Fast, Low-Overhead Encryption for Apache Hadoop*

The Intel® Distribution for Apache Hadoop* software enables real-time analytics on massive data sets with enterprise-class data protection

Solution Brief

Intel® Xeon® Processors

Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI)

"The Intel® Distribution for Apache Hadoop* software on Intel® Xeon® based servers provides breakthrough data security for big data analytics. By taking advantage of Intel® AES-NI, which reduces the burden of encryption by up to 19x, this important data protection capability is delivered without sacrificing performance."

—Iddo Kadim, Director of Datacenter Technologies, Intel Corporation

Some of today's most successful companies are using Apache Hadoop* to create a competitive advantage by extracting real-time insights from petabytes of structured, semi-structured, and unstructured data. Yet data security is a concern for many of these businesses. How much personally identifiable information (PII) and intellectual property is buried within these companies' massive data sets, and what would a data breach do to their business?

Big data repositories often contain sensitive information, much of it obtained by monitoring customer behavior and employee output. A data breach could be catastrophic, yet strong security safeguards have not been integrated into widely used Apache Hadoop distributions, and for good reason. Big data technologies are implemented precisely because they enable fast analysis of large data sets; yet data encryption, the method of choice for protecting sensitive business data, is a compute-intensive process that can slow analysis, negating the primary value of an Apache Hadoop cluster.

The Intel® Distribution for Apache Hadoop software provides an answer to this dilemma. It includes built-in support for enterprise-class access controls (see the sidebar, Fine-Grained Access Controls). It is also optimized for Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI), a technology that is built into Intel® Xeon® processors. As described in this paper, Intel performance tests have shown that Intel AES-NI can accelerate encryption performance in an Apache Hadoop cluster by up to 5.3x and decryption performance by up to 19.8x.[1]

With this built-in support for access controls and fast, low-overhead encryption, companies can take full advantage of Apache Hadoop, while simultaneously protecting sensitive data and maintaining compliance with regulatory requirements, such as the Payment Card Industry security standard or the Health Insurance Portability and Accountability Act (HIPAA). They can safely process all their data, including PII, to make better, faster decisions about business-critical processes, such as portfolio management, telecommunications service delivery, and medical treatments.

Fine-Grained Access Controls

The Intel® Distribution for Apache Hadoop* software provides a flexible and efficient framework for managing and controlling user access to data and services using existing Kerberos* authentication solutions. Administrators can use Intel® Manager for Apache Hadoop software to create and manage access control lists (ACLs) and to authorize individual users for specific data tables and services. A variety of integrated features, such as wizard-based setup and encrypted key exchange, simplify the otherwise complex task of establishing strong, cluster-wide security safeguards.

An Enterprise-Ready Big Data Solution

The Intel Distribution for Apache Hadoop software delivers real-time big data processing and analytics for enterprise customers, with an integrated software environment that is optimized to deliver superior performance, security, and manageability on servers powered by Intel Xeon processors. The software package contains core components of the Apache Hadoop framework, including MapReduce*, Apache Hadoop Distributed File System* (HDFS*), Apache Hive* data warehouse infrastructure, Apache Pig* data flow language, and Apache HBase* database (Figure 1). It also includes Intel® Manager for Apache Hadoop software to simplify deployment and management.

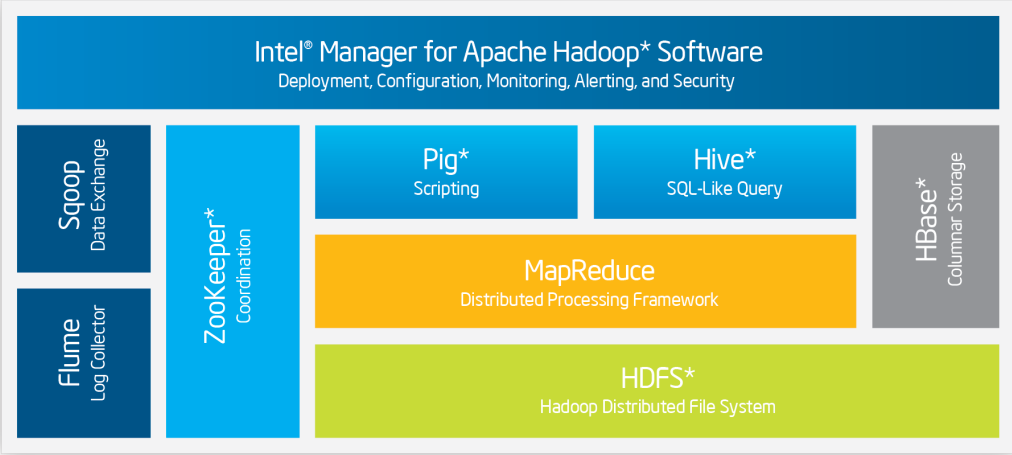

Figure 1. The Intel® Distribution for Apache Hadoop* software provides a complete solution for capturing, processing, and analyzing large data sets

The software stack is tuned to support scalable, real-time analytics on massive data sets, with the stability and reliability needed for mission-critical enterprise deployments. Intel is working closely with the open-source, academic, and vendor communities to deliver new capabilities, including:

- **Query federation**, so a single query can be distributed across multiple data repositories to provide more accurate and complete results.
- **True SQL support**, so data analysts can integrate big data analytics more effectively with their traditional analytics environments and use familiar analytic skills to extract higher value. The Intel Distribution for Apache Hadoop software already provides support for a variety of SQL-based queries using an HBase coprocessor application that parses and compiles SQL queries to run as MapReduce jobs.
- **Performance acceleration** through optimized, in-memory support for MapReduce applications.
- **Integrated security**, with access and data protection at each layer and with accelerated encryption capabilities.
- **Enhanced analytics**, with the ability to support graphing functionality across distributed data in an Apache Hadoop environment.

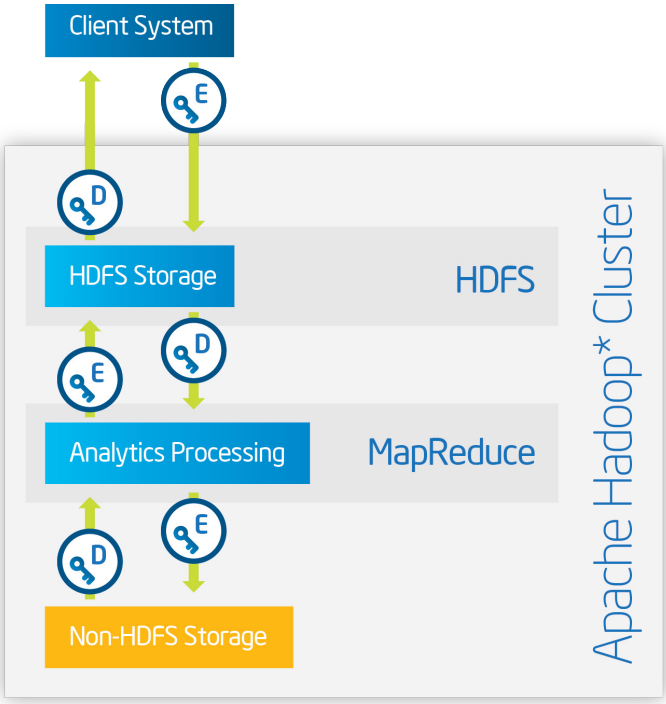

Figure 2. The Intel® Distribution for Apache Hadoop* software provides integrated, transparent, end-to-end support for data encryption

Intel is developing these innovative capabilities to extend and enhance open-source Apache Hadoop solutions. New functionality will be introduced first in the Intel Distribution for Apache Hadoop software and then offered as contributions to the open-source Apache Hadoop project.

Built-in Support for Data Encryption

The Intel Distribution for Apache Hadoop software provides built-in support for end-to-end data protection. Encryption is transparent to users, can be applied on a file-by-file basis, and works in combination with external key management applications. Java KeyStore* is currently supported, and future versions will support a broader range of standards-based key management solutions.

To take advantage of these capabilities, sensitive files must be encrypted by external security applications before they arrive at the Apache Hadoop cluster and are loaded into HDFS. Each file must arrive with the corresponding encryption key. This supports best practices for data security. If files were encrypted only after arrival, they would reside on the cluster in their unencrypted form, which would create vulnerabilities.

When an encrypted file enters the Apache Hadoop environment, it remains encrypted in HDFS. It is then decrypted as needed for processing and re-encrypted before it is moved back into storage (Figure 2). The results of the analysis are also encrypted, including intermediate results. Data and results are neither stored nor transmitted in unencrypted form, even if they are stored within the cluster in a file system other than HDFS.

Hardware Acceleration for Fast, Low-Overhead Encryption

Encryption and decryption are compute-intensive processes that traditionally add considerable latency and consume substantial processing resources. The Intel Distribution for Apache Hadoop software running on Intel Xeon processors helps to eliminate much of the latency and greatly reduce the load on the processors.

Encryption and decryption are performed using OpenSSL 1.0.1c*. This version of OpenSSL has been optimized by Intel engineers for Intel AES-NI. Intel AES-NI provides seven instructions that help to accelerate the most complex and compute-intensive steps of the AES algorithms. It also helps to make encryption stronger by protecting against "side channel" snooping attacks, which use sophisticated techniques, such as statistical analysis, to break encryption codes. [2]

To verify the performance benefits of Intel AES-NI in a big-data analytics environment, Intel engineers measured encryption performance for the Intel Distribution for Apache Hadoop software running on a server powered by the Intel Xeon processor E5 family (Table 1). Performance was measured with and without Intel AES-NI enabled. Test results showed that Intel AES-NI boosted performance for AES encryption by up to 5.3 times and boosted decryption by up to 19.8 times when running on the Intel Xeon processor E5 product family (Figure 3).

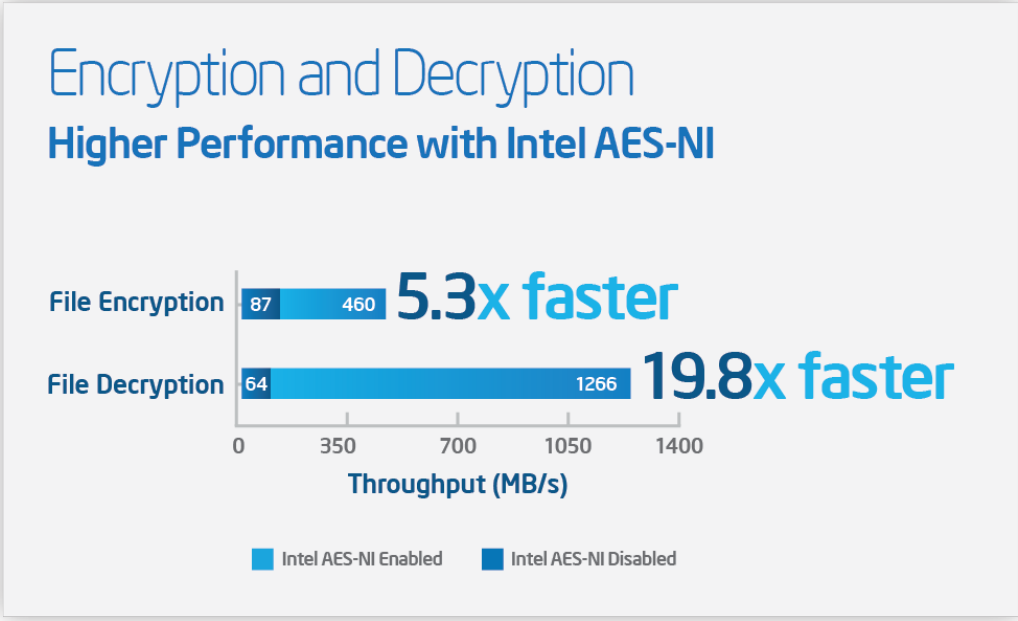

Figure 3. The Intel® Distribution for Apache Hadoop* software supports Intel® Advanced Encryption Standard New Instructions, which improves encryption performance dramatically when running on servers powered by the Intel® Xeon® processor E5 family

Table 1. Test environment: Encryption and decryption

Processor	4 x Intel® Xeon® processor E5-2690 (32 core system, 1 core used in testing)
Memory	32 GB
Operating System	CentOS 6.3*
Encryption Software	OpenSSL 1.0.1c* (with and without Intel® AES-NI enabled)
File System	Apache Hadoop Distributed File System* (HDFS*)—namemode, datanode, and the test program were all run on the same server
Storage	240 GB Intel® Solid-State Drive (SSD) 320 Series
Test Input	1 GB text file

Proven Performance in Realistic Scenarios

Data compression has become a best practice for Apache Hadoop implementations. Compression not only improves storage utilization, but also accelerates performance by allowing more data to be held in main memory, where it can be accessed much faster than from disk storage.

When encryption and compression are used together, data is both decrypted and decompressed as it is loaded into memory for processing, and then re-compressed and re-encrypted when it is transferred back into storage. To verify the benefits of Intel AES-NI in this more demanding environment, Intel engineers measured performance for:

1. Decrypting, decompressing, and reading data files from HDFS into main memory.
2. Compressing, encrypting, and writing data files from main memory into HDFS.

All tests were performed using the Intel Distribution for Apache Hadoop software running on a server powered by the Intel Xeon processor E5 family (Table 2). The Snappy* library was used for compression and decompression.[3] Tests were run with Intel AES-NI enabled and not enabled to quantify the performance benefits of hardware-assisted encryption.

The test results showed that Intel AES-NI improves data file transfer performance by 3.3x when decrypting, decompressing, and reading a data file from HDFS; and by 1.5x when compressing, encrypting, and writing a data file from memory back into HDFS (Figure 4). With these improvements, the performance penalties of encryption are greatly reduced, enabling IT organizations to provide strong protection for sensitive data, while maintaining the performance levels required to support real-time, big data analytics.

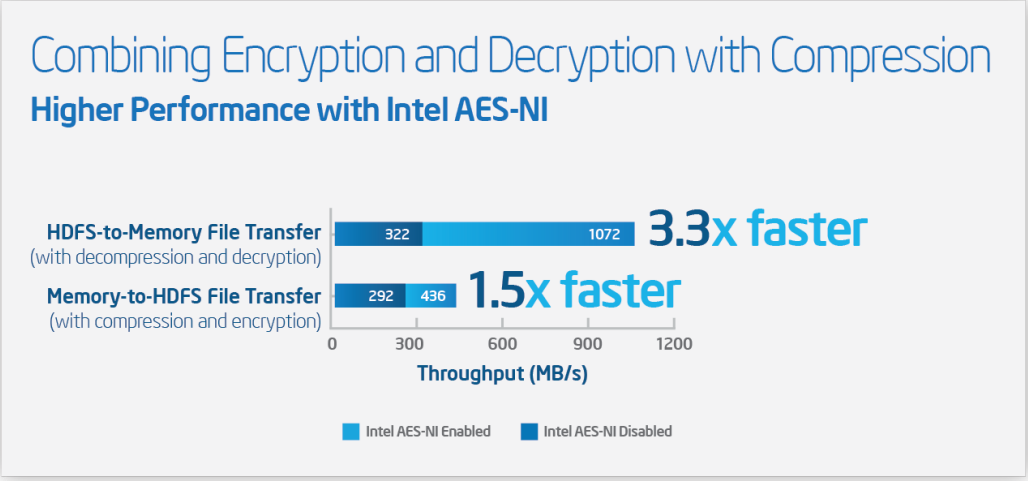

Figure 4. With the Intel® Distribution for Apache Hadoop* software and Intel® AES-NI, both file encryption and file compression can be implemented in Apache Hadoop environments—without sacrificing performance

Table 2. Test environment: Encryption with compression and decryption with decompression

Processor	4 x Intel® Xeon® processor E5-2690 (32 core system, 1 core used in testing)
Memory	32 GB
Operating System	CentOS 6.3*
Encryption Software	OpenSSL 1.01c* (with and without Intel® AES-NI enabled)
File System	Apache Hadoop Distributed File System* (HDFS*)—one namemode, seven datanodes, and the test program were all run on the same server
Storage	1 TB Hard Disc (ATA-ST1000NM0011)
Test Input	1 GB text file

Conclusion

Apache Hadoop offers a powerful tool for analyzing large and diverse data sets, yet the lack of integrated support for strong data security has been a serious roadblock to implementation for many businesses. The Intel Distribution for Apache Hadoop software provides an answer: a comprehensive, enterprise-ready software platform for big data analytics that is highly optimized for performance, stability, manageability, and security when run on servers powered by the Intel Xeon processor E5 family.

By taking advantage of Intel AES-NI technology, the Intel Distribution for Apache Hadoop software accelerates data encryption by up to 5.3 x and data decryption by up to 19.8x, so IT organizations no longer have to choose between performance and security. It also provides enterprise-class support for access controls. With these integrated capabilities, businesses can achieve the competitive advantages of big data analytics with less risk and with the confidence that their most sensitive data is protected.

[1] Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

[2] Side channel attacks can be contrasted with more traditional "brute force" attacks, which attempt to take advantage of any weaknesses that can be found within the encryption software algorithms.

[3] There are a number of compression and decompression libraries available. The Snappy* library is becoming the preferred library for Apache Hadoop due to its performance advantages.

Intel® Advanced Encryption Standard–New Instructions (Intel® AES-NI) requires a computer system with an AES-NI-enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® Core™ processors. For availability, consult your system manufacturer. For more information, see http://www.intel.com/content/www/us/en/architecture-and-technology/advanced-encryption-standard--aes-/data-protection-aes-general-technology.html.

Printed in USA 0213/KB/PRW/PDF Please Recycle 328649-001US