Networks & Communications Group

Enhancing Big Data Security

ADVANTECH

Enabling an Intelligent Planet

Enhancing Big Data Security

With the advent of Big Data comes the risk of greater security breaches as data volumes increase. Many companies are still trying to evaluate the potential of Big Data, let alone investigate the risks associated with Hadoop and the Cloud.

In the quest for new ways to house and exploit increasing amounts of unstructured data, companies need to ensure they have mechanisms in place which allow them to meet government compliancy regulations for data protection. Concerns about the security of stored data represent a significant barrier to the widespread adoption of Big Data, and in response, a number of companies are emerging with new products that secure data in ways which are practically transparent to the user.

One fundamental method is software and hardware encryption technology that operates on selected data on the fly or across an entire disk. However, software-based encryption adds significant extra load on a database server's CPU. This increases costs and overall complexity, particularly when the solution is required to scale.

The Intel® Communications Chipset 89xx Series, optimized for use with the Intel® Xeon® processor E5-2600 product family, offers a hardware-based data encryption and compression solution which alleviates this problem. When used in conjunction with additional Intel Communications Chipset 89xx Series units on hardware platforms such as Advantech's ATCA-based Netarium systems, or in servers deploying the Advantech PCIE-3214 PCIe card with four Intel® Communications Chipset 89xx Series devices, significant performance and cost savings can be achieved.

This paper investigates current trends in Big Data and Big Data Analytics, and looks at encryption techniques that provide the necessary integrated support for stronger data security.

ADVANTECH
Enabling an Intelligent Planet

Enhancing Big Data Security

So just how big is Big Data?

Einstein would probably say it's all relative, depending on where you're standing in time and space. If you're in today's traditional data storage business, then right here and right now, big data is huge and is tending towards becoming super massive – nobody needs an oracle to dispute that fact, do they? That is, until another scientific breakthrough happens and the storage world gets turned upside down.

Perhaps that may happen sooner than we think. According to a recent scientific discovery, we may all soon be walking around with the entire planet's filmography in a DNA device the size of a teacup. In 2012, a team of researchers headed by Nick Goldman and Ewan Birney at the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI) vividly demonstrated the potential of using DNA to store and transport human-made data. Vast amounts of data can be encoded and transported in microscopic volumes. The research team says that data can be written so efficiently that every film and TV program ever made could be stored in HD quality in just one cupful of DNA and kept for significantly longer than is possible today.

According to the team, DNA has already proven to be a robust and durable way to store information because we can extract it from bones of mammoths, which date back tens of thousands of years, and make sense of it. What's more, it's also incredibly small, dense, and does not need power for storage, so shipping and keeping it is easy, and it's less susceptible to failures. Today, however, the high cost and slow speed of writing and reading DNA does not make it very practical, but within a decade the technique is expected to be cheap and fast enough to make DNA data storage extremely viable.

From the article describing the experiment, which appeared in Nature magazine[1], it would appear that scientists at Agilent Technologies Inc. in California made a synthetic string of DNA that encoded an mp3 of Martin Luther King's "I have a dream" speech, a jpeg photo, a pdf of Watson and Crick's DNA research paper, and a text file of every one of Shakespeare's sonnets. The company then sent the DNA to the EMBL scientists near Cambridge, who then "read" the DNA code and reconstructed the digital files without any errors. Anybody with the correct software key could use a standard DNA reading machine to decode the information, the scientists say.

This paper sets out to analyze some of the issues we face today to secure current big data and goes on to suggest commercially available methods to address them. However, the vision of someone walking around the planet with all of earth's data in their smart phone or implanted in their body certainly implies we have some serious thinking to do. The scalability story may be about to take a considerable shrink, Moore's law could be surpassed on the grandest of scales, and anyone with the right software key could gain access to unlimited power beyond all expectation.
Does it make a difference if the world's data is spread across a network of authenticated servers or in the palm of everyone's hand?
You bet it does!

To quote a phrase attributed to Henley Stanley Haskins in a 1940s book on Wall Street: ***"What lies behind us and what lies before us are tiny matters compared to what lies within us."***

What exactly is Big Data?

Exactitude is also relative. When we're talking bytes, bits are often important. When we're discussing zettabytes: processing power, latency and security are more important.

Big data typically refers to collections of data sets with sizes beyond the ability of commonly used software tools such as database management tools or traditional data processing applications to capture, curate, manage, and analyze within a tolerable elapsed time. Big data sizes are constantly increasing, ranging from a few dozen terabytes in 2012 to today many petabytes of data in a single data set. To meet the demands of handling such large quantities of data, new platforms of "big data" tools are being developed. In a 2001 research report[2] and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e., increasing in volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data.[3] In 2012,Gartner updated its definition as follows: "Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."[4]

According to the Wikipedia entry for Big Data, the trend toward larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to, among many things, "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions." As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 quintillion (2.5×1018) bytes of data are created. The challenge for large enterprises is determining who should own Big Data initiatives that straddle the entire organization. Big Data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers." What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

Where does Big Data come from?

Some examples of Big Data sources are cited by Tom White in ***"Hadoop: The Definitive Guide"*** [5]**:**

- The New York stock exchange generates about 1 terabyte of new trade data each day.
- Facebook hosts approximately 10 billion photos, taking up one petabyte of storage—and millions more photos are added every week..
- Ancestry.com the genealogy site stores around 2.5 petabytes of information.
- The Internet Archive stores around 2 petabytes of data and is growing at a rate of approximately 20 terabytes per month.

- The Large Hadron Collider at CERN near Geneva produces about 15 petabytes of results from its experiments every year.

Other interesting examples of Big Data sources include Big Science, web logs, RFID, sensor networks, social networks, social data (due to the social data revolution), Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and often interdisciplinary scientific research. There's also military surveillance, map and forecasting applications for drive times, medical records, photography archives, video archives, and large-scale e-commerce.

What can you do with it?

There are many applications where Big Data could have far-reaching business impact. In the mobile phone business, it can help operators better analyze customer churn by identifying dissatisfaction triggers, discovering their cause, and potentially initiating remedial action before unhappy customers change carriers. Moreover, the results can be put to use for process improvement and personnel training. The importance of analyzing data is essential to identifying the points of rupture and subsequently understanding their cause. Boyd Davis, vice president and general manager of Intel's Datacenter Software division, gave a real-world example of the challenge of managing Big Data. One of the things cellular operators want to do is manage call data records, so people can keep track of their cellphone minutes and who they called. According to Davis, that was a challenge faced by a very large mobile operator in China who Intel had been working with. They wanted to put the call data records into a storage environment and make the data available to their consumers online. But they had hundreds of millions of users. When they put all that data into traditional databases, the databases simply broke from the scale and volume of the data. It was billions of records — the databases couldn't scale to handle it.

But with the Intel® Distribution for Apache Hadoop which is now broadly available, Intel was able to make that data accessible to customers in just a second or two. Customers type in a query, hit "enter," and it pops up in what we call "human real time."

But that's just the beginning. Now that the mobile operator has all the call data in its Hadoop framework, they can now do things like ask, "Which of our smartphones are the most profitable in terms of data plans and usage?" Or, "What's the most popular smartphone at a given time of the year, so we can direct manufacturing to reflect that?"

Intel® IT Uses Big Data to Bring Chips to Market Faster

EXTRACT from Chip Shot

Using Big Data and predictive analytics, Intel found it could achieve a 25 percent decrease in chip design validation time, accelerating the time to market for new processors. This is one of the highlights of the 2012-2013 Intel® IT Performance Report, which also showcases results for IT initiatives in cloud computing, BYOD and other areas.

A link to further information can be found here.

The market for Big Data is growing fast — it's forecasted to top $18 billion in 2013 and $47 billion by 2017, according to Wikibon. Lurking behind those big numbers is the reality that organizations often struggle to access and make use of the information within Big Data. Apache Hadoop, the open-source software framework, has emerged as an important technology for managing huge volumes of data. Intel introduced its Hadoop distribution software on February 26th, 2013. Boyd Davis, vice president and general manager of Intel's Datacenter Software division, explained the initiative in an interview on intelfreepress.com:

Hadoop: Getting Big Answers from Big Data

"Hadoop is a framework for managing Big Data. It's got three primary components. First, it's a way of storing data on a large scale. Second, it's a way to organize that data so that it can be accessible via a variety of different tools. And third, it is a set of tools that allows you to gain insights from the data.

Also, it's open source, so there's a community of programmers around the world contributing to it, as does Intel. So the Hadoop framework is not a single product or project. And because it's so versatile, we believe it has the potential to be a transformative technology. But, while a lot of organizations like the idea of downloading and using open source Hadoop code — because it's free — once they go into production with an application or service, they want somebody who can back them up. That's where we come in.

Intel's Hadoop distribution takes advantage of the fact that we have the most intimate knowledge of the underlying hardware, like our Intel® Xeon® processors. And we are getting substantial performance gains because of that."

The Hadoop framework was created by Doug Cutting and Michal J. Cafarrela. Cutting, who worked at Yahoo at the time, allegedly named it after his son's toy elephant.

He went on to add. "We have an example of the gains that can be made when you're sorting, say, a terabyte of data. Using a standard benchmark, on the previous-generation Xeon platforms using hard-disk drives and 1 gigabit Ethernet connections, and just the standard Hadoop distribution, it would take about 4 hours to sort that data. Now, we add in the newest-generation Xeon and we can cut that in half. Then you add solid-state drives , and that drops it down another 80 percent. You go from 1 Gig Ethernet to our new, faster 10 Gig Ethernet connection, and you drop it another 50 percent. And then if you use the Intel® Hadoop Distribution, you drop it another 40 percent.

Suddenly, from greater than 4 hours, you're down to just 7 minutes to run that workload. So yeah, there's this huge link to Intel® hardware which delivers the optimizations, and the software framework that takes advantage of the hardware, and all of a sudden you're giving a lot of value to customers.

Big Data Security

Before addressing the security issues related to Big Data, it may be interesting to note that you can actually use Big Data Analytics to secure it: "Big Data Security for Dummies" by Solera Networks goes into the details of how to harness the power of Big Data to detect advanced threats and targeted attacks by collecting digital evidence in order to streamline incidence response and help companies integrate Big Data into existing security fabrics. According to Solera Networks, security-conscious organizations are turning to Big Data Security as the newest weapon in their cybercrime arsenals.

By collecting all available digital evidence, including raw packets, flow data and files, organizations can uncover advanced targeted attacks traditional security defenses sometimes miss. Organizations are learning to use internal data sources they never knew existed and to extend the value of known data sources by integrating their Big Data Security solutions into their existing security fabric.

Enhancing Big Data Security

In the quest for new ways to store and exploit Big Data, companies need to ensure they have mechanisms in place which allow them to meet government compliancy regulations for data protection, especially for data at rest. Implementations must essentially involve two things. First, secure encryption technology must be used to protect confidential data, in particular Personally Identifiable Information (PII) and Protected Health Information (PHI), but also a company's own Intellectual Property (IP). Second,, careful management of access to the cryptography keys which unlock the encrypted data must be put in place.

Growing concerns about the security of stored data are creating new opportunities for IT vendors, and a number of companies are emerging with new products that secure data in ways which are practically transparent to the user. One method being applied is software and hardware encryption technology operating on selected data on the fly or across an entire disk of data at rest. However software-based encryption adds significant extra load on a database server's CPU and costs notably increase, along with complexity, when a solution is required to scale.

Data Protection Methods for Apache Hadoop

To address this, the Intel Distribution for Apache Hadoop software includes built-in support for enterprise-class access controls by providing a flexible and efficient framework for managing and controlling user access to data and services by means of 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.

The Intel Distribution for Apache Hadoop software 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 the Intel solution brief, "Fast, Low-Overhead Encryption for Apache Hadoop", 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.

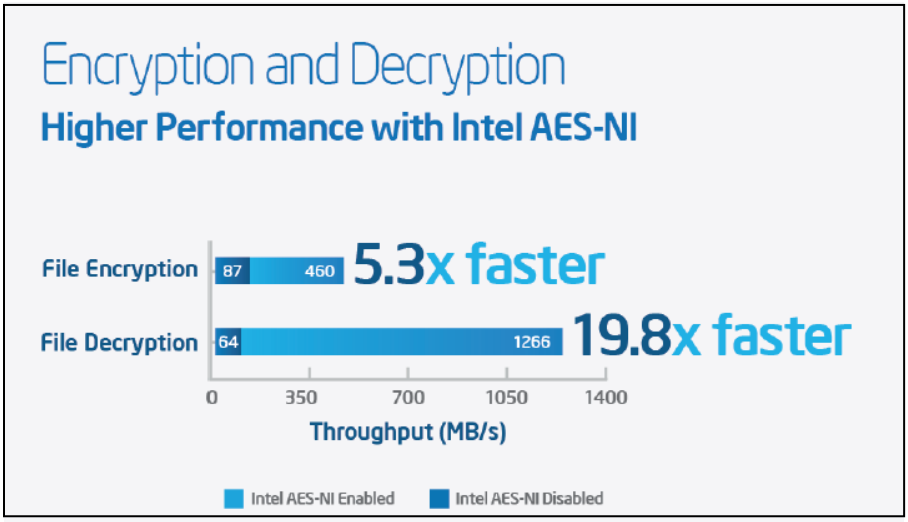

Figure 1. The Intel® Distribution for Apache Hadoop software supports Intel® Advanced Encryption Standard New Instructions. This improves encryption performance dramatically when running on servers powered by the Intel® Xeon® processor E5 family

Hardware versus Software 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 the Hadoop Distributed File System (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. 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.

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, which provides seven instructions that help to accelerate the most complex and compute-intensive steps of the AES algorithms. Intel AES-NI also helps make encryption stronger by protecting against "side channel" snooping attacks which use sophisticated techniques, such as statistical analysis, to break encryption codes.

The Intel Communications Chipset 89XX Series optimized for use with the Intel Xeon Processor E5-2600 series offers a hardware-based data encryption and compression solution which significantly accelerates encryption and compression. When used in conjunction with additional Intel® DH8910 Series chipsets, either on ATCA blades like the MIC-5333 shown below and used in Advantech's ATCA-based Netarium™ systems (Fig 4), or in servers deploying Advantech's quad Intel® DH8910 Series PCIe card (Fig 6), even greater performance increases can be achieved.

MIC-5333 Dual Intel® Xeon® E5-2600 Series Processor Blade

Advantech's MIC-5333 is based on the Intel® Platform for Communications Infrastructure. In addition to its two high performance dual Intel Xeon E5-2600 Series processors, its Intel® Communications Chipset 8910 incorporate acceleration and offload features for encryption and enhanced security. One unique advantage of the MIC-5333 is three available Fabric Mezzanine Module (FMM) sites for add-on modules. With two FMM Type I sites and one FMM type II connected to the front panel, the flexibility in blade function personalization is extensive. If higher encryption/decryption performance is required, up to 4 additional Intel DH8910 Series chipsets can be added.

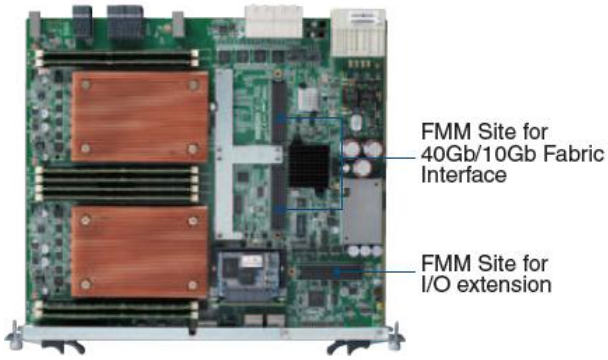

Figure2. MIC-5333 with 3 FMM sites

When coupled with the RTM-5104 Rear Transition Module shown in Figure 3, a fourth FMM site is available, making the standard COTS front ATCA blade and RTM highly configurable with an abundance of I/O and acceleration options. The RTM-5104 is a single slot (6HP) ATCA rear transition module for I/O extension of Advantech ATCA CPU blades.

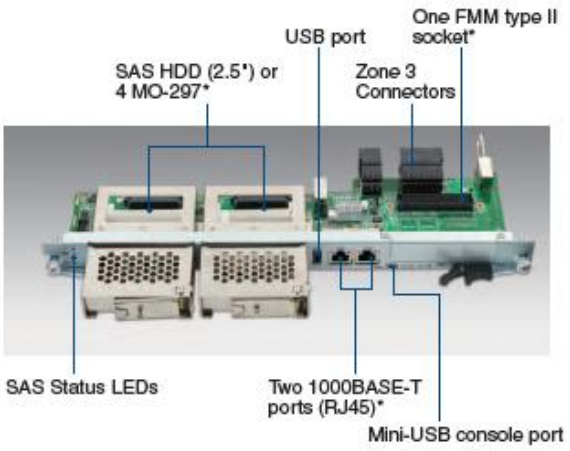

Figure 3. RTM-5104 Rear Transition Module with one FMM site, storage and I/O

Advantech Netarium™ ATCA System Platforms

Figure 4. Netarium™ AdvancedTCA Series 2-14 Slots, 16-200 Intel® Xeon® Cores. Configurations with up to 10x MIC-5333 and 40G, 80G or 160Gbps backplane connectivity per MIC-5333. Throughput >1Tbps. Supports Customized COTS (C^2OTS) framework.

FMM and PCIe Card

The FMM approach to integrating Intel Communications Chipset 89xx Series devices on an ATCA blade is shown below. Each FMM-based accelerator device is connected to the Intel Xeon Processor E5-2600 series through the PCIe Gen 2 x16 interface supporting up to 64Gbps. Typical configurations on an ATCA blade will provide support for over 80 Gbps security processing.

Figure 5. FMM-5006 Intel® Communications Chipset 89xx Series based Fabric Mezzanine Module

Figure 6 shows the Advantech PCIE-3214 PCIe card with four Intel® Communications Chipset 89xx Series devices. This card can be used to provide encryption and compression acceleration to any system or server with a standard PCIe slot.

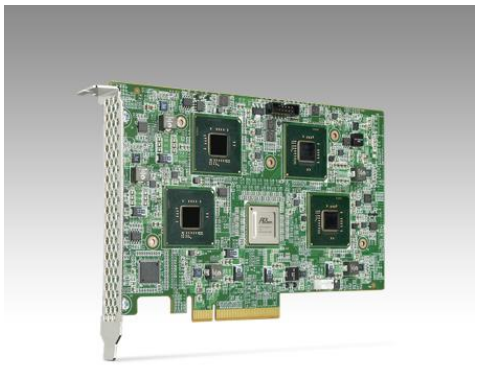

Figure 6. Advantech PCIE-3214 PCIe card with four Intel® Communications Chipset 89xx Series devices

The PCIE-3214 is a full height, half length PCI Express card supporting hardware acceleration for Intel®QuickAssist technology. Four onboard DH 8910 Series chipsets are complemented by a PCIexpress Gen 3 switch to fully utilize the

bandwidth offered by the latest Intel Xeon E5 processor family. Packaged in a standard full-height, half-length PCIe form factor, the PCIE-3214 is a perfect fit for hardware acceleration and offloading in high performance, high I/O throughput servers and appliances. Offering acceleration for common security and crypto offloads such as AES, 3DES, Kasumi and SNOW, the PCIE-3214 can supplement the CPU throughput for the termination of standard security protocols such as IPSEC and SSL, freeing up valuable cores and CPU cycles for application processing. With 20Gbps bulk crypto throughput and 28k RSA decrypt ops per accelerator device, the PCIE-3214 featuring more than 100k RSA decrypt ops offers best-in-class performance per watt at an outstanding price-performance ratio. Complemented with 7Gbps compression offload (LZS, Deflate) and even higher decompression offload per accelerator device, the PCIE-3214 can be of great benefit in Big Data analytics and storage applications.

The PCIE-3214 supports simultaneous crypto and compression offloading, making it an ideal choice for securing Big Data analytics as well as demanding applications such as Secure Storage, WAN and traffic optimization, and Secure Web Servers.

Fully supported by Intel® QuickAssist Libraries and the Intel® Data Plane Development Kit (Intel® DPDK), customers can use application software without modifications across Intel® platforms, minimizing time to market, total cost of ownership, and resource investment. Complementing Advantech's offering of standard blades, servers and appliances with built-in and scalable Intel QuickAssist offload, the PCIE-3214 rounds out the portfolio by bringing high performance offload to whitebox servers and proprietary platforms.

In Conclusion

Concerns about the security of Big Data have slowed the uptake of analytics projects across many businesses. However suppliers are now emerging with new products which secure data in ways which are almost transparent to the user. One of the methods being applied is software and hardware encryption technology operating on selected data on the fly or across entire disks of data at rest. However software-based encryption adds significant extra load on a database server's CPU and notably increases costs, along with complexity, when the solution is required to scale. Intel Distribution for Apache Hadoop software now offers a solution, providing an enterprise-ready software platform for Big Data analytics that is highly optimized for performance, stability, manageability, and security when run on systems 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.

Advantech platforms based on Intel Architecture for Communications Infrastructure provide a broader product offering with greater configurability, scalability and performance for hardware acceleration across multiple form factors. This enables business to achieve the competitive advantages of Big Data analytics with less risk and with the confidence that their most sensitive data is protected.

[1]http:/www.nature.com/nature/journal/vaop/ncurrent/full/nature11875.html

[2] Douglas, Laney. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. Retrieved 6 February 2001.

[3] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived retrieved 13 July 2011

[4] Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

[5] White, Tom (10 May 2012). *Hadoop: The Definitive Guide*. O'Reilly Media. p. 3. ISBN 978-1-4493-3877-0.

Advantech Contact Information

Hotline Europe: 00-800-248-080 | Hotline USA: 1-800-866-6008
Email: NCG@advantech.com.tw
Regional phone numbers can be found on our website at **http://www.advantech.com/contact/**

www.advantech.com/nc

ADVANTECH
Enabling an Intelligent Planet