Voltage security

WHITE PAPER

Big Data, Meet Enterprise Security

Will Data Security and Compliance Issues Put Big Data Developments on Hold?

Large organizations worldwide are working to develop and deploy Big Data analytical facilities alongside their established business intelligence infrastructure. These initiatives are motivated in nearly equal parts by the conviction that new business insights and opportunities are buried in the avalanche of new data, by the knowledge that conventional business intelligence systems are unequal to the task, and by the fear that competitors will be first to master and exploit the available new data streams.

Because the phenomenon of Big Data analytics is only a few years old, few standards exist to ensure that these new systems and the analytical activities they support are successfully integrated into the existing policy frameworks that ensure governance, compliance and security. One of those critical policy domains—data security—has the potential to arrest many of these developments and block the realization of their business benefits if not adequately addressed. This paper presents a comprehensive solution that cost-effectively ensures the security of sensitive information in Big Data environments without impairing their operational flexibility or computational performance.

A Big Step Up for Data-driven Decision-making

Most accounts now distinguish Big Data from the established domain of enterprise management information by three characteristics first noted by Gartner: volume, velocity and variety.

Volume – Very large data sets—think terabytes and petabytes of information—are not a new phenomenon, but the rise of ecommerce and social media, the global distribution of machine intelligence in business networks and personal electronic devices, and the exponential growth of commercial and scientific sensor networks are making them commonplace. There are now many organizations with volumes of data that exceed the ability of conventional methods to organize, search and analyze in meaningful time intervals.

Velocity – One reason these data sets are so large is their unprecedented growth rate. In a recent Harvard Business Review article[1], Andrew McAfee and Erik Brynjolfsson report that:

- As of 2012, approximately 2.5 exabytes of data are created every day, a number that is expected to double roughly every 40 months.
- More data now crosses the internet each second than was stored in the entire internet just 20 years ago.
- It is estimated that Wal-Mart collects 2.5 petabytes of customer transaction data every hour.

Application: Healthcare Informatics

An agency of the U.S. government is using a Big Data facility secured by Voltage SecureData to share 100 TBs of patient healthcare information. Independent R&D institutes across the country access and analyze the data for emerging risks and patterns. All data is de-identified at the field level before release in full HIPAA/HITECH compliance. When researchers identify health risks that may impact a population with living members represented in the data, the agency is able to quickly and securely re-identify and contact the affected individuals for proactive treatment.

[1] *Big Data: The Management Revolution,* by Andrew McAfee and Erik Brynjolfsson, Harvard Business Review, October 2012

Variety – Big Data includes a wide and growing range of data types, many of them new: text messages, social media posts, ecommerce clickstreams, GPS location traces, machine logs and sensor measurements. Structured, unstructured or semi-structured, much of this data is incompatible with the relational database repositories at the heart of most business intelligence facilities.

Rapid Rise, Quick Commercialization

Until 2004, the three Vs seemed to put Big Data beyond the reach of practical commercial analysis. That's when Jeff Dean and Sanjay Ghemawat published their seminal paper on the MapReduce programming model developed at Google for parallel, distributed processing of large datasets on scalable commodity clusters. The model was quickly embraced by the open source community, leading to the Apache Hadoop project and the development of a complete software framework for distributed analytical processing. This success promptly launched startups like Cloudera and Hortonworks to commercialize the new technologies.

The combination of Big Data analytics based on the MapReduce programming model, open source software, and commodity hardware clusters offers some extremely appealing business benefits for organizations with large data sets at their disposal, including:

- The ability to derive business insights and competitive advantages from data streams that cannot be addressed with conventional BI tools. To ask questions that were previously unanswerable.
- The ability to respond more quickly and intelligently to changing business environments and emerging opportunities
- A game-changing cost differential of up to 20:1 relative to proprietary business intelligence solutions

The conspicuous success of online companies like Google, Yahoo and Facebook in using Big Data techniques to manage and query very large data volumes has stimulated intense interest and accelerating adoption in other industries. While up to 45 percent of annual investment remains targeted at social media, social network and content analytics[2], the majority of spending now represents a diverse range of market sectors, including financial services, communications, pharmaceuticals, healthcare and government. Each of these segments brings its own interests in sensitive data types: Social Security and national ID numbers, payment card account numbers, personal health records—each with its own set of security mandates.

Data Security: A Sinkhole in the Big Data Roadmap

As sensitive data flows into new Big Data facilities, many of them still pilot-stage developments, the issue of data security becomes an increasingly urgent problem for business sponsors eager to bring them into production. Unless these systems can be rendered compliant with the full range of global data security and privacy regulations, their potential business impacts may remain a matter of purely academic interest.

But data security in Big Data environments is no small challenge. Their processing and storage clusters typically encompass hundreds or thousands of nodes. The software stack is entirely open source, with many alternatives for most key components, most of them still in very active development. Compared to a proprietary business intelligence infrastructure, a Big Data facility presents a large attack surface with all the vulnerabilities associated with rapid, ongoing change.

The one similarity is the extreme sensitivity of administrators and business users alike to any imposition by security on query response times.

Existing Security Solutions: A Gap Analysis

In these environments, none of the conventional approaches to system and data security are satisfactory or sufficient:

Perimeter security and user access controls are essential starting points but inadequate on their own. Even the best solutions are sometimes defeated by today's blended, persistent threats.

File-system encryption only protects data at rest. Sensitive data is immediately exposed to theft or compromise as soon as it is decrypted for transmission and use by an application. Decryption on access is required because the encryption process destroys the original data formats, rendering it useless to applications without extensive recoding. Needless to say, this approach also introduces significant processing overhead for the continuous write encryption and read decryption.

Data masking is typically a one-way conversion technique that destroys the original data values. It is useful in de-identification for testing and development, but problematic when used in many analytic use cases. For example, if masked data is used in a financial fraud detection

[2] Gartner Research Note "Big Data Drives Rapid Changes in Infrastructure and $232 Billion in IT Spending Through 2016" 12 October 2012, ID: G00245237, by Mark A. Beyer, John-David Lovelock, Dan Sommer, Merv Adrian.

application it may be possible to identify suspicious transactions, but not to quickly recover the relevant user and account identities for corrective action. Data masking also requires the creation and maintenance of large lookup tables which quickly become a significant management project in their own right.

Needed: Data Security that's High Strength, Low Impact

What's needed to ensure the viability of Big Data analytics is a data-centric security solution that:

- Protects sensitive data wherever it is stored, moved or used, with no exposure between storage, transmission, and processing.
- Enables compliance with most global data security, privacy and data residency mandates
- Integrates quickly and affordably with existing infrastructure and adapts flexibly to new analytical applications and data sources.
- Allows quick policy-based retrieval of original data values by properly authorized and authenticated users and applications
- Imposes no significant overhead on analytical performance
- Preserves the formats and referential integrity of protected data, so that existing analytics and ad-hoc queries don't need to change

Application: Financial Regulation

A regulatory body within the financial service industry is using Hadoop analytics and SecureData to monitor brokerage transaction patterns for compliance violations, data breach risks, and financial market risks. Data is de-identified as it enters Hadoop, providing protection for data at rest AND data in use by the organization's data scientists. With the ability to perform analysis on data in its encrypted state, there is no need for constant encrypt/decrypt operations, and virtually no impact on performance.

The Data-Centric Solution: Voltage SecureData for Hadoop

There is a solution that fulfills all these requirements and is already successfully deployed in Big Data facilities. Voltage SecureData™ for Hadoop from Voltage Security® delivers high performance data security with extensibility, scalability and adaptability based on groundbreaking cryptographic technologies that protect sensitive data independently of the systems that use it. Any type of sensitive data—structured, unstructured, semi-structured—can be encrypted at the time it is acquired by the organization, or at the time it is loaded into the Hadoop file system. Thereafter, most applications use the encrypted data, which retains its original format. Decryption occurs only when necessary, ensuring continuous protection in storage, in transit, in use, and at all points in between. Essentially, Voltage SecureData for Hadoop functions as an on-demand encryption-decryption service that can be called at any time by any properly authorized and authenticated user or application, such as Hive.

Voltage SecureData for Hadoop is part of the Voltage Security data protection portfolio, which also includes Voltage SecureData™ Enterprise, Voltage SecureData Web™ and Voltage SecureData Payments™.

Voltage Format-Preserving Encryption

Voltage SecureData for Hadoop uses Voltage Format-Preserving Encryption™ (FPE), an innovative implementation of FFX-mode AES encryption, to provide high-strength encryption of data without altering the original data format. Most applications can then use the protected data in normal processing, eliminating both the security vulnerabilities and processing overhead of routine decryption-re-encryption workloads. The technique is reversible and deterministic, meaning that multiple encryptions of the same plain text will always yield the same ciphertext. This ensures that not only format is preserved, but referential integrity across all instances of the same encrypted data. Database joins and other routine operations execute normally using FPE encrypted data.

Unlike some other solutions, Voltage SecureData for Hadoop does not encrypt entire files at the Hadoop file system (HDFS) layer. Instead, it protects only designated fields that hold sensitive data. It integrates easily with custom data ingestion scripts and with leading ETL solutions, allowing on-the-fly encryption of aggregated data as it loads into Hadoop, or selectively inside MapReduce jobs. Voltage SecureData virtual appliances deliver linear scalability to support very large clusters, and encryption processing can be offloaded to the Hadoop nodes for very high throughput performance with multimillion encryption operations per second.

Voltage FPE technology is patented, and recognized by the US National Institute of Standards and Technology (NIST).

Voltage Stateless Key Management

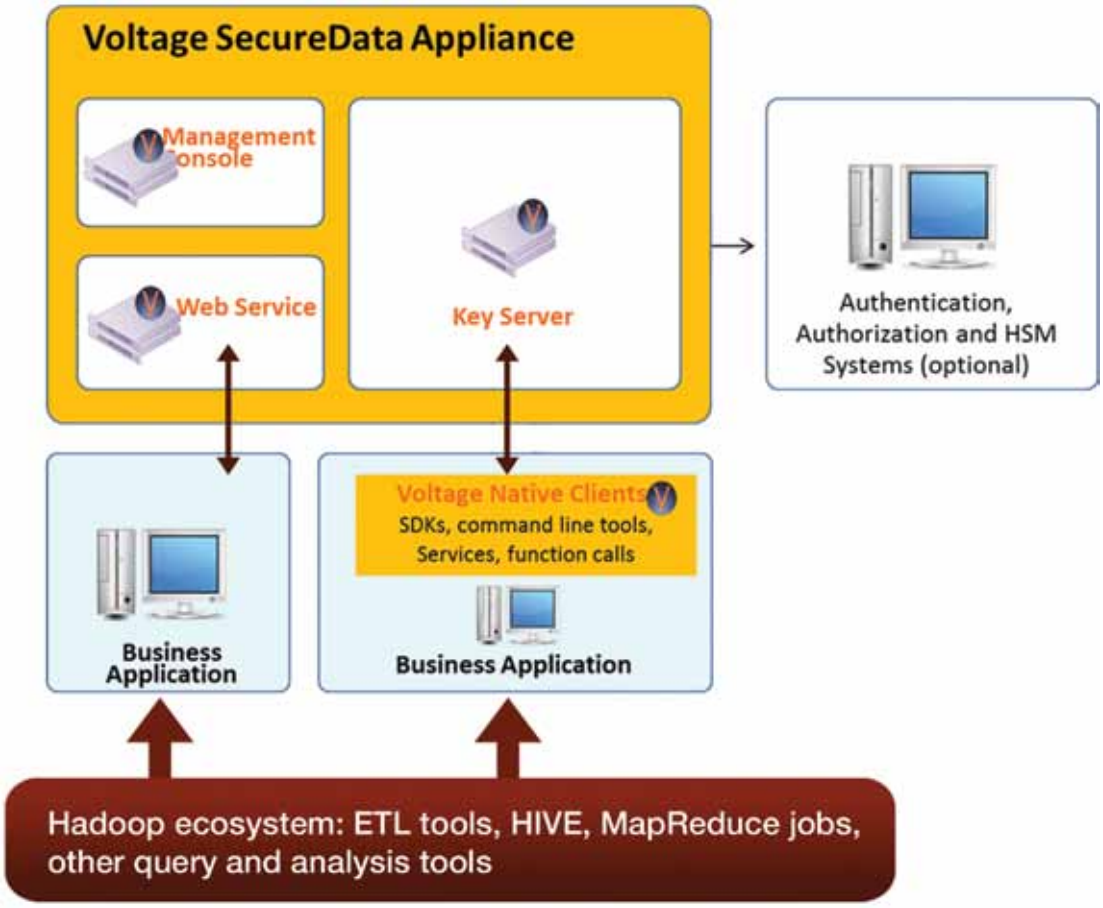
Voltage Stateless Key Management provides keys automatically with no storage or database management issues because database synchronization and frequent backups are not required. Key management can be linked to existing identity management infrastructure, including external LDAP directories. Permission to decrypt or de-tokenize can be assigned on an application or user basis, and can incorporate user roles and groups to simplify management. The result is role-based access to data at a data field level, mapping directly to enterprise data access rules and policies.

Voltage Secure Stateless Tokenization (SST)

Voltage Secure Stateless Tokenization™ (SST) technology is an advanced, patent pending data security solution that provides enterprises, merchants and payment processors with a new approach to help protect payment card data. Voltage SST technology is stateless because it eliminates the token database which is central to other tokenization solutions, and removes the need for storage of cardholder or other sensitive data. Voltage Security has developed an approach to tokenization that uses a set of static, pre-generated tables containing random numbers created using a FIPS random number generator. These static tables reside on virtual appliances—commodity servers—and are used to consistently produce a unique, random token for each clear text Primary Account Number (PAN) input, resulting in a token that has no relationship to the original PAN. No token database is required with SST technology, thus improving the speed, scalability, security and manageability of the tokenization process. The SST solution eliminates the token database, and with that, the complexities of database synchronization, back-up and recovery – with no need for synchronization, there is a 100% guarantee of data integrity and no data collisions. Voltage SST technology's foundations in proven principles and independent validation give enterprises and auditors the assurance and peace of mind that security and scope reduction are proven, not just claimed.

Swift, Simple Integration

Voltage SecureData integrates quickly and easily with virtually any application, from legacy custom systems to the latest enterprise applications. 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 key components of the Hadoop software framework.

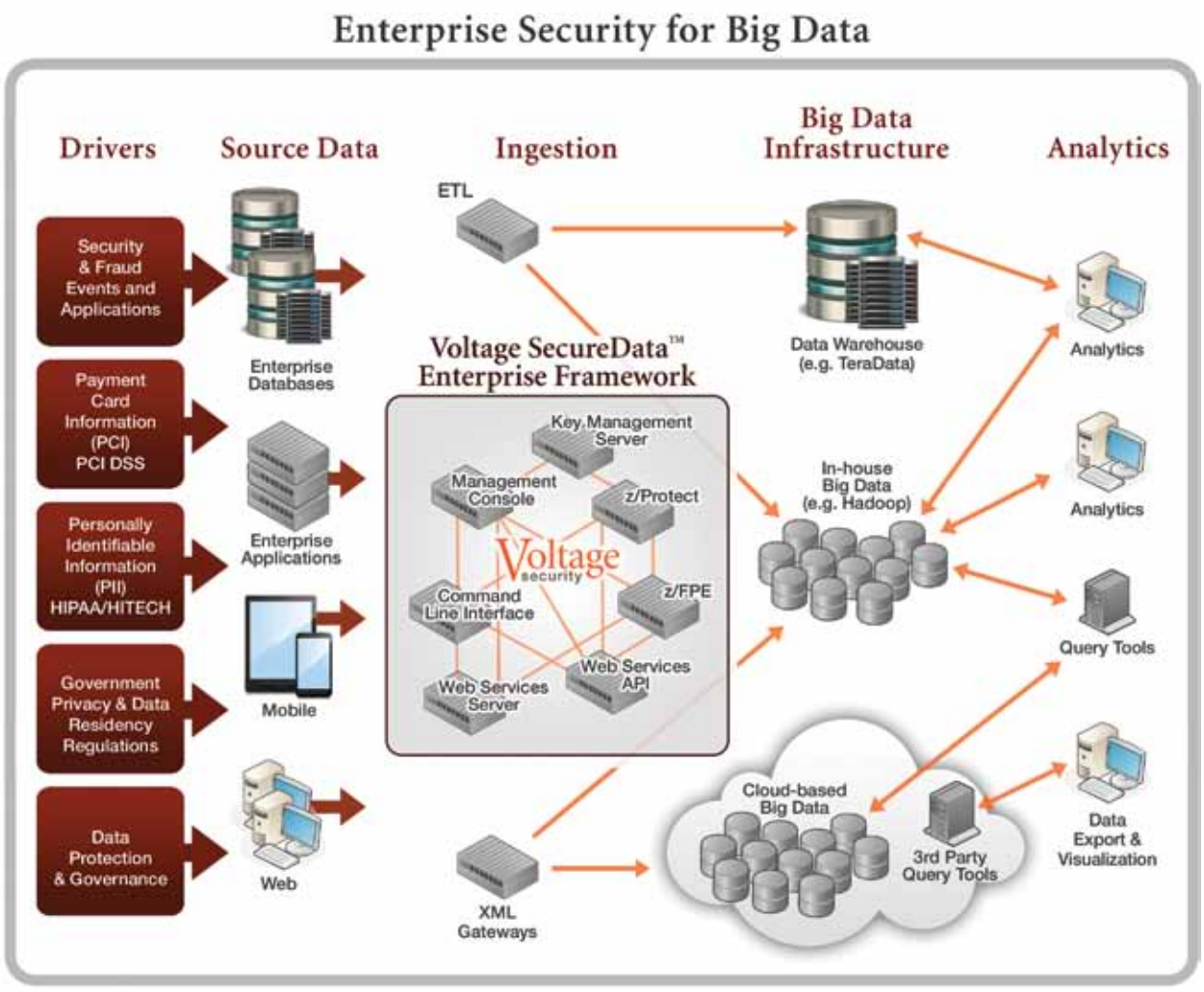


Voltage SecureData deploys as a set of services running on one or more virtual appliances, providing scalable key management, a management console, and a web services interface to business applications. A java-based SDK and four high-level APIs (in addition to web services) make it easy to add SecureData function calls to any application.

Core components of the Voltage SecureData platform include:

- The Voltage SecureData Management Console, which enforces data access and key management policies and eliminates the need to configure each application because flexible policies are centrally defined and reach all affected applications.
- The Voltage Key Management Server, which eliminates key storage and management because keys are dynamically derived. The key server integrates seamlessly with existing identity management and authorization systems permitting FIPS 140-2 hardware key management through hardware security modules.
- The Voltage SecureData Web Services Server, which provides centralized encryption and tokenization for Service Oriented Architecture environments, enterprise applications and middleware.
- The Voltage SecureData Simple API, which maximizes efficiency on a broad range of application servers through native encryption on HP/UX, HP NonStop, Solaris, Linux, AIX, and Windows. Specialized versions of Simple API plug into Teradata and Hadoop.
- Voltage SecureData z/Protect, which maximizes performance on mainframe systems through native z/OS support.
- Voltage SecureData z/FPE, a mainframe data processing tool to fast track integration into complex record management systems such as VSAM, QSAM, DB2 and custom formats.
- Voltage SecureData Command Line, a scriptable tool for bulk encryption to easily integrate encryption in files and databases.

SECUREDATA ENTERPRISE OPERATION IN A HADOOP ENVIRONMENT



In a typical SecureData for Hadoop implementation, one or more SecureData appliances provide key management services to all applications and infrastructure components that use or support the environment. These may include an ETL system, query tools such as Hive, analytical tools and custom MapReduce jobs, each communicating with the key server through a Voltage plug-in.

All sensitive data in the Hadoop file system is stored in protected form, having been encrypted either during ingestion or upstream in the originating systems. Because format is preserved during encryption, most analytical processing uses the data without decryption. When a job does require decrypted values, the application plug-in requests the appropriate key from the SecureData appliance, which authenticates the job (or user) to the enterprise identity management system before responding. Decryption processing then leverages the parallel execution resources of the full Hadoop compute cluster to optimize performance.

In an enterprise implementation, business applications would utilize the SecureData service in the same way to encrypt sensitive data at the point and time of acquisition. That data would then remain protected throughout its entire lifecycle, regardless of how many times it might be copied, transmitted, replicated, repurposed, or analyzed, including within a Hadoop environment.

A Better Way to Secure Big Data

SecureData for Hadoop delivers a single framework that protects sensitive information at the data level and throughout the data lifecycle. It provides safety in storage, transmission, aggregation and use, ensuring that any type of data can be safely aggregated into a Big Data analytical environment, and that any pilot-stage Hadoop facility can be safely moved into production.

For projects that must meet industry or national standards for data security, residency and privacy, Voltage SST and patented FPE techniques enable dramatic reductions in audit scope, costs and complexity. Where data is aggregated from global sources and across national boundaries, jurisdiction-specific policies can be applied on decryption access to remove data residency restrictions. With SecureData, Hadoop environments can meet and exceed most security requirements for Personally Identifiable Information (PII) and Personal Health Information (PHI), and be rendered out-of-scope for Payment Card Industry Data Security Standard (PCI DSS) audits.

Finally, Voltage SecureData for Hadoop eliminates the high-frequency full-file decryption and re-encryption workloads that sap application performance and confine protection to data at rest. And unlike traditional “data masking” solutions that perform one-way de-identification, Voltage technologies allows original data to be retrieved when needed. Sensitive data elements can be FPE-encrypted at the source or when loaded into the Hadoop file system, then used for Big Data analytics without decryption. SecureData integrates into existing environments in a fraction of the time required by conventional data security solutions, and enables flexible, cost-effective adaptation of the data management infrastructure as business requirements evolve.

Protecting the Business Potential of Big Data

Voltage SecureData delivers comprehensive data security across the enterprise and within the Big Data environment. For organizations now launching a Big Data program, Voltage SecureData for Hadoop can be up-and-running in weeks to secure sensitive data before it enters Hadoop environments. Voltage SecureData helps IT organizations respond successfully to the business need for near-term returns on Big Data investments, without risk to sensitive information or regulatory compliance. Enterprise Security for Big Data from Voltage Security removes the top obstacles to moving forward with Big Data initiatives, and enables integration of Big Data analytics and insights broadly, throughout the extended enterprise.

ABOUT VOLTAGE SECURITY

Voltage Security®, Inc. is the leading data protection provider, delivering secure, scalable, and proven data-centric encryption and key management solutions, enabling our customers to effectively combat new and emerging security threats. Leveraging breakthrough encryption technologies, our powerful data protection solutions allow any company to seamlessly secure all types of sensitive corporate and customer information, wherever it resides, while efficiently meeting regulatory compliance and privacy requirements

For more information, please visit **www.voltage.com**.

v03-19-2013