practice is to never allow sensitive information to reach the HDFS in its live and vulnerable form. De-identified data in Hadoop is protected data, and even in the event of a data breach, yields nothing of value, avoiding the penalties and costs such an event would otherwise have triggered.

The Solution—Voltage SecureData for Hadoop and IoT

SecureData for Hadoop and IoT provides maximum data protection with industry-standard, next generation Format-Preserving Encryption (FPE), (see NIST SP-800-38G) and Voltage Secure Stateless Tokenization (SST) technologies.

With Voltage FPE and SST, protection is applied at the data field and sub-field level, preserves characteristics of the original data, including numbers, symbols, letters and numeric relationships such as date and salary ranges, and maintains referential integrity across distributed data sets so joined data tables continue to operate properly. Voltage FPE and SST provide high-strength encryption and tokenization of data without altering the original data format.

Micro Focus® Voltage SecureData encryption/tokenization protection can be applied at the source before it gets into Hadoop, or can be evoked during an ETL transfer to a landing zone, or from the Hadoop process transferring the data into HDFS. Once the secure data is in Hadoop, it can be used in its de-identified state for additional processing and analysis without further interaction with SecureData. Or the analytic programs running in Hadoop can access the clear text by utilizing the SecureData high-speed decryption/de-tokenization interfaces with the appropriate level of authentication and authorization.

If processed data needs to be exported to downstream analytics in the clear—such as into a data warehouse for traditional BI analysis—there are multiple options for re-identifying the data, either as it exits Hadoop using Hadoop tools or as it enters the downstream systems on those platforms.

To implement data-centric security requires installing the Voltage SecureData infrastructure components and then interfacing with the appropriate applications and data flows. SDKs, APIs and command line tools enable encryption and tokenization to occur natively on the widest variety of platforms, including Linux, mainframe and mid-range, and supports integration with a broad range of infrastructure components, including ETL, databases, and programs running in the Hadoop environment, and is available for any Hadoop distribution. Voltage has technology partnerships with Hortonworks, MapR, Cloudera and IBM, and certifications to run on each of these. Voltage SecureData is integrated with the Teradata Unified Data Architecture (UDA), and with the Vertica Big Data Platform.

Rapid Evolution Requires Future-Proof Investments

Implementing data security can be a daunting process, especially in the rapidly evolving and constantly changing Hadoop space. It's essential for long-term success and future-proofing investments, to apply technology via a framework that can adapt to the rapid changes ongoing in Hadoop environments. Unfortunately, implementations based on agents frequently face issues when new releases or new technology are introduced into the stack, and require updating the Hadoop instance multiple times. In contrast, Voltage SecureData for Hadoop and IoT provides a framework that enables rapid integration into the newest technologies needed by the business. This capability enables rapid expansion and broad utilization for secure analytics.

Securing the Internet of Things

Failure to protect sensitive data in the Hadoop environment holds major risk of data breach, leaking sensitive data to adversaries, and non-compliance with increasingly stringent data privacy regulations such as the General Data Protection Regulation (GDPR). Big Data use cases such as real-time analytics, centralized data acquisition and staging for other systems require that enterprises create a “data lake”—or a single location for the data assets.

While IoT and big data analytics are driving new ways for organizations to improve efficiencies, identify new revenue streams, and innovate, they are also creating new attack vectors which make easy targets for attackers. This is where perimeter security is critical, but also increasingly insufficient—it takes, on average, over 200 days before a data breach is detected and fixed.

As the number of IoT connected devices and sensors in the Enterprise multiplies, the amount of sensitive data and Personally Identifiable Information collected at the IoT Edge and moving into the back-end in the data center--is growing exponentially.

The data generated from IoT is a valued commodity for adversaries, as it can contain sensitive information such as Personally Identifiable Information (PII), payment card information (PCI) or protected health information (PHI). For example, a breach of a connected blood pressure monitor’s readings alone may have no value to an attacker, but when paired with a patient’s name, it could become identity theft and a violation of (HIPAA) regulations.

IoT is here to stay. A recent Forbes article predicted that we will see 50 billion interconnected devices within the next 5-10 years. Because a multitude of companies will be deploying and using IoT technologies to a great extent in the near future, security professionals will need to get ahead of the challenge of protecting massive amounts of IoT data. And, with this deluge of sensitive IoT data, Enterprises will need to act quickly to adopt new security methodologies and best practices in order to enable their Big Data projects and IoT initiatives.

New Threats Call for New Solutions—NiFi Integration

A new approach is required, focused on protecting the IoT data as close to the source as possible. As with other data sources, sensitive streaming information from connected devices and sensors can be protected with Voltage FPE to secure sensitive data from both insider risk and external attack, while the values in the data maintain usability for analytics.

However, Apache NiFi, a recent technology innovation, is enabling IoT to deliver on its potential for a more connected world. Apache NiFi is an open source platform that enables security and risk architects, as well as business users, to graphically design and easily manage data flows in their IoT or back-end environments.

Application: Supply Chain Data and Analytics Solutions Provider

- A technology company that provides real-time supply chain data and analytics for retailers, manufacturers and trading partners, is using Hortonworks HDP Open Source Enterprise Apache Hadoop Platform, with Voltage SecureData for Hadoop to de-identify sensitive data at the field level.
- The company delivers pharmacy claims reconciliation for top retailers, grocery and pharmacy chain stores, and is responsible for Personally Identifiable Information (PII), Protected Health Information (PHI) subject to HIPAA/HITECH regulations, and data ingested from thousands of hospitals and healthcare facilities, such as insurance identification, date information, procedure codes, etc.
- Their data science team performs analytics on the de-identified claims data inside the Hadoop environment using MapReduce and Hive, to produce usage trending, market basket insights, and identification of new products and services. Additionally, when specific health risks or need for a procedure or medication are identified through their data analysis, the individual can be quickly and securely re-identified, and that information provided back out to the healthcare provider.

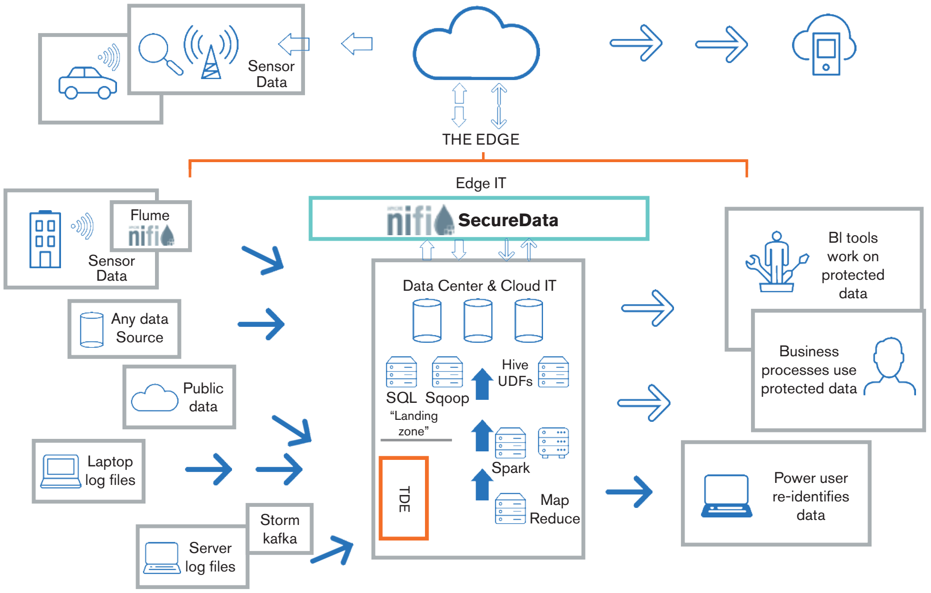

Figure 1. Threats in the IoT Space—Pushing Protection to the Edge

When asked about the role of data security in IoT initiatives, 78% of buyer-side respondents indicate lack of data security will impact their IoT progress.[1]

www.microfocus.com

SecureData for Hadoop and IoT is designed to easily secure sensitive information that is generated and transmitted across Internet of Things (IoT) environments, with Format-preserving Encryption (FPE). The solution features the industry's first-to-market Apache NiFi integration with NIST standardized and FIPS compliant format-preserving encryption technology to protect IoT data at rest, in transit and in use.

The Voltage SecureData NiFi integration enables organizations to incorporate data security into their IoT strategies by allowing them to more easily manage sensitive data flows and insert encryption closer to the intelligent edge. This capability is included in the SecureData for Hadoop and IoT product. In addition, it is certified for interoperability with Hortonworks DataFlow (HDF).

With this industry first, the SecureData for Hadoop and IoT solution now extends data-centric protection, enabling organizations to encrypt data closer to the intelligent edge before it moves into the back-end Hadoop Big Data environment, while maintaining the original format for processing and enabling secure Big Data analytics.

Packages

The Voltage SecureData for Hadoop and IoT is available in two pre-configured packages. Use the Starter Edition to get started, protecting sensitive data for pilot projects and small deployments, which includes licensing for up to 5 Hadoop nodes. Use the Enterprise Edition with full, production level Voltage SecureData infrastructure and licensing for up to 20 Hadoop nodes. Each package includes an unlimited number of applications running directly on Hadoop or used by an ETL or batch process transferring directly into or out of Hadoop. Protection for additional Hadoop nodes can be added to these packages to meet the exact data protection needs for any Enterprise Hadoop and multi-platform environments.

Voltage SecureData for Hadoop and IoT Starter Edition	Voltage SecureData for Hadoop and IoT Enterprise Edition
▪ 1 Key server and Web services server for production	▪ Dual key servers and Web services servers for production
▪ Installation kit for Linux platform	▪ Installation kits for Linux & Windows platforms
▪ Usage license for up to 5 Hadoop nodes	▪ Usage license for up to 20 Hadoop nodes
▪ Developer templates for Hive, MapReduce and Sqoop	▪ Developer templates for Hive, MapReduce and Sqoop
▪ SecureStorage Volume Encryption	▪ SecureStorage Volume Encryption
▪ One year premium support	▪ One year premium support
▪ Voltage SecureData Installation, Configuration and Setup	▪ Voltage SecureData Installation, Configuration and Integration Assistance

Learn More At
microfocus.com/sdhadoop

☎

Micro Focus
UK Headquarters
United Kingdom
+44 (0) 1635 565200

U.S. Headquarters
Rockville, Maryland
301 838 5000
877 772 4450

Additional contact information and office locations:
www.microfocus.com

[1] *Database Trend & Applications (DBTA) Internet of Things Market Survey, January 2017*