

### Highlights

- Analyze massive data from all sources in real time
- Support advanced modeling and simulation
- Accelerate business processes
- · Help reduce storage costs significantly

## TerraEchos Kairos on IBM PowerLinux servers

A world leader in stream computing harnesses the power of real-time associative analytics for extreme workflow optimization in the big data arena

#### Why big data processing is important

Businesses and consumers generate almost inconceivable volumes of data every day. U.S. retail giant Wal-Mart handles more than one million customer transactions per hour, feeding databases estimated at over 2.5 petabytes in size. That is the equivalent of 167 times the volume of data contained in all the books held in the Library of Congress.<sup>1</sup> The New York Stock Exchange captures about 1 terabyte of trade information daily.<sup>2</sup> Google processes about 24 petabytes of data in a single day.3 And every 20 minutes, users upload 2.7 million photographs to Facebook.<sup>4</sup> To sum it up, 2.5 quintillion bytes of data are created every single day; in fact, 90 percent of the data existing in our world today was created in the last two years. To make matters even more challenging, because of the explosion of data generated by social media sources, 80 to 85 percent of all the world's data is now unstructured (text, audio, video, click streams, log files and so on). Today's businesses and industries must discover new and enormously more efficient ways to analyze unprecedented volumes of data that are derived from nontraditional, unstructured data sources.

<sup>4</sup> NW Linux: Linux and Technology News. December 31, 2010. http://nwlinux. com/facebook-data-and-post-statistics-over-20-minutes-time



<sup>1 &</sup>quot;Data, data everywhere." *The Economist.* February 25, 2010. www.economist. com/node/15557443?story\_id=15557443

<sup>2</sup> Henschen, Doug. "New York Stock Exchange Ticks on Data Warehouse Appliances." *Information Week*. May 16, 2008. www.informationweek.com/news/software/ bi/207800705

<sup>3</sup> Dean, Jeffrey and Ghemawat, Sanjay. "MapReduce: simplified data processing on large clusters." *Communications of the ACM*. Volume 51, Issue 1, January 2008. http://portal.acm.org/citation.cfm?doid=1327452.1327492

IBM® has been working with the U.S. government and strategic IBM Business Partners for a decade to develop a radical new approach to data analysis that enables high-speed, scalable, and complex analytics of heterogeneous data streams in motion. This environment is referred to as big data. IBM InfoSphere® Streams is integral to this new generation of big data analytic-processing methods. As a highly scalable, agile software infrastructure, it enables businesses to perform in-motion analytics on a huge variety of relational and nonrelational data types at unprecedented volumes and speeds — from thousands of real-time sources.

"TerraEchos not only understands that the world is moving toward data in motion. We understand how to harness and glean actionable insights from these data streams. We view PowerLinux hyperthreading, parallel file system and other features as being right on target with our Linux server needs."

## – Dr. Alex Philp, Founder and CTO TerraEchos, Inc.

TerraEchos, Inc., headquartered in Missoula, USA, already a world leader (and award-winning IBM Business Partner) uses InfoSphere Streams as a core data processing component. For example, in one high-security environment that involved miles of optical sensors, cameras, and satellite feeds, an early version of the TerraEchos Kairos<sup>™</sup> offering was able to reduce the time required to process 275 MB of data from hours to a mere fourteenth of a second. This performance improvement in analyzing data not only provided enormous time savings, but more importantly, enabled reaction times that were otherwise impossible to achieve before. (As a side benefit, the application also removed the need to store all of this data, thus reducing costs even more dramatically.)

# How PowerLinux performs in the big data arena

The IBM PowerLinux<sup>™</sup> platform is important to this scenario because TerraEchos has observed big data throughput and analysis that is as much as two times faster than the remarkable speed just mentioned. Dr. Alex Philp, founder and CTO of TerraEchos explains additional reasons for using PowerLinux servers for Kairos implementations, "PowerLinux improves the run time for data analysis by 50 percent while delivering a cost per transaction that is astoundingly low. The PowerLinux JVM is highly optimized to make it three times faster than some other JVMs on the market, and PowerLinux disk and storage performance is better by a factor of two. Coupled with the reliability, scalability, and manageability of IBM Power Systems<sup>™</sup> servers, we view PowerLinux as a best-server solution in many environments."

#### TerraEchos Kairos on PowerLinux

Originally developed for military, government, and high-value civilian environments such as nuclear facilities and smart grids, the rack-mounted Kairos compute appliance can be customized to a client's specific location and needs for the purpose of analyzing and acting on data from a wide variety of sources — cyber and physical — instantaneously, providing security against cyber-attacks. However, Kairos is now proving its huge value in many other big data environments.

One engineering application for Kairos involves the complexity of oil and gas exploration, where a typical well generates multiple TB of data per day. This is an enormous amount of sensor data, and because the drilling effort is so costly, there is no time to index, match, and reduce the data in an offline, batch-processed manner. In contrast, Kairos is capable of processing streaming data so fast that critical tuning of the drilling process can happen in real time. "Embedding IBM InfoSphere Streams and operating on PowerLinux enables Kairos to provide a user-friendly solution for performing complex, large-scale analytics. The TerraEchos partnership with IBM means that enterprises of all sizes can manage the massive volume, variety, and velocity of data that consumer and businesses create every day – and they can do so reliably and cost-effectively."

## - Dr. Alex Philp

By their very nature, Kairos big data workloads are ideal for PowerLinux servers. The reasons for this are that Kairos exploits the PowerLinux hyperthreading technology to maximize high throughput. Additionally, Kairos requires high I/O performance and uses a parallel-file system (PFS) across multiple servers. PowerLinux provides a PFS that is synergistic with the Apache Hadoop Distributed File System (HDFS) framework. PowerLinux servers allow Kairos to respond to millions of events per second while running thousands of tasks in parallel to deliver real-time analytics services.

#### PowerLinux blasts through data analytics and helps save money

Kairos literally turns huge volumes of highly disparate, streaming data into immediately viable insights (for example, security threats, exploratory drilling feedback, and market reactions).

Kairos integrates with and taps directly into the backbone of any IT network — localized or cloud-based — and works similar to a central nervous system for processing and analyzing structured and unstructured data as it streams across networks and the cyber infrastructure. Its mission is to detect anomalies in real time, even as the data streams are moving at thousands of megabytes per second. Computationally challenging analytics are conducted in a blink, and anomalies are verified on the fly to eliminate false positives. Kairos adapts to the monitoring criteria and its environment over time. As mentioned before, a side benefit of this essentially instant analysis is the elimination of the need for expensive space-consuming data storage.

Kairos uses IBM InfoSphere Streams to collect data from multiple sensor types, thus enabling associated streams of structured and unstructured data that can be integrated into a system that detects, classifies, correlates, predicts, and communicates discreet patterns and trends. All of this is accomplished by means of a service-oriented architecture (SOA). Based on this technology, TerraEchos provides one of the most robust classification systems in the industry, and was the first company to license InfoSphere Streams from IBM as the computational platform for sensor data analytics (referred to as *sensor as a service*).

#### Why PowerLinux

IBM PowerLinux Big Data Solutions are deeply optimized systems, from the hardware, to the Linux® operating system and IBM software offerings. IBM is in a unique position to use two decades of Linux and open-source experience, as well as IBM research technologies and hardware expertise from the microprocessor up. These are among the reasons that readers of the Linux Journal selected IBM as the winner in the Best Linux Server Vendor category in its 2011 Readers' Choice Awards.<sup>5</sup>

IBM has invested more than US\$1 billion in the Linux and open-source community. IBM is consistently among the top commercial contributors of Linux code, with more than 600 IBM<sup>6</sup> developers involved in over 100 open-source projects and thousands of dedicated development and support personnel supporting all of the IBM products and customers on Linux. IBM not only supports this community, IBM is part of it.

<sup>5</sup> Linux Journal award, www.linuxjournal.com/slideshow/readerschoice-2011, December 1, 2011

<sup>6</sup> http://go.linuxfoundation.org/e/6342/ho-writes-linux-2012/879kf/221839922

IBM has completed tens of thousands of Linux customer engagements, and facilitated thousands of migrations to Linux and offers the widest range of hardware, middleware, and services products for Linux in the industry. IBM supports Linux on all modern IBM Systems and backs PowerLinux with a full line of implementation, support, and migration services.

#### For more information

To learn more about TerraEchos and IBM PowerLinux servers, contact your IBM marketing representative or IBM Business Partner, or visit the following web sites:

ibm.com/partnerworld/powerlinux and www.terraechos.com.



© Copyright IBM Corporation 2012

IBM Corporation IBM Systems and Technology Group 3039 Cornwallis Road RTP, NC 27709

Produced in the United States of America June 2012 All Rights Reserved

IBM, the IBM logo and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at ibm.com/legal/copytrade. shtml

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Other product, company or service names may be trademarks or service marks of others.



Please Recycle