Success Story

Secure Analytics in Hadoop - Retail/Pharma Supply Chain

HP Security Voltage

"HP SecureData allows us to take our entire Hadoop cluster out of scope for the purposes of HIPAA compliance audits. Without that ability, we would have had to build a new logging process, which would have taken us months or even a year, delaying our time to market."

– Leading Supply Chain Data Analytics Solution Provider

Long Story Short

A leading supply chain data analytics solution provider needed to provide its large retail pharmacy clients with insights to enable product price optimization, store performance management, and targeted digital promotions to store customers. To do this it needed to analyze vast amounts of customer health, pharmacy claims, and purchasing data–much of it sensitive Personally Identifiable Information (PII) and HIPAA regulated data–on the fly. The provider determined that Hadoop was the only environment in which it could achieve that kind of performance on complex data sets. Aware of the data privacy and regulatory requirements for protection of sensitive data, the company also realized it required a security solution that would provide protection for data at rest, in use and in transit. It found that solution in HP SecureData for Hadoop.

Pressing Business Challenges

- Design a Hadoop "data lake" to support the competitive objective to transform the company's services through advanced technologies
- Enable data scientists to do real-time analytics on huge data sets, including raw transactional data, to produce insights for client companies
- Ensure compliance with HIPAA, as well as internal and complex regulatory data privacy policies, in a cost-effective manner for multiple client companies
- As much as possible, keep the Hadoop cluster out of scope for HIPAA compliance, in order to contain compliance management costs
- De-identify sensitive data to protect against cyber-attack
- Enable analytics on protected data and maintain referential integrity for database joins

Solution Considerations

The provider sought a data-centric security solution that would:

- Accelerate time to market for new analytic data services offerings
- Reduce time required to produce actionable analytics to improve clients' business performance
- Scale to enable lightning fast analytics on huge, complex data sets
- Protect multiple types of data from multiple sources and real-time feeds and ensure end-to-end data protection in transit, in use and at rest
- Ensure the usability of protected/ de-identified data for analytics designed to produce actionable business insights
- Be compatible with existing internal and external platforms and systems

Solution

Given its need to perform real-time analysis on 150+ terabyte data sets, the provider realized it would not be able to achieve that objective within its existing RDBMS systems. It decided to deploy a Hadoop cluster, initially comprising 14 data nodes and 3 management nodes. Additional nodes and clusters could be easily added as its data services business grew. Also, because much of the data it would be analyzing was subject to HIPAA regulations, it needed a data protection solution to ensure compliance. It researched available data protection solutions and found there was only one solution that could mitigate the new and unique data risk management challenges introduced by the implementation of Hadoop: HP SecureData for Hadoop.

Employing HP Format-Preserving Encryption (FPE) and HP Secure Stateless Tokenization (SST) technologies, HP SecureData for Hadoop de-identifies data before it reaches the HDFS so that even in the event of a data breach, nothing of value is revealed to unauthorized users and the data cannot be monetized by cyber-thieves.

By implementing HP SecureData, protection is applied at the data field and sub-field levels while the characteristics of the original data (including numbers, symbols, letters and numerical relationships such as date and salary ranges) are preserved. Referential integrity across distributed data sets is maintained so that joined data tables continue to operate properly.

HP SecureData encryption/ tokenization protection can be:

- Applied at the source before it gets into Hadoop,
- Evoked during an ETL transfer to a landing zone, or
- Evoked from the Hadoop process transferring the data into HDFS (for example, using Sqoop).

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 the HP Security Voltage system.

Hadoop and HP SecureData Deliver Solid Results

- Data scientists can now securely analyze terabyte-, or even petabyte-scale, real-time data sets and produce targeted customer promotional and store performance information on a weekly basis, compared to competitors' analyses on year-old data.
- By de-identifying data before it ever enters the Hadoop Data File System (HFDS) and other downstream analytics applications, the provider can reduce HIPAA compliance audit scope, saving significant time and money.
- Because HP SecureData maintains the referential integrity of data to be analyzed, there is no need for data analysts and scientists to ever see live production data in order to target customers for specific promotions or to provide pharmacy clients' store by store performance and trends.
- The provider has evolved from a business services provider (mediating coupon redemption, supply chain management and pharmacy reconciliation) to a strategic, integrated, information services partner that helps its clients grow and become more profitable.

"Based on previous experiences, I had some concerns about the impact of encryption on our analytics systems performance. In fact, we were pleasantly surprised to find that the combination of Hadoop with HP Format-Preserving Encryption allowed us to achieve super fast analytic performance while ensuring the protection of the sensitive data entrusted to us by our clients"

– Leading Supply Chain Data Analytics Solution Provider