White Paper

Big Data Analytics

Turn Big Data into Big Value

A Practical Strategy

ABSTRACT

Some of today's most successful companies achieve game-changing business advantages by capturing, analyzing, and acting upon vast amounts of diverse, fast-moving "big data." This paper describes three usage models that can help you implement a flexible and efficient big data infrastructure to realize competitive advantages in your own business. It also describes Intel innovations in silicon, systems, and software that can help you to deploy these and other big data solutions with optimal performance, cost, and energy efficiency.

The Big Data Opportunity

Big data is often compared to a tsunami. Today's five billion cell phone users and nearly one billion Facebook* and Skype* users generate unprecedented volumes of data, and they represent only part of the global online population. Intel estimates that more than 1,500 exabytes (EB) of data—1,500 billion gigabytes—flowed through the cloud in 2012. To put that in perspective, the total number of words spoken in all of human history is estimated at about 5 EB.

Nor have the flood waters of big data begun to level out. We are moving quickly toward the "Internet of things," in which vast numbers of networked sensors in businesses, homes, cars, and public places drive data generation to almost unimaginable levels (Figure 1). Yet comparing big data with a tsunami misses the most important point.

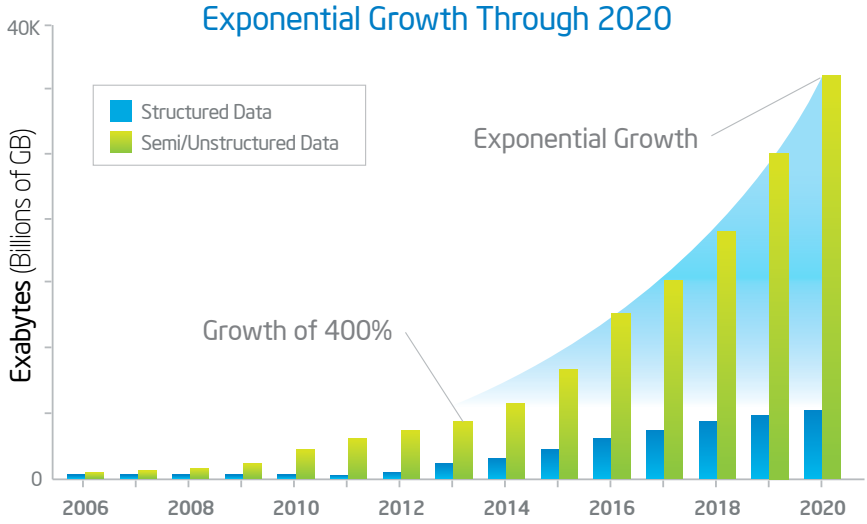

Figure 1. Current and forecasted growth of big data. Source: Philippe Botteri of Accel Partners, Feb. 2013.

While a tsunami is destructive, big data holds tremendous potential value. With the right tools and strategies, businesses can extract insights that deliver game-changing competitive advantages. A number of public and private organizations do that today.

- **Retailers** analyze social media trends in real time to offer the hottest products to the most likely buyers, and they do this at volumes and with levels of granularity that have never before been possible.
- **Financial firms** analyze credit card transactions, bill payments, and bank account activity to detect and prevent fraud in real time and to improve the recovery of lost funds.
- **Content providers** analyze subscriber selections in real time, so they can recommend related content and offer new products and services in ways that improve both revenue and customer satisfaction.
- **Cities** use big data to predict and alleviate traffic congestion, and to stave off costly road expansions.
- **Utilities** load-balance their energy grids through real-time monitoring of energy usage, so they can transmit power more efficiently and reliably and put off major infrastructure build outs.

Using big data to achieve these kinds of benefits requires new approaches to data management. Big data differs from traditional business information. Although transactional data comprises a portion of it, big data is multi-structured, fast moving, and may come in greater amounts than your infrastructure can handle.

- Big data sets dwarf traditional business data—petabytes instead of terabytes.
- Big data includes structured and unstructured content in many different formats, such as e-mail, social media, video, images, blogs, sensor data, "shadow data" such as access journals and Web search histories, and many other types.
- Big data is generated constantly, and instantaneous insights can improve outcomes in real-time business scenarios. Although batch analytics can still be valuable, on demand queries from live or streaming data offer game-changing potential.

Because the value of big data stretches across vast amounts of complex, fast-moving content, deriving meaningful insights often requires extensive mining and deep analysis that go beyond traditional Business Intelligence (BI) queries and reports. Machine learning, statistical modeling, graph algorithms and other emerging techniques can unveil valuable, actionable insights that deliver significant competitive advantages.

Extracting Business Value from Big Data

This paper explores three usage models to extract value from big data. They apply across a wide variety of organizations. Each usage model builds on the previous one to deliver increasing value.

- **Usage Model 1—Extract, Transform, and Load (ETL).** Before you can analyze data, you must aggregate, pre-process, and store it in a database. ETL does that, but big data can overwhelm traditional ETL tools and strategies. Apache Hadoop* offers a cost-effective answer to that challenge.
- **Usage Model 2—Interactive Queries.** Recent technology innovations dramatically increase the performance and scalability of traditional data warehousing models. With these improvements, real-time analytics can operate on much larger and more varied data sets to extend the value of existing BI investments and to integrate more effectively with new big data solutions such as Hadoop.
- **Usage Model 3—Predictive Analytics.** New analytic techniques go beyond data mining and visualization to determine not only what has happened and why, but to predict what is likely to happen based on all available data, including real-time streams from external sources. This last usage model builds on the two previous ones to create a more unified and extensible analytics environment.

Usage Model 1—ETL using Apache Hadoop*

Like traditional data, big data must be extracted from external sources, transformed into structures that fit operational needs, and loaded into a database for storage and management. Traditional ETL solutions cannot handle the demands of poly-structured data, so Hadoop software has emerged as the de facto platform for addressing this need (Figure 2).

The distributed storage and processing environment of a Hadoop cluster works well for big data ETL. Hadoop breaks-up incoming streams into pieces and applies simple operations in parallel to rapidly process large amounts of data. It supports all data types and can operate across tens, hundreds, or even thousands of servers to provide massive scalability. The Hadoop Distributed File System (HDFS) stores the results on low cost storage devices directly attached to each server in the cluster—ready for immediate up-loading to the enterprise data warehouse or unstructured data stores.

Hadoop can process poly-structured data for analysis, even when that data is not predefined. In other words, Hadoop supports a Schema on Read model as opposed to the Schema on Write model used in traditional ETL processes. This enables Hadoop to load large amounts of data in a short time, making that data quickly available for analysis, visualization, and other uses.

Infrastructure Considerations

Dual-socket servers based on the Intel® Xeon® processor E5 family provide an optimal balance of capability versus cost for most Hadoop deployments. These servers offer more cores, cache, and memory capacity than previous generation servers. They also provide up to twice the I/O bandwidth with 30 percent lower I/O latency.[1] These resources sustain high throughput for larger numbers of data-intensive tasks executing in parallel.

Lightweight, I/O-bound workloads, such as simple data sorting operations, may not require the full processing power of the Intel Xeon processor E5 family. Such workloads run economically on high-density, low-power servers based on the Intel® Xeon® processor E3 family or the Intel® Atom™ processor-based System on a Chip (Intel Atom SoC). With power envelopes as low as 6 watts, the 64-bit x86-based Intel Atom SoC provides unprecedented density and energy-efficiency in a server-class processor.

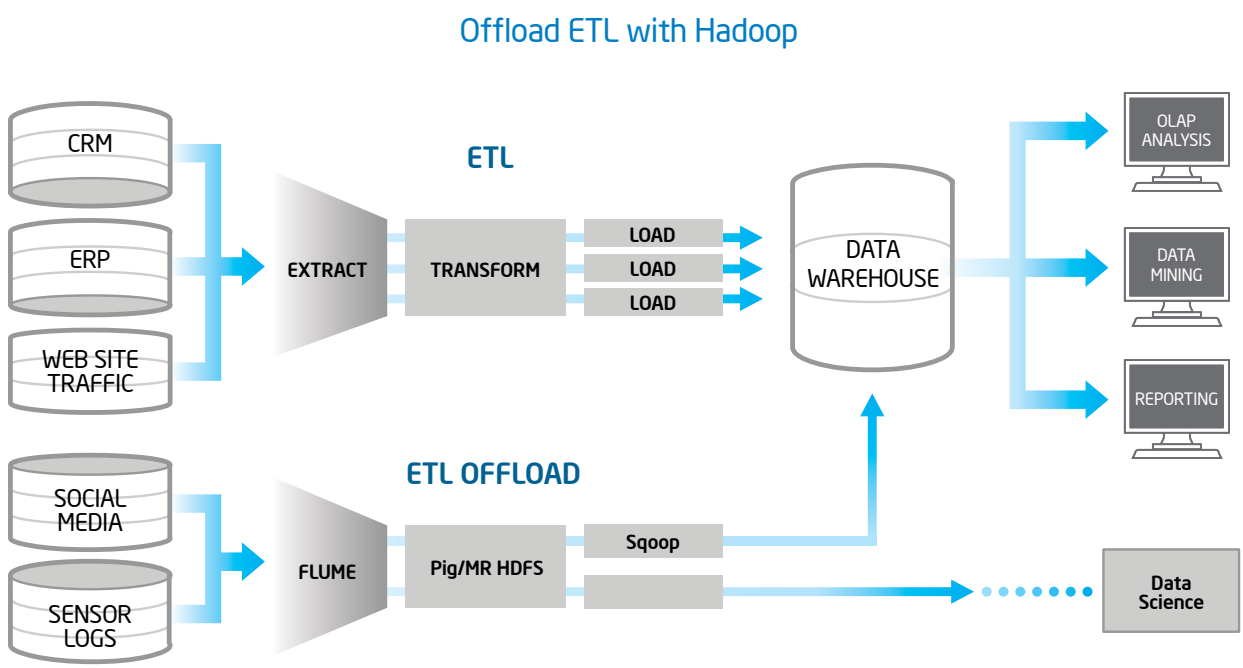

Figure 2. Using Apache Hadoop,* organizations can ingest, process, and export massive amounts of diverse data at scale.

All servers in a Hadoop cluster require substantial memory and a relatively large number of storage drives to meet the demands of data-intensive Hadoop workloads. Sufficient memory is required to support high throughput for the many operations performed in parallel. Multiple storage drives (two or more per core) deliver the aggregate I/O throughput needed to avoid storage bottlenecks. Storage performance improves considerably with at least one Intel® Solid State Drive (Intel® SSD) in each server node.

By processing data near where it is stored, Hadoop greatly reduces the need for high-volume data movement. Nevertheless, fast data import and export requires sufficient network bandwidth. In most cases, each rack of servers should use a 10 Gigabit Ethernet (10 GbE) switch and each rack-level switch should connect to a 40 GbE cluster-level switch. As data volumes, workloads, and clusters grow, it may be necessary to interconnect multiple cluster-level switches or even to uplink to another level of switching infrastructure.

For more detailed information, see the Intel white paper: "Extract, Transform & Load (ETL) Big Data with Apache Hadoop*" posted in the Intel Developer Zone at software.intel.com.

Usage Model 2—Interactive Queries

A data warehouse provides a central repository for business data and BI functions, such as online analytical processing (OLAP) and data visualization. New and historical data is gathered from disparate sources and prepared for interactive queries and other types of analysis.

Big data can overwhelm traditional data warehouse capabilities and resources. Vendors have responded with a variety of advances in performance and scalability. For example:

- **In-memory databases** eliminate the latencies and overhead associated with shuttling data back and forth between servers and storage systems. This approach reduces data access times from milliseconds to nanoseconds, which practically eliminates a bottleneck that impeded database performance for decades. Oracle TimesTen,* SAP HANA,* Microsoft IMDB,* IBM solidDB,* VMware vFabric SQLFire,* and a number of open source solutions use this strategy to speed-up the processing and management of incoming data streams.
- **Data warehouse appliances** combine servers, storage, operating systems, database management systems, and supporting components into pre-built, highly-optimized, turnkey systems to simplify integration and improve performance. Many data warehouse appliances support in-memory databases, and some include proprietary data filtering technologies that accelerate data flow. Most of these appliances come as large-scale, symmetric multi-processor (SMP) systems or massively-parallel processing (MPP), extensible blade systems. Examples include IBM Netezza,* HP EDW Appliance,* Oracle Exadata,* Teradata DW Appliance,* Dell Parallel DW,* and the Pivotal (formerly EMC Greenplum) Data Computing Appliance.*

Businesses looking to implement a powerful and cost-effective big data platform should consider combining a large-scale SQL data warehouse with a Hadoop cluster. The cluster can quickly ingest and process large, diverse, and fast-moving data streams. Appropriate data sets can then be loaded into the data warehouse for ad hoc SQL queries, analysis, and reports. Users also can query multi-structured data sets that reside in the Hadoop cluster using software such as Apache HBase,* Spark,* Shark,* SAP HANA,* Apache Cassandra,* MongoDB,* Tao,* Neo4J,* Apache Drill,* or Impala.* This hybrid strategy offers a foundation for faster, deeper insights than either solution alone can achieve.

Similar processes apply whether you use a traditional data warehouse or a more modern system designed for larger volumes and faster data streams: gather data from external sources then cleanse and format the data to fit into the warehouse data model. This can be done prior to loading the data into the warehouse or it can be done on-the-fly as streaming data sources are fed into the warehouse.

With the data loaded, analysis can begin. Modern data warehouses support ad hoc queries, enabling access on-demand for data with any meaningful combination of values. This contrasts with more-traditional data warehouses that generate only pre-defined reports based on known relationships.

Infrastructure Considerations

Whether you integrate your own SQL data warehouse solution or evaluate appliances, the following considerations can help you optimize scalability, reliability, and total cost of ownership (TCO).

The complex analytics performed in SQL data warehouses do not typically scale well across large numbers of clustered nodes, so individual data warehouse servers must deliver high-performance and scalability. Four-socket, eight-socket, and larger servers based on the Intel® Xeon® processor E7 family provide the scalable performance needed to handle demanding analytics workloads. For example, enterprise data warehouse appliances, including large-scale SMP and blade-based MPP systems, use these processors to maximize overall performance and throughput. Each Intel Xeon E7 processor provides up to 10 cores, 20 threads, and 30 MB of last-level cache. These processors also support DIMMs as large as 32 GB—up to 4 terabytes of total memory in an eight-socket server—so they can host very large in-memory databases.

Since a data warehouse typically runs on a single server, system uptime is particularly important. The Intel Xeon processor E7 family includes advanced reliability, availability, and serviceability (RAS) features built into the silicon to support mission-critical levels of reliability and to protect data. All key interconnects, data stores, data paths, and subsystems integrate active and passive error monitoring.

Self-healing features proactively and reactively repair known errors and also reduce the likelihood of future errors by acting automatically based on configurable error thresholds. Intel works extensively with hardware, operating system, virtual machine monitor (VMM), and application vendors to help ensure tight integration throughout the hardware and software stack.

As data volumes skyrocket, new strategies help scale data storage capacity more efficiently and cost-effectively, both within and beyond the data warehouse. The following strategies can work together to meet diverse needs at lower total cost.

- **Scale-out storage architectures** deliver affordable high capacity and support federation across private and hybrid clouds. These solutions scale dynamically, and you can provision them faster than traditional storage systems. They also help to improve data management efficiency.
- **Low-latency, proximity storage** is a good fit for data-intensive applications that perform better when co-located with the data storage devices. Examples include business processes, decision support analyses, and high-performance computing workloads, as well as collaborative processes, applications, and web infrastructure running on virtualized servers.
- **Centralized storage** aggregated as logical pools in storage area networks (SANs) support high-performance business databases. When optimized for affordable capacity rather than high performance, centralized solutions provide efficient storage for backup, archive, and object store requirements.

Higher storage efficiency can help to contain costs in the face of rapid data growth. Many storage vendors integrate Intel Xeon processors into their storage solutions to support advanced data management functions that help to improve efficiency. According to IDC's June 2013 Worldwide Storage and Virtualized x86 Environments 2013–2017 Forecast, about 80 percent of worldwide, enterprise-class storage solutions for corporations, cloud, and HPC run on Intel architecture. Look for storage platforms that support data-efficiency technologies, including:

- **Data de-duplication** to conserve capacity.
- **Data compression** to increase throughput.
- **Thin provisioning** to improve utilization, by enabling storage to be provisioned on demand, instead of overprovisioning capacity based on projected needs.
- **Intelligent tiering** to optimize performance versus cost, by automatically moving "hot" data to faster storage devices and "cold" data to higher capacity, lower cost drives. With this approach, a small number of high-speed drives, such as Intel® SSD 710 Series SATAs, can deliver substantial performance improvements at relatively low cost.

Loading data sets into data warehouses quickly and efficiently enables analytics applications to provide business insights in a timely manner. Efficient ETL processing is one component of the solution. Another is a fast and efficient network to drive the growing business value of analytics throughout the enterprise. Intel® Ethernet products integrate technologies to address these requirements.

- **Near-native performance in virtualized environments.** Virtualization improves infrastructure flexibility and utilization—important for containing costs as big data solutions grow. Intel® Virtualization Technology for connectivity (Intel® VT-c) helps to reduce I/O bottlenecks and improve overall server performance in virtualized environments. Its Virtual Machine Device Queues (VMDQ) technology offloads traffic sorting and routing to dedicated silicon in the network adapter. Its PCI-SIG Single Root I/O Virtualization (SR-IOV) technology allows a single Intel® Ethernet Server Adapter port to support multiple, isolated connections to virtual machines.
- **Unified 10 GbE networking.** Consolidating data center traffic onto a single, high-bandwidth network helps to reduce cost and complexity and provides the performance and scalability needed to address rapidly growing needs. Intel Ethernet Converged Network Adapters support fiber channel over Ethernet (FCoE) and iSCSI to simplify implementation and reduce costs when consolidating local area network (LAN) traffic and storage area network (SAN) traffic.
- **Simpler, faster connections to iSCSI SANs.** Intel Ethernet Converged Network Adapters and Intel Ethernet Server Adapters provide hardware-based iSCSI acceleration to improve performance. They also take advantage of native iSCSI initiators integrated into leading operating systems to simplify iSCSI deployment and configuration in both native and virtualized networks.

For more detailed information, see the Intel SQL Data Warehousing Usage Model white paper.

Usage Model 3—Predictive Analytics on the Hadoop Platform

Predictive analytics extracts higher value from data by capturing relationships from past events and using them to predict future outcomes (Figure 3). Retailers use predictive analytics to deliver more compelling offers to individual customers, healthcare organizations use it to select best-fit treatment protocols, and financial services organizations use it to increase investment returns and reduce risk.

Although predictive analytics can aid in strategic business planning, its greatest value may come from tactical guidance at the point of decision and operational guidance at the point of execution. Centralized teams of data scientists, database administrators, and software developers work together to provide customized solutions for the most critical business operations. As businesses integrate this capability more widely into their operations, they must provide optimized decision tools for a wider range of users and automated systems.

Predictive analysis falls into two main categories: regression and machine learning.

- **Regression techniques** compare current data with historical models to forecast the most probable outcome.
- **Machine learning** uses artificial intelligence with little or no human intervention. The system analyzes a representational data set to extract relationships, and it generalizes from that to make predictions based on new data. Optical character recognition (OCR) is a classic example, but new applications exploit big data across a wide range of scenarios.

Intel IT began its own trailblazing big data analytics effort in 2010, and recommends combining the two usage models already discussed in this paper to create a hybrid analytics infrastructure (Figure 4).

1. **Deploy a data warehouse appliance** based on an MPP architecture to perform complex predictive analytics quickly on large data sets. A number of vendors have incorporated the Intel Xeon processor E7 family into blade-based appliances that deliver the required performance at relatively low cost. These systems fit into existing enterprise BI solutions and provide integrated support for advanced analytics tools and applications, such as R, an open-source statistical computing language that is popular among data scientists.
2. **Add a Hadoop cluster** for fast, scalable, and affordable ETL for the data warehouse. Hadoop also runs other data processing and analytics functions that perform well in a distributed processing environment. The Hadoop ecosystem offers a growing variety of tools and components to address these needs.

Infrastructure Considerations

To provide maximum flexibility, the data warehouse and the Hadoop cluster should use a high-speed data loader and link together using 10 GbE or another high-bandwidth networking technology. This allows you to move data quickly between the two environments, so you can use the most effective analytics techniques based on specific data types, workloads, and business needs.

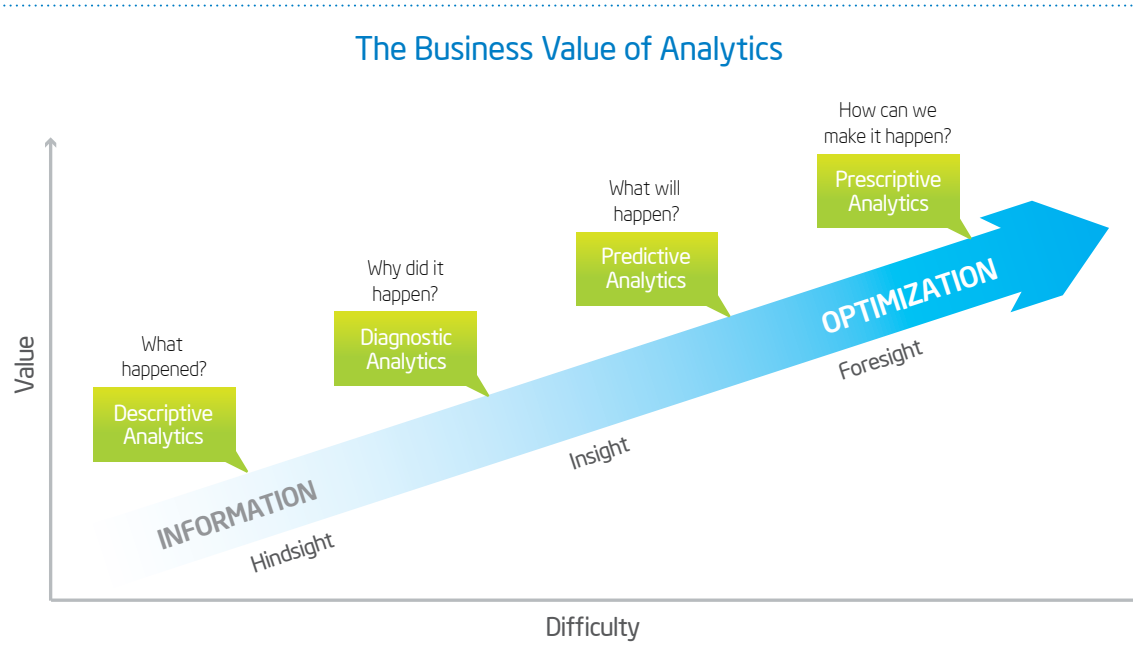

Figure 3. According to Gartner, the difficulty and business value of analytics both increase as the focus moves from hindsight to foresight.

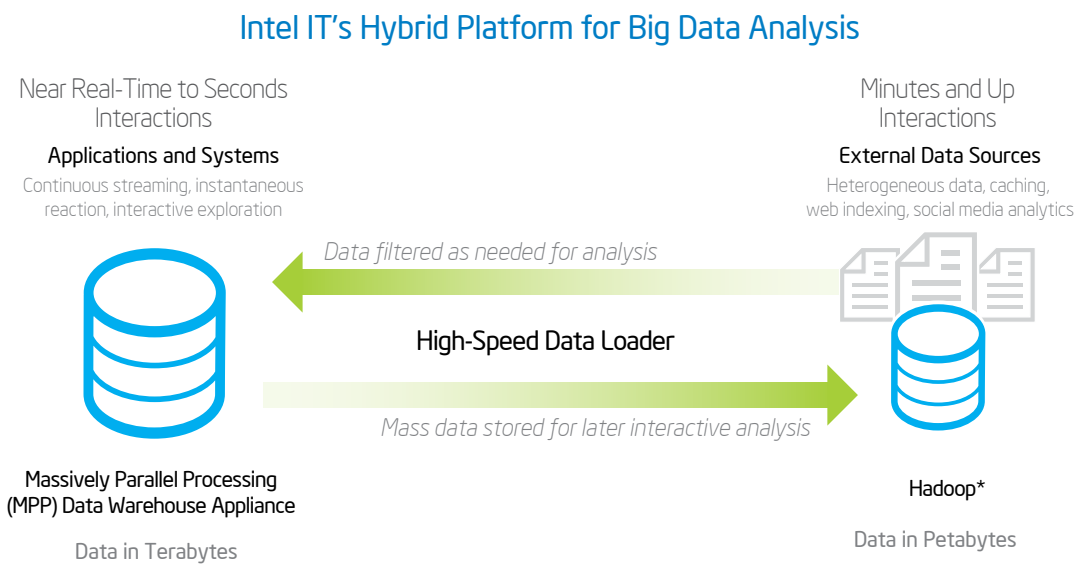

Figure 4. Intel IT's big data platform provides a flexible foundation for analytics—including predictive analytics—by using a high-speed data loader to connect a massively parallel processing (MPP) data warehouse appliance with clusters of industry-standard servers running Hadoop software.

Creating a Better Foundation for Big Data Analytics

As big data technologies and solutions advance, Intel products and technologies help speed-up innovation throughout the ecosystem. By working with hardware, software and service providers to ensure broad support, Intel helps businesses integrate these new capabilities more simply and affordably on a standards-based, connected, managed, and secure architecture.

Processor Advances for Performance and Security

Intel processor advances deliver increasing performance and value for next-generation big data solutions. Ongoing improvements in per-thread performance, parallel execution, I/O throughput, memory capacity, and energy efficiency help businesses address rapidly growing needs using affordable, mainstream computing systems.

Intel also integrates advanced security technologies that protect data more effectively, so you can integrate sensitive data into your big data analytics environment. Current security technologies in Intel Xeon processors provide the following advantages.

- **Strong workload isolation on trusted infrastructure**. Intel® Trusted Execution Technology (Intel® TXT) and Intel® Virtualization Technology (Intel® VT) help to protect systems and software more effectively in virtualized and cloud environments. Intel VT provides silicon-assisted workload isolation. Intel TXT can establish trusted infrastructure pools by ensuring that Intel® Xeon® processor-based servers boot only into "known good states."
- **Fast, low-overhead data encryption.** Intel® Advanced Encryption Standard New Instructions (Intel AES-NI) provides hardware acceleration for encryption to protect data in latency-sensitive analytics environments without sacrificing performance. Intel performance tests show that Intel AES-NI can accelerate encryption performance in a Hadoop cluster by up to 5.3x and decryption performance by up to 19.8x when used in combination with the Intel Distribution for Apache Hadoop software (Intel Distribution).[2] Intel Xeon processors and the upcoming Intel Atom SoC support Intel AES-NI.

New Tools and Optimized Software

Intel works both independently and in collaboration with leading software vendors and the open-source community to provide optimized software stacks and services for big data analytics. These efforts help to deliver new and advanced functionality throughout the big data ecosystem. They also help to ensure the best possible performance for big data applications running on Intel architecture.

Intel also delivers software products that help address some of the most critical needs within the big data ecosystem.

- **Performance benchmarking for Hadoop clusters and applications.** The Intel® HiBench suite includes 10 benchmarks that IT organizations and software vendors use to measure performance for specific, common tasks, such as sorting and word counting, and for more comprehensive real-world functions, such as web searching, machine learning, and data analytics. Intel engineers use the Intel HiBench suite to help with upstream Hadoop optimizations for Intel Architecture as well as with Java* optimizations for Hadoop.

IMPLEMENTING BIG DATA ANALYTICS

Intel is integrating predictive big data analytics into its existing business intelligence (BI) environment to help improve business efficiency and performance. A number of big data proof-of-concept deployments are underway in partnership with Intel business groups. Current focus areas include malware detection, chip design validation, market intelligence, and recommendation systems.

To learn about Intel IT strategies and best practices for implementing big data analytics, read the Intel IT white paper, "Mining Big Data in the Enterprise for Better Business Intelligence."

- **An enterprise-ready distribution of Hadoop.** The Intel Distribution provides the most up-to-date optimizations for Intel architecture in a software package that simplifies deployment and supports enterprise-class requirements for security and manageability. Many optimizations first go into the Intel Distribution and subsequently get submitted to the open-source Apache Hadoop project.
- **A fast, massively-scalable, distributed file system.** Intel® Luster storage software, an Intel optimized distribution of the Lustre* distributed file system, supports large-scale cluster computing. This software scales to support tens of thousands of client systems and tens of petabytes of storage. It delivers more than a terabyte per second of aggregate I/O throughput.

Advanced Power Management for Lower Operating Costs

Storing and analyzing big data requires substantial infrastructure build-outs for most organizations, and that requires managing energy consumption to contain total costs. The energy-efficiency of Intel Xeon processors and Intel Atom SoCs can help. No matter which you choose, software supports both Intel Xeon processor and Intel Atom processor families without recompilation to help you avoid the complexity of managing multiple architectures and code bases.

Intel offers tools to help you manage power consumption more effectively.

- **Efficient data center power management.** Intel Data Center Manager (Intel DCM) plugs into existing management consoles and takes advantage of built-in instrumentation in Intel processors to provide advanced power and thermal management, from individual servers and blades, to racks, rows, and entire data centers.
- **Integrated energy-management in Linux* environments.** The Running Average Power Limit (RAPL) Linux kernel software driver developed by Intel provides support for monitoring, managing, and capping power consumption for the Intel Xeon processor E5 family.

Conclusion

The ability to capture, store, and analyze data from all sources offers game-changing competitive advantage across a wide range of industries, yet the tsunami of big data introduces tough new infrastructure challenges. The three usage models presented in this paper provide a model that enterprises can use and adapt to turn big data into business value.

- Deploy Hadoop to ingest and prepare big data for analysis.
- Connect your Hadoop cluster to a fast, scalable data warehouse for interactive query capabilities using mixed data.
- Add predictive analytics and machine learning applications to make accurate predictions and act on them in real time.

Intel innovations in silicon and software provide optimizations and targeted functionality to help you implement these and other big data usage models more simply and effectively.

For more information visit these links on intel.com:

Big Data Intelligence | The Intel IT Center | Private Cloud Solutions

[1] Source: The claim of up to 32% reduction in I/O latency is based on Intel internal measurements of the average time for an I/O device read to local system memory under idle conditions for the Intel® Xeon® processor E5-2600 product family versus the Intel® Xeon® processor 5600 series. 8 GT/s and 128b/130b encoding in the PCIe 3.0 specification enables double the interconnect bandwidth over the PCIe 2.0 specification. For more information, read the PCI-SIG* press release, "PCI-SIG releases PCI Express 3.0 Specification."

[2] For details, see the Intel solution brief, "Fast, Low-Overhead Encryption for Apache Hadoop*". 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.

Printed in USA 0713/DF/HBD/PDF Please Recycle 329261-001US