opentext™ | Cybersecurity

Solution Flyer

Voltage SecureData Enterprise for Hadoop

Protect sensitive data in and beyond the data lake

Voltage SecureData Enterprise for Hadoop at a Glance

- High performance and scalability match Hadoop cluster sizes and speeds.
- Broad platform and application support inside and outside of Hadoop across Cloudera, and legacy Hortonworks and MapR distributions.
- Support for Hadoop ecosystem technologies includes MapReduce, Sqoop, Hive, Spark, Kafka, Storm, NiFi, TDE.
- Protects data close to source, retaining usability for applications and analytics, with selective re-identification by authorized actors.
- Encryption, tokenization, hashing, and data masking protection techniques backed by security proofs and standards.
- Secure stateless technologies remove overhead of storage and management of keys and token tables
- Privacy regulation anonymization and pseudonymization guidance supported

The Need to Secure Sensitive Data in the Hadoop Ecosystem

Apache Hadoop is an open source platform that provides a software framework for highly reliable, scalable, distributed storage and processing of large data sets. Operational efficiencies gained through the use of clusters of low-cost, high-speed, commodity computers enable organizations to ingest and analyze massive amounts of structured, semi-structured, and unstructured data.

In enterprise environments, data security is paramount. Failure to protect sensitive data incurs a major risk of data breach, leaking sensitive data to adversaries, and non-compliance with increasingly stringent data privacy laws, such as the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and the Health Insurance Portability and Accountability Act (HIPAA). Big Data use cases such as real-time analytics, centralized data acquisition, and staging for downstream systems require that enterprises create a “data lake”—a single location for enterprise data assets.

Hadoop poses unique challenges in securing its data lakes, however, which include accounting for the automated, complex replication of data across multiple storage nodes following ingestion into the Hadoop Distributed File System (HDFS). Of course, infrastructure controls should be used, including protecting the perimeter of the computing environment, and monitoring the activities of users and networks. But, as has been demonstrated time and time again, traditional IT security alone cannot protect an organization from cyber-attacks or prevent data exfiltration in even the most tightly controlled environments.

Hadoop is vulnerable. Its architecture is open and its aggregation of sensitive corporate and personal data in a low-trust environment makes it a prime target for hackers and data thieves. Clearly, techniques for protecting data at petabyte scales are essential if big data breaches are to be mitigated or eliminated. But, equally so, such techniques must not prevent or otherwise render infeasible the analytic processing for which the Hadoop data lake was created.

Traditional Data Protection Is Insufficient

The obvious answer to the Hadoop data security question is to augment infrastructure controls with protection of the data itself. But while traditional data protection methods, such as storage level encryption and data masking, can be deployed to improve security in the Hadoop environment, these approaches are limited when considered in relation to big data analytics.

Storage-level encryption protects data at rest at the disk volume level. This technology prevents attackers who have simply obtained physical access to the disk from being able to read it. While this may be a useful control for Hadoop clusters or large data stores where there are frequent disk repairs and swap-outs, it does not protect the data from anyone who has obtained legitimate access credentials. Such attackers can freely extract all the data on the disk in its unprotected form.

Data masking is a useful technique for obfuscating sensitive data, most often used to create functional substitutes of live production data for test, development, and user training. However, masking breaks relationships in the data and thus also the

ability to glean insights from such relationships. Masked data is also irreversible, destroying its value for analytic and post-processing scenarios in which access to the unprotected data is required. Moreover, masking transforms may fail to fully anonymize certain data against re-identification, particularly when correlations against other data in the Hadoop data lake are possible.

Voltage SecureData Enterprise for Hadoop Is the Solution

Voltage SecureData Enterprise by OpenText™ for Hadoop provides a set of advanced security solutions, including Voltage SecureData Format-Preserving Encryption* (FPE), Format-Preserving Hash (FPH), Secure Stateless Tokenization (SST), and Stateless Key Management, that enable the pseudonymization or anonymization of sensitive data at field and sub-field levels while preserving their format, behavior, and meaning. Characteristics of the original data, including character types, alphabets, and numeric relationships, such as date and salary ranges, can be maintained along with their referential integrity across distributed data sets, while avoiding the requirements to manage the storage of encryption keys or token tables that traditional solutions incur.

Due to the lack of inherent security controls in Hadoop, a best practice is to never allow sensitive data to reach HDFS in its clear, unprotected form. Voltage SecureData Enterprise protection can be applied at the source before it is imported into Hadoop, evoked from an extract-transform-load (ETL) process as it is transferred to a Hadoop landing zone, or from a Hadoop process as it is written to HDFS. Such de-identified forms of the data can be used in applications, analytic engines, data transfers, and data stores as they are, without further modification, and yet a Hadoop breach that exposes such data yields nothing of value to the attackers, avoiding the penalties and costs such an event would otherwise trigger.

Big Data analytics that need to re-identify pseudonymized data can still be authorized to do so, of course, by authenticating to Voltage SecureData Enterprise's high-speed interfaces. And if processed data needs to be exported for downstream storage and analytics—such as into an enterprise data warehouse for traditional business intelligence (BI) analysis—there are multiple options for re-identifying the data as it exits the data lake or is imported by downstream platforms, such as OpenText™ Vertica™ and Teradata, both of which are already integrated with Voltage SecureData Enterprise.

Rapid Technology Evolution Requires Flexible Solutions

It's essential for the long-term security of Big Data investments to apply solutions that can adapt to the rapid evolutions occurring in the Hadoop technology space. In contrast to agent-based implementations that create management and operational issues when new or updated technologies are introduced, Voltage SecureData Enterprise for Hadoop provides a framework that enables rapid integration with the latest tools and broad utilization for secure analytics.

Software Development Kits (SDKs), Application Programming Interfaces (APIs), User-Defined Functions/Extensions (UDFs/UDxs), integration code samples, and command line tools enable Voltage SecureData security to occur natively on a wide variety of platforms, and support integration with a broad range of infrastructure components, including ETL tools, databases, and programs running throughout the Hadoop environment. Supported technologies include MapReduce, Sqoop, Hive, Spark, Kafka, Storm, NiFi, and TDE. Supported distributions include Cloudera CDH, Hortonworks HDP, and MapR.

Securing the Internet of Things

As the number of Internet connected devices in the Enterprise continues to multiply, the volume of data generated and transferred into Big Data systems like Hadoop is growing exponentially *(see Figure 1)*. Data generated from this Internet of Things (IoT) is a valued

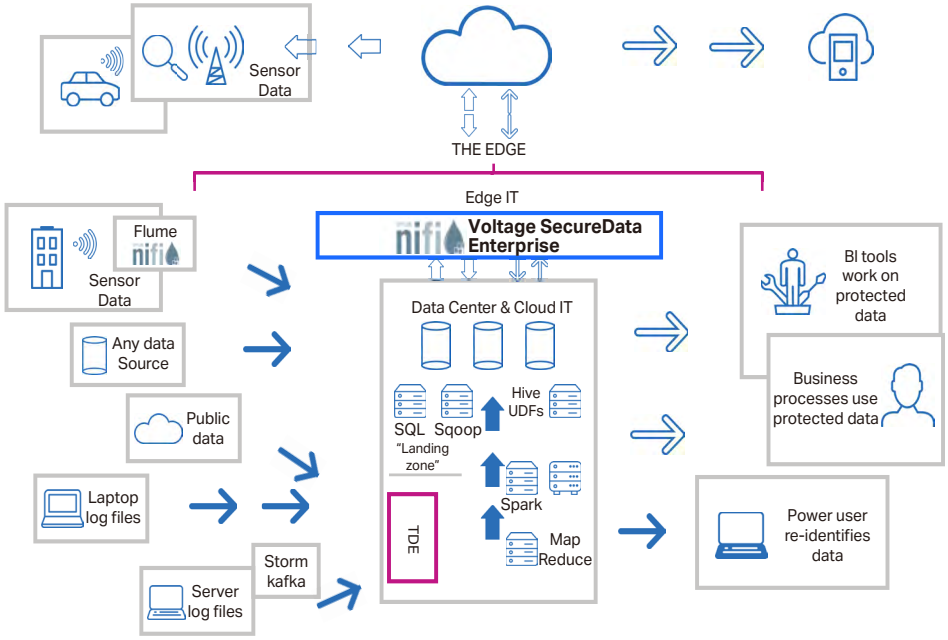

Figure 1. Threats in the IoT Space—Pushing Protection to the Edge

*NIST SP-800-38G

commodity for adversaries, as it may contain an organization's Intellectual Property (IP) and sensitive information, such as personally identifiable information (PII), payment card information (PCI), or protected health information (PHI).

Our existing suite of advanced security solutions provided through Voltage SecureData Enterprise for Hadoop, including FPE, FPH, and SST, easily secures sensitive information generated and transmitted across large-scale IoT deployments, such as are implemented by mobile operators and auto manufacturers. And our industry-first integration with Apache NiFi—an integrated data logistics platform that enables the graphical design and management of data flows—further supports the incorporation of data-centric security into IoT and back-end environments closer to the intelligent edge.

Packages

Voltage SecureData Enterprise for Hadoop is available in Starter and Enterprise Editions. Starter Edition includes licensing for up to 5 Hadoop nodes and is intended for pilot projects and small deployments. Enterprise Edition includes full, production-level infrastructure and licensing for up to 20 Hadoop nodes. Each package includes licensing for 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 your exact data protection needs.

Voltage SecureData Enterprise for Hadoop Starter Edition

- 1 Key Server and Web Services Server for production
- Installation kit for Linux platform
- Usage license for up to 5 Hadoop nodes
- Integration templates for MapReduce, Hive, Sqoop, Spark, NiFi, Kafka, Storm
- One-year premium support
- Voltage SecureData Installation, Configuration and Setup

Voltage SecureData Enterprise for Hadoop Enterprise Edition

- 2 Key Servers and Web Services Servers for production
- Installation kits for Linux & Windows platforms
- Usage license for up to 20 Hadoop nodes
- Integration templates for MapReduce, Hive, Sqoop, Spark, NiFi, Kafka, Storm
- One-year premium support
- Voltage SecureData Enterprise Installation, Configuration, Setup, and Integration Assistance

Connect with Us

www.opentext.com

A technology company that provides real-time supply chain data and analytics for retailers, manufacturers, and trading partners, is using Voltage SecureData for Hadoop to de-identify data ingested from thousands of hospitals and healthcare facilities. The company's delivery of pharmacy claims reconciliation for grocery and pharmacy chain stores subject it to both HIPAA and HITECH (Health Information Technology for Economic and Clinical Health) regulations for PII and PHI, such as insurance identification, date information, and procedure codes. Voltage SecureData Enterprise for Hadoop enables the company's data science team to perform analytics on claims data inside the Hadoop environment and produce insights on usage trends, market baskets, and the identification of new products and services without exposing the team to sensitive data. But when the team's analysis identifies specific customer health risks or the need for a procedure or medication recommendation, the customer can be re-identified by authorized personnel.

opentext™ | Cybersecurity

OpenText Cybersecurity provides comprehensive security solutions for companies and partners of all sizes. From prevention, detection and response to recovery, investigation and compliance, our unified end-to-end platform helps customers build cyber resilience via a holistic security portfolio. Powered by actionable insights from our real-time and contextual threat intelligence, OpenText Cybersecurity customers benefit from high efficacy products, a compliant experience and simplified security to help manage business risk.