The Need to Secure Sensitive Data in Hadoop

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 zeta-byte 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

Benefits

Voltage’s 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.
- 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.
- High performance, high scalability well-matched with Hadoop speeds.
- Broad platform and application support – inside and outside Hadoop.
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 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.

Voltage Security provides such a solution.

The Solution – Voltage SecureData™ Suite for Hadoop

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

With Voltage SecureData 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.

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

“CISOs should not treat big data security in isolation, but require policies that encompass all data silos to avoid security chaos. New data-centric audit and protection solutions and management approaches are required.”

– Gartner,
“Big Data Needs a Data-Centric Security Focus”,
Brian Lowans, Earl Perkins,
26 March 2014

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 2.1 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 health care 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.
processing and analysis without further interaction with the Voltage system. Or the analytic programs running in Hadoop can access the clear text by utilizing the Voltage 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 Voltage 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.

Voltage Security has technology partnerships with Hortonworks, MapR, Cloudera and IBM, and the Voltage SecureData Suite for Hadoop is certified to run on each of these.

How it Works

Seven specific options Voltage provides for protecting sensitive data used in Hadoop are shown below.

- **Option 1: Apply data protection at source applications** – This is the ideal scenario, so that the sensitive data is fully protected wherever it flows, including Hadoop. It ensures that the Hadoop system is not brought into scope for PCI and other compliance policies.

- **Option 2: Apply data protection during import into landing zone (ETL process)** – This is a good intermediate option that does not require interfacing at the source applications while still preserving the Hadoop system’s status as outside the scope of compliance requirements.

- **Option 3: Apply data protection during Hadoop import processing (e.g. Sqoop, MapReduce)** – This option utilizes the Voltage interfaces running directly within Hadoop jobs, providing potentially the highest-speed data import and the opportunity to integrate with other Hadoop pre-processing tasks. This is a good approach for handling streaming data into Hadoop. Data already residing
in HDFS can also be de-identified and re-written with this same MapReduce interface.

- **Option 4: Using de-identified data within Hadoop** – Once the data has been imported into Hadoop, the ideal scenario is performing all analytics and processing on the protected data, which avoids the need for decryption or de-tokenization.

- **Option 5: Using and exporting re-identified data from Hadoop (Hive, MapReduce)** – There will be situations in which some data fields need to be processed in their clear, unaltered form, and this is possible using the same interfaces as discussed in Option #3 above. Voltage’s stateless key architecture enables the distributed, high-speed re-identification of the data so as not to slow down the Hadoop processing.

- **Option 6: Exporting data and re-identifying outside Hadoop (ETL process)** – Hadoop data can also be bulk transferred and re-identified during ETL processes for downstream processing outside of Hadoop. A common use case for this is when the Hadoop analysis is used as pre-processing for input to traditional BI applications.

- **Option 7: Using storage-level encryption within Hadoop** – Voltage also offers the option of encryption at the volume level, which is the exact equivalent of most of the “Hadoop encryption” options available in the industry today. It is unnecessary protection for any fields protected by the “data-centric security” options above, but is very useful as an extra level of defense for the entire data set stored in Hadoop, particularly unstructured data or sensitive data you might not yet be aware of. Voltage’s storage-level encryption uses the same key management infrastructure used for Voltage FPE and SST, so one key management solution protects physical disks in case of loss as well as sensitive data fields.

## Packages

The Voltage SecureData Suite for Hadoop 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 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.

<table>
<thead>
<tr>
<th>Voltage SecureData Suite for Hadoop Starter Edition</th>
<th>Voltage SecureData Suite for Hadoop Enterprise Edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 1 Key Server &amp; Web Services Server for Production</td>
<td>- Dual Key Servers &amp; Web Services Servers for Production</td>
</tr>
<tr>
<td>- Installation Kit for Linux platform</td>
<td>- Dual Key Servers &amp; Web Services Servers for Test &amp; Dev</td>
</tr>
<tr>
<td>- Usage License for up to 5 Hadoop nodes</td>
<td>- Installation Kits for Linux &amp; Windows platforms</td>
</tr>
<tr>
<td>- Developer Templates for Hive, MapReduce &amp; Sqoop</td>
<td>- Usage License for up to 20 Hadoop nodes</td>
</tr>
<tr>
<td>- SecureStorage Volume Encryption</td>
<td>- Developer Templates for Hive, MapReduce &amp; Sqoop</td>
</tr>
<tr>
<td>- One year premium support</td>
<td>- SecureStorage Volume Encryption</td>
</tr>
<tr>
<td>- Voltage Installation, Configuration and Set-up</td>
<td>- One year premium support</td>
</tr>
</tbody>
</table>

## ABOUT VOLTAGE SECURITY

Voltage Security®, Inc. is the world leader in data-centric encryption and tokenization. Voltage provides trusted data security that scales to deliver cost-effective PCI compliance, scope reduction and secure analytics. Voltage solutions are used by leading enterprises worldwide, reducing risk and protecting brand while enabling business.

For more information, please visit [www.voltage.com](http://www.voltage.com).