IBM PowerLinux Big Data Analytics Solutions

Watson-inspired analytics solutions for businesses of all sizes

Data is exploding from sources like sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and cell phone GPS signals, to name a few sources. Every day, we create 2.5 quintillion bytes of data, so much that 90 percent of the data in the world today has been created in the last two years alone. This data is big data.

Today’s businesses want to leverage the data explosion to create competitive advantages. In fact, 89 percent of CEOs say they want better insight through business intelligence and analytics. An infrastructure that is designed for data means extending beyond traditional sources of data to generate insight by leveraging new forms of information.

Both structured and unstructured data will continue to grow at astronomical rates. Today’s organizations are challenged with managing the volume of data they have. It’s difficult to keep pace with the growing amounts of incoming data that could provide vital information to their business to make timely decisions and achieve business goals.

Smarter companies are thinking differently about how to deal with big data that is growing exponentially within their organizations. To meet this challenge, many are deploying Linux-based IT infrastructures designed to capably handle both structured and unstructured data, making it easier to capture, manage and analyze information to drive better-informed business decisions.
IBM PowerLinux Big Data Analytics solutions help businesses gain new insights with scalable, powerful solutions, using Apache Hadoop and IBM InfoSphere BigInsights software to analyze data-at-rest, and InfoSphere Streams software to analyze data-in-motion. The deep integration and optimization of analytics workload performance on PowerLinux enables businesses to run thousands of tasks in parallel to deliver analytics services faster.

**PowerLinux Big Data solutions are industry standard, tuned to the task**

Today organizations are only tapping in to a small fraction of the data available to them. The challenge is figuring out how to analyze ALL the data, and find insights with new and unconventional data types.

IBM stepped up to the challenge with new developments from IBM Research extending available open source technologies. Based on PowerLinux and leveraging Apache Hadoop technology, Watson embodies optimized system design by combining open source Hadoop software running on a cluster of POWER7 systems with IBM’s DeepQA analytics software and the domain knowledge and expertise of IBM’s Research team. Watson was tuned to the task of answering questions posed in natural language in less than three seconds by executing thousands of complex analytical tasks simultaneously. The result—a brilliant landmark win on TV’s Jeopardy game show—was impressive, but merely hinted at the real-world value of Linux running on workload-optimized systems. And, these same principles that enabled Watson to master natural language queries can be put to work for today’s businesses. We understand how to design a system that is optimized across the full technology stack to maximize performance and efficiency. IBM had a choice of systems to use when designing Watson. We chose PowerLinux servers because they are fit for this purpose. IBM PowerLinux Big Data Solutions leverage Apache Hadoop and IBM InfoSphere BigInsights for Hadoop-based data-at-rest analytics and IBM InfoSphere Streams for real time data-in-motion analytics.

Health insurer WellPoint is implementing a PowerLinux Big Data solution to help medical professionals make better decisions right at the point of care for patients with serious illnesses. These workload-optimized systems are helping reduce the time to diagnose and treat illnesses.

Apache Hadoop is an open source framework that allows for the distributed processing of large data sets across clusters of computers. It is architected to scale out from single servers to thousands of machines, each offering local computation and storage and can help organizations perform analysis swiftly, including unstructured data which isn’t possible to analyze with conventional systems.

For companies wanting an enterprise-ready Hadoop-based solution, IBM InfoSphere BigInsights is IBM’s powerful and versatile solution for managing and analyzing Internet-scale volumes of structured and unstructured data-at-rest. Built on the open source Apache Hadoop software framework, InfoSphere BigInsights enhances this technology to withstand the demands of your enterprise, adding administrative, workflow, provisioning, and security features, along with
sophisticated analytical capabilities including text analytics and IBM BigSheets for data exploration. The result is that you get a more developer- and user-friendly solution for complex, large-scale analytics.

IBM InfoSphere Streams enables continuous analysis of massive volumes of streaming data with sub-millisecond response times. The offering provides a highly scalable and agile infrastructure that can support a wide variety of both structured and unstructured data types. These capabilities help you improve your organization’s insights and decision-making, providing an opportunity to respond to events as they happen.

PowerLinux Big Data Solutions for Apache Hadoop, IBM InfoSphere BigInsights and InfoSphere Streams enable businesses to deliver analytics services faster by taking advantage of the workload-optimizing features of the POWER7 architecture that specifically apply to big data:

IBM was among the select companies that Forrester invited to participate in its 1Q12, Forrester Wave report, Enterprise Hadoop Solutions. Technologies evaluated were IBM InfoSphere BigInsights (IBM’s Hadoop-based offering), and IBM Netezza®. In this evaluation, IBM was placed in the leadership category of the Wave and achieved the highest possible score in both the Strategy and Market Presence segments. In the third segment, Current Offering, IBM received the second highest score.

The report by analyst James Kobielus states, “IBM has the deepest Hadoop platform and application portfolio.”

Run thousands of tasks in parallel with

- Eight high frequency cores per socket
- Four intelligent threads per core
- Larger on-chip cache (eDRAM)

Achieve massive scale-out flexibility with

- Choice of PowerLinux 7R1 one-socket server in a dense rack, 7R2 two-socket server in a dense rack or POWER7-based Flex System p24L compute node
- Internal or external storage options
- High speed, low latency interconnect
- Comparable pricing to x86 rack/compute node

Exploit extreme memory bandwidth

- 2x the bandwidth of other commercially available systems at 500 GB per chip

Workload accelerators designed for key big data functions are planned² to further accelerate data compression and text analytics.

IBM PowerLinux Big Data Solution for Apache Hadoop

Apache Hadoop is emerging as the preferred way of handling Internet-scale structured and unstructured information at rest. With the explosion of data generated by social media sources, 80 to 85 percent of the world’s data is now unstructured (text, audio, video, click streams, log files, etc.)
“Big data and scalability go hand-in-hand. Big data solutions must be capable of handling Terabytes to Petabytes of data for real-time, deep analytical and complex workloads. IBM PowerLinux Big Data Solutions provide the best performance and total cost of ownership to meet the growing needs of our big data customers.”

—Anjul Bhambhri, Vice President Big Data, IBM Information Management Software

IBM Watson was based on the distributed Hadoop architecture, running on Linux and built on 90 commercially available POWER7 servers with 15 terabytes of RAM and 2,870 processor cores. Two-socket PowerLinux 7R2 servers are designed to scale out, complementing the distributed architecture of the Hadoop framework and file system. Based on POWER7 processors, IBM PowerLinux servers with Apache Hadoop help businesses apply Watson-like solutions to real-life analytics projects of any scale. The PowerLinux 7R2 server is the basis for the IBM PowerLinux Big Data Solution for Apache Hadoop, providing an innovation platform for mining structured and unstructured data.

IBM PowerLinux Big Data Solution for InfoSphere BigInsights
IBM has embraced the open source movement because we believe Apache Hadoop offers a way for organizations to manage and analyze massive volumes of data. For customers wanting an out-of-the-box, enterprise-ready Hadoop-based solution, IBM InfoSphere BigInsights enhances Hadoop with performance, reliability, security and administrative features, including a sophisticated text analytics module and IBM BigSheets for data exploration.

The IBM Big Data Solution for InfoSphere BigInsights analytics platform brings the power of Hadoop to the enterprise. Complementary IBM innovations deliver massive scale-out data processing and analysis with built-in resiliency and fault tolerance. Organizations can run large-scale, distributed analytics jobs on clusters of cost-effective PowerLinux server hardware. Designed to store data at rest, this solution leverages Hadoop’s MapReduce framework to tackle very large data sets by breaking up the data across many nodes and coordinating data processing across a massively parallel environment. The Hadoop environment with MapReduce and additional open source technologies such as Hive and Pig are integrated and delivered with this robust solution.

The IBM PowerLinux Big Data Solution for InfoSphere BigInsights is optimized to withstand the demands of your enterprise, adding administrative, workflow, provisioning, and security features, along with best-in-class analytical capabilities from IBM Research. The result is a more developer- and user-friendly solution for complex, large scale analytics, allowing enterprises of all sizes to manage the massive volume, variety and velocity of data that consumers and businesses create every day, cost-effectively turning data into insight.

IBM PowerLinux Big Data Solution for InfoSphere Streams
The IBM PowerLinux Big Data Solution for InfoSphere Streams enables continuous analysis of massive volumes of streaming data with sub-millisecond response times. Stream computing changes when, where, and how much of your business data you can analyze. Designed to manage stream flows and apply various analytics—mining, mathematical, video, etc.—against that streaming data, companies can store less, analyze more, and make better decisions, faster. By extracting insight from data as it is streaming into your organization, you can react to events as they are happening to change business outcomes. For example, financial institutions can inspect real-time credit card usage to detect and prevent fraudulent transactions.
The IBM PowerLinux Big Data Solution for InfoSphere Streams performs complex real-time analytics on data-in-motion. Enterprises extend their data warehouse data mining solution to score data in real-time as it flows into the organization—and scale to any size PowerLinux cluster. This streaming analytics solution benefits from the high memory bandwidth of the POWER7 platform. The deep integration of this analytics workload on PowerLinux enables businesses to run thousands of tasks in parallel to deliver real-time analytics services.

Exploiting the data explosion to analyze streaming data with multiple data types and respond to millions of events per second as they happen enables smarter companies to drive innovation and set themselves apart in the crowded marketplace.

Whether your business requires real-time analytics on streaming data, or uncovering insights from historical data, or both, IBM PowerLinux Big Data solutions help answer those business questions that have previously been beyond reach.

**Why IBM?**

Forward-thinking organizations are transforming their IT infrastructures to enable themselves to more effectively apply analytics to achieve better business outcomes. The same deep analytics capabilities that have been demonstrated by Watson can be applied to countless businesses and industries. Turning to IBM software, based on industry-standard Linux and open source applications, running on workload-optimized systems can improve overall business performance.

IBM PowerLinux Big Data Solutions are deeply optimized systems, from the hardware, to the Linux operating system and software applications, through IBM Software offerings. IBM is in a unique position to leverage two decades of Linux and open source experience, as well as IBM Research technologies developed for Watson, and hardware expertise from the microprocessor up. We believe that 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.³

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 IBM³ developers involved in over 100 open source projects and thousands of dedicated development and support personnel supporting all of IBM’s products and customers on Linux. We not only support this community, we’re part of it.

IBM has completed tens of thousands of Linux customer engagements, and facilitated thousands of migrations to Linux. We offer the widest range of hardware, middleware, and services products for Linux in the industry. We support Linux on all modern IBM Systems. And we back PowerLinux with a full line of implementation, support and migration services.
For more information
To learn more about IBM PowerLinux Big Data Solutions, please contact your IBM representative or IBM Business Partner, or visit the following website:
ibm.com/power/powerlinux

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. We'll partner with credit-qualified clients to customize an IT financing solution to suit your business goals, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward. For more information, visit: ibm.com/financing

© Copyright IBM Corporation 2012
Systems and Technology Group
Route 100
Somers, NY 10589
Produced in the United States of America
July 2012

IBM, the IBM logo, ibm.com, Power Systems, PowerLinux, POWER7, IBM Watson, and InfoSphere are 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

Netezza is a registered trademark of IBM International Group B.V., an IBM Company.

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

PowerLinux uses the registered trademark Linux pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the Linux mark on a world-wide basis.

Apache Hadoop is a trademark of Apache Software Foundation.

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


2 IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion. Information regarding potential future products is intended to outline our general product direction and should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.


4 http://go.linuxfoundation.org/e/6342/ho-writes-linux-2012/879kf/221839922