Intel IT Center

JUNE 2012

Big Data 101: Unstructured Data Analytics

A Crash Course on the IT Landscape for Big Data and Emerging Technologies

What's the buzz about big data and unstructured data analytics really about? Should you be worried about it? This brief provides you with a crash course on big data: why it matters, the impact on IT, emerging technologies for unstructured data analytics, and how Intel can help.

Why Big Data Matters

Data is exploding at an astounding rate. While it took from the dawn of civilization to 2003 to create 5 exabytes of information, we now create that same volume in just two days![1] By 2012, the digital universe of data will grow to 2.72 zettabytes (ZB) and will double every two years to reach 8 ZB by 2015. For perspective: That's the equivalent of 18 million Libraries of Congress.[2] Billions of connected devices—ranging from PCs and smartphones to sensor devices such as RFID readers and traffic cams—generate this flood of complex structured and unstructured data.

Big data refers to huge data sets characterized by larger *volumes* (by orders of magnitude) and greater *variety* and complexity, generated at a higher *velocity* than your organization has faced before. These three key characteristics are sometimes described as the three Vs of big data.

Unstructured data is heterogeneous and variable in nature and comes in many formats, including text, document, image, video, and more. Unstructured data is growing faster than structured data. According to a 2011 IDC study,[3] it will account for 90 percent of all data created in the next decade. As a new, relatively untapped source of insight, unstructured data analytics can reveal important interrelationships that were previously difficult or impossible to determine.

Big data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights into customers, partners, and the business—and ultimately gaining competitive advantage. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly, and jump on new business opportunities.

Impact of Big Data on IT

Big data is a disruptive force presenting opportunities as well as challenges to IT organizations. To realize its full potential to the organization, big data analytics requires a new approach to capturing, storing, and analyzing data.

The three Vs characterize what big data is all about, but also define the major issues IT needs to address:

- **Volume.** The massive scale and growth of unstructured data outpace traditional storage and analytical solutions.
- **Variety.** Big data is collected from new sources that haven't been mined for insight in the past. Traditional data management processes can't cope with the heterogeneity and variable nature of big data, which comes in formats as different as e-mail, social media, video, images, blogs, and sensor data—as well as "shadow data" such as access journals and Web search histories.
- **Velocity.** Data is generated in real time with demands for usable information to be served up as needed.

The confluence of the three Vs drives a fourth: **Value**. For any enterprise to succeed in driving value from big data, volume, variety, and velocity have to be addressed in parallel. Partial credit is not an option.

Infrastructure Challenges

Emerging technologies such as Hadoop* and MapReduce are designed to address the three Vs of big data. They also put significant demands on infrastructure to support the distributed processing of unstructured data analytics, including requirements for:

- Infrastructure that is built for large-scale, distributed, data-intensive jobs that spread the problem across clusters of server nodes
- Storage that is efficient and cost-effective enough to capture and store terabytes, if not petabytes, of data, with intelligent capabilities to reduce your data footprint such as data compression, automatic data tiering, and data deduplication
- Network infrastructure that can quickly import large data sets and then replicate it to various nodes for processing
- Security capabilities that protect highly-distributed infrastructure and data
- Human resource skill sets needed for identifying opportunities through the use of statistics, algorithms, mining, and visualization

The Rise of the Data Scientist

Finding skilled personnel is one of the major challenges associated with big data analytics. Successful big data analytics initiatives involve close collaboration between IT, business users, and "data scientists" to identify and implement the analytics that will solve the right business problems. Data science is an emerging field, and data scientists are a new kind of professional with a unique skill set. Data scientists are responsible for modeling complex business problems, discovering business insights, and identifying opportunities. Demand is high for people who can help make sense of the massive streams of digital information pouring into organizations.

Emerging Technologies for Big Data Analytics

New technologies are emerging to make unstructured data analytics possible and cost-effective. The new approach redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources. It utilizes easily scalable "shared nothing" architecture, distributed processing frameworks, and nonrelational and parallel relational databases.

Shared-nothing architecture is stateless, with no nodes sharing memory or disk storage, and is made possible because of the convergence of advances in hardware, data management, and analytic applications technologies.

- **Hardware architecture.** Clusters of commodity servers, such as those based on Intel® Xeon® processors, provide the compute power and speed for massively parallel processing across a distributed grid.
- **Analytics applications architecture.** New data processing systems make the computing grid work by managing and pushing the data out to individual nodes, sending instructions to the networked servers to work in parallel, collecting individual results, and then reassembling them to produce meaningful results. Processing the data where it resides is faster and more efficient than first transporting it to a centralized system.
- **Data architecture.** To handle the variety and complexity of unstructured data, databases are shifting from relational to nonrelational. Unlike the orderly world of relational databases, which are structured, normalized, and densely populated, nonrelational databases are scalable, network oriented, semistructured, and sparsely populated. NoSQL database solutions do not require fixed table schemas, avoid join operations, and scale horizontally.

Distributed Frameworks: The Emergence of Apache* Hadoop*

Apache* Hadoop is evolving as the best new approach to unstructured data analytics. Hadoop is an open-source framework that uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management. In addition to offering high availability, Hadoop is more cost-effective for handling large unstructured data sets than conventional approaches, and it offers massive scalability and speed.

As more and more enterprises recognize the value and advantages associated with big data insights, adoption of Hadoop is growing. Apache released the first full production version of Apache Hadoop 1.0 in January 2012. For more about Hadoop deployments, see the *Intel® Cloud Builders Guide to Cloud Design and Deployment on Intel® Platforms: Apache* Hadoop**.

The Hadoop Ecosystem

Commercial versions of Hadoop are also on the rise. The Hadoop ecosystem is a complex landscape of vendors and solutions that includes established players and several newcomers. Numerous vendors offer their own Hadoop distribution, packaging the basic stack with other Hadoop projects such as Hive*, Pig*, and Chukwa*. Some of these distributions can integrate with data warehouses, databases, and other data management products, enabling the analytics engine to access and query data across multiple sources.

Hadoop Infrastructure: Big Data Storage and Networking

Hadoop clusters are made possible by dramatic improvements in mainstream compute and storage resources and are complemented by 10 gigabit Ethernet (10 GbE) solutions. The increased bandwidth associated with 10 GbE is key to importing and replicating the large data sets across servers. Intel® Ethernet 10 Gigabit Converged Network Adapters provide high-throughput connections, and Intel SATA Solid State Drives are high-performance, high-throughput hard drives for raw storage. To enhance efficiency, storage needs to support advanced capabilities such as compression, encryption, automated tiering of data, data deduplication, erasure coding, and thin provisioning—all of which are supported with the Intel Xeon® processor E5 family today.

What about Big Data and Cloud?

As a result of cloud computing, organizations now have access to large grids of commodity computers within their own data centers of networked servers and in public cloud infrastructure services such as Amazon* Web Services. In the era of big data, the cloud offers a potential self-service consumption model for data analytics. Both cloud computing and big data analytics are extensions of virtualization technologies and grid computing models, making the cloud an agile data platform to support the business at a cost significantly less than traditional data platforms. Hadoop is fast evolving as the de facto framework for big data in the cloud.

How Intel Can Help

Intel, the company that makes the technologies that underpin your data center infrastructure—servers, networking, storage, databases, and data warehouses—can help make big data analytics work for you by:

- Delivering optimized technology built to scale for big data analytics projects
- Helping you advance your new big data analytics projects faster
- Addressing tomorrow's challenges with a vision for distributed analytics on the edge

Intel Resources to Learn More

The Intel IT Center provides straightforward, fluff-free, unbiased information that addresses each of the ways Intel can help IT pros implement strategic projects like big data analytics. For planning guides, peer research, real-world customer references, vendor spotlights, and live events about big data analytics, visit intel.com/bigdata.

1 "Google Chief Eric Schmidt on the Data Explosion." *I-Global Intelligence for the CIO* (August 4, 2010). www.i-cio.com/features/august-2010/eric-schmidt-exabytes-of-data

2 "Big Data Infographic and Gartner 2012 Top 10 Strategic Tech Trends." *Business Analytics 3.0* (blog) (November 11, 2011). practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and-gartner-2012-top-10-strategic-tech-trends/

3 "Extracting Value from Chaos." *IDC IView*, EMC Corporation (June 2011). www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Share with Colleagues

0612/RF/ME/PDF-USA 327439-001

Sponsors of Tomorrow.™