**IT@Intel White Paper**

**Intel IT**
IT Best Practices
Business Intelligence
July 2012

# Mining Big Data in the Enterprise for Better Business Intelligence

The ability to mine and analyze big data gives organizations deeper and richer insights into business patterns and trends, helping drive operational efficiencies and competitive advantage in manufacturing, security, marketing, and IT.

## Executive Overview

**To improve Intel's business intelligence (BI), Intel IT is putting in place the systems and skills for analyzing big data. A major portion of this data consists of the large, unstructured data sets that constitute up to 90 percent of enterprise data. The ability to mine and analyze big data, in any form, from many sources, gives us deeper and richer insights into business patterns and trends, helping drive operational efficiencies and competitive advantage in manufacturing, security, marketing, and IT.**

Developing the organizational skill to mine and process big data to perform predictive and prescriptive analytics will be a key driver of performance in the future, enabling Intel to:

- Make better decisions
- Increase business velocity
- Accelerate the pace of innovation
- Discover and tap new markets

Advances in parallel computing now make it possible to handle big data, to the point where it is now becoming standard practice to capture and store information well before its value is completely understood, and tackle many business problems that have been previously too large to handle.

Intel IT is deploying a big data platform in 2012 in close partnership with Intel business groups in proofs of concept to demonstrate its utility in providing BI within the enterprise.

Our big data projects for 2012 include:

- Malware detection
- Chip design validation
- Market intelligence
- Recommendation system

Intel is in the early stages of developing the capacity for better BI through big data, though we anticipate fast growth in these capabilities across research and development, cybersecurity, design, manufacturing, operations, market development, and human resource management.

Moty Fania
Advanced Business Intelligence Solutions, Intel IT

John David Miller
Principal Engineer, Intel IT Labs

## Contents



## IT@INTEL

The IT@Intel program connects IT professionals around the world with their peers inside our organization – sharing lessons learned, methods and strategies. Our goal is simple: Share Intel IT best practices that create business value and make IT a competitive advantage. Visit us today at www.intel.com/IT or contact your local Intel representative if you'd like to learn more.

# BUSINESS CHALLENGE

**Worldwide, the amount of raw data is growing exponentially, due in part to the explosion of connected devices, Internet services, social media, cameras, sensors, and user-generated content. Moreover, up to 90 percent of corporate data, including documents, web pages, and email, is unstructured. The sheer volume and complexity of data is overwhelming typical database software, and this situation is calling for a new approach.**

## Drowning in Data, Starved for Knowledge

According to the report "Big data: The next frontier for innovation, competition, and productivity" from the McKinsey Global Institute, 15 out of 17 business sectors in the United States have more data stored per company than the U.S. Library of Congress.[1] Wal-Mart is a good example: The retail giant handles more than 1 million customer transactions every hour, importing this data into databases estimated to contain more than 2.5 petabytes (PBs) of data—equivalent to 167 times the information contained in all the books in the Library of Congress.

Most big data derives from the billions of transactions and other bits of information that enterprises such as Intel log every day about their customers, suppliers, and operations. Formerly considered chiefly a storage problem, big data is now becoming recognized as the newest strategic asset, a gold mine for actionable insights into every aspect of one's business.

We recognize two basic categories of use cases for big data:

- **Big databases.** Contain structured data that is simply too large for a traditional relational database management system (RDBMS) to handle.
- **Deep analytics.** Used to search for answers to complex, open-ended problems. Typically, such answers are not directly codified in the source data. Instead, big data visualization and analytics tools help gain valuable insights through successive refinement and abstraction.

In the past, most companies could only try to aggregate the data for analysis, or take samples and try to extrapolate meaning from them. This is still the status quo. Gartner predicts that "through 2015, more than 85 percent of Fortune 500 organizations will fail to effectively exploit big data for competitive advantage."[2] Leading-edge organizations though are already implementing big data analysis capabilities and could see significant results. According to Gartner, these companies are moving fast to update business intelligence, data mining, and business analytics practices with the new tools and skill sets that big data offer.

Professor Eric Brynjolfsson, director of MIT's Center for Digital Business Research, conducted research on 179 large publicly traded firms and found that companies that use "data-driven decision making" are about 5 percent more productive and profitable than their competitors. He concluded that "there is a lot of low-hanging fruit for companies that are able to use big data to their advantage."[3]

[1] McKinsey Global Institute, May 2011. www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[2] "Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond," Gartner press release, December 1, 2011. www.gartner.com/it/page.jsp?id=1862714

[3] MIT Sloan Experts: Commentary on today's business issues, February 14, 2012. www.mitsloanexperts.com/2012/02/15/erik-brynjolfsson-on-big-data-a-revolution-in-decision-making-improves-productivity

## The Need for Trailblazing

Today's IT managers are being challenged to develop systems to analyze big data and help their companies make intelligent decisions based on it. Because big data is a new field, its practitioners and best practices are not well established. Companies that offer big data training, consulting, and other services will be able to help fill the void.

About two years ago, we began to look into how to tap our big data. Intel's big data efforts represent a key element in our overall roadmap for transforming Intel's business with advanced analytics.

# SOLUTION

**Intel IT is developing several big data proofs of concept to prove the effectiveness of big data in solving high-value business problems.**

Based on our research and industry best-known methods, we began in 2012 to implement an internal, hybrid, and cost-effective big data platform based on the following:

- An SQL-based massively parallel processing (MPP) data warehouse appliance
- Hadoop*, for distributed processing of large data sets across clusters of computers

## MPP Database Management System Platform

Traditional business analytics solutions employ data warehouse systems designed for online transaction processing instead of analytics. These systems are built using general purpose database, server, and storage platforms that lack the specialization required to handle terabytes (TBs) of constantly growing and changing data.

In comparison, today's MPP platforms are no longer limited to SQL programming, and it is not unusual for these platforms to support development in programming languages such as Java*, C/C++, and R. When equipped with rich advanced analytics and in-database data-mining capabilities, these platforms provide the flexibility to leverage existing BI and SQL skills and acquire new skills based on using the most suitable programming language for a specific purpose.

The solution we are using is a third-party data warehouse solution with an asymmetric massively parallel architecture that analyzes huge data volumes up to 100 times faster than traditional systems. Such systems are available from a number of vendors.

We selected our solution based on several factors.

- Cost effectiveness in terms of best price per performance and capacity
- Simplicity and rapid time-to-value
- Scalability in storage and performance from TBs to PBs
- Built-in advanced analytics and integral support of the open source R statistical computing language
- Affinity or overall fit with our existing BI ecosystem
- Interoperability with an enterprise ecosystem

Expressly built for analytics, this system combines proprietary data filtering technology with Intel® Xeon® processor E7 family-based blades and commodity disk drives to deliver big data performance at a low cost and requiring little maintenance. The system is designed to scale performance and capacity through the addition of blades. Each blade is connected to multiple disk drives in close proximity that stream data in parallel to greatly reduce access latencies compared to solutions that use separate data storage systems.

Our selection process included a paper study of 11 data warehouse appliance vendors, followed by a request for proposal (RFP) from five vendors. One vendor was selected based on the RFP analysis and a technical evaluation.

## Hadoop

Hadoop is an open source framework for processing massive volumes of data. Instead of one large supercomputer, Hadoop coordinates local storage and computation across multiple servers that act as a cluster, with each server working with a subset of the data.

Hadoop is a top-level open source project of the Apache Software Foundation. Numerous commercial distributions are also available.

By itself, Hadoop is merely the distributed computing OS, providing two basic services:

- **Hadoop Distributed File System.** This distributed file system provides UNIX*-like file system storage distributed across all nodes in the Hadoop cluster. Hadoop can also use other file systems.
- **MapReduce.** This distributed computation feature is the cornerstone of Hadoop. MapReduce coordinates each of the servers in the cluster to operate on a part of the overall processing task in parallel.

On top of this core are numerous commercial and open source applications, toolkits, and data layers, including:

- **Hive.** An SQL language for querying Hadoop data
- **HBase.** A high-speed, read/write, column-oriented database, able to handle billions of rows times millions of columns
- **Pig.** An interactive scripting environment for processing data
- **Mahout.** A machine-learning library that provides algorithms for clustering, collaborative filtering, and recognizing similarity
- **Sqoop.** An import/export exchange with RDBMS databases
- **Oozie.** A workflow environment for coordinating complex data processing operations
- **Cassandra.** A document-oriented database

Hadoop has the ability to scale linearly. For example, doubling the number of machines in a cluster can cut the processing time roughly in half—or process twice the amount of data in the same time.

Hadoop is written in Java and runs on Linux*. Hadoop applications are also typically written in Java, though other languages are also possible. Some Hadoop tools, such as Hive and Pig, run on a client computer and generate MapReduce programs on-the-fly.

Because Hadoop aggregates the storage of all the servers in its cluster, and these servers can use commodity hard drives, the cost-per-TB of storage is very low, and the amount of storage in a cluster can grow to accommodate PBs of data. Thus, Hadoop makes it cost effective to capture and keep data that was previously thrown away. It also makes it feasible to capture and store data that isn't yet understood, but could be of value. Domains such as text analytics have demonstrated that more data—not less data—yields the best results, even when using simpler algorithms. In domains such as cybersecurity, Hadoop's massive capacity enables analysis across longer time frames.

Hadoop and its related technologies are, in general, not intended to replace online transaction-process ing systems or other traditional RDBMSs. Hadoop's strength is in the batch processing of TBs and PBs of data.

## Hybrid Platform Advantages

Combining the third-party data warehouse appliance with its asymmetric massively parallel architecture components and Hadoop (see Figure 1), we have assembled a big data platform that is cost effective, highly scalable, and plays to each component's strengths. Having the components co-located and connected with a fast network connection and a high-speed data loader allow the big data platform to more effectively move portions of data between the platforms when needed.

## Developing Big Data Skill Sets and Expertise

One of the biggest challenges in big data is addressing the lack of skilled experts. According to the McKinsey Global Institute report cited earlier, by 2018 the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills, as well as 1.5 million managers and analysts with the knowledge of how to use the analysis of big data to make effective decisions.

Skills in the science of big data—such as statistics, mathematics, machine learning, and visual analytics—are essential to have, but also important are the skills required to align the data with the business and turn findings into positive business outcomes. This means that IT customers, such as Intel business groups, need to develop internal big data consumption skills in order to profit from big data.

Many big data technologies such as Hadoop are open source and developed by Internet companies to process large volumes of structured and unstructured data in a cost-effective way. These are maturing rapidly but currently require deeper technical skills in areas such as Linux, Java development, and distributed computing. To deploy big data technologies, companies will also have to develop these skills and more.

### CLOSING THE GAPS IN KNOWLEDGE AND SKILLS

Acquiring new skills can be more difficult than implementing the technology. Intel IT and Intel business groups are working to close this gap in big data expertise and sophistication through

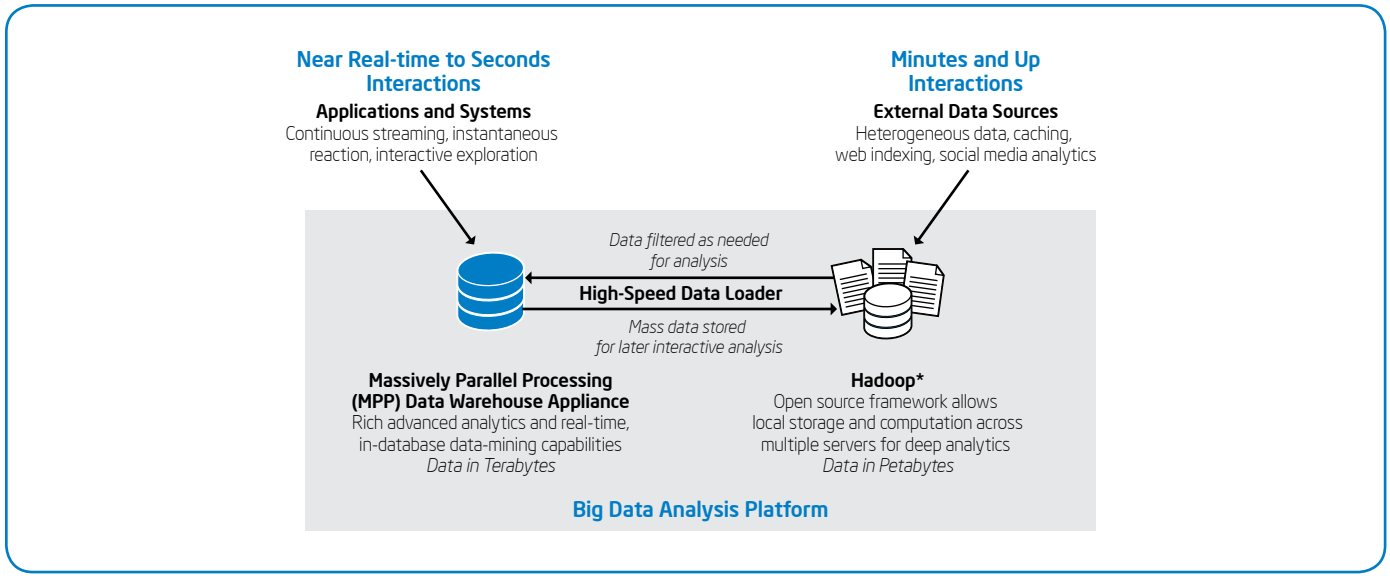


Figure 1. Intel IT's business intelligence big data platform is based on a combination of a massively parallel processing (MPP) data warehouse appliance and clusters of industry-standard servers running Apache Hadoop*.

research and training, hiring people with the essential skill sets needed to work with big data, initiating testing and conducting proof of concept (PoC) efforts such as the ones reported here, and implementing specific use cases.

In the absence of direct experience, working with the data and performing big data analytics is the way to gain this experience. We believe the end result of being able to use big data for predictive and prescriptive BI will be well worth the investment in hardware, software, training, and time.

## PROOFS OF CONCEPT

**Intel has numerous proofs of concept for big data currently in progress. In this section, we discuss four of them.**

### Malware Detection

Cyber threats continue to increase, as attackers and tools become more sophisticated. The goal of security detection is to uncover threats in time for users and security responders to take action against them.

Signature-based file scanning, the most common method of dealing with malware threats in the past, is becoming less effective because of the volume of malware being produced. A better approach is to stay a step ahead of malware, looking deeply into what the malware is doing, where it came from, and even predict where it might arise in the future. This kind of deeper monitoring and prediction requires constantly observing server activity for anomalies at every level: system, network, and application.

The patterns that point to these threats are often hidden in various network and server logs, such as proxy, DNS, DHCP, and VPN, which contain potentially massive amounts of data. Anomalies can manifest as anything from typical malware signatures to broader, global patterns of suspicious activity and behavior, such as communication with URLs that are known to be bad or unusual searches. Analysis requires a complex series of steps, including correlating data from many sources and defining a baseline of normal network activity and patterns in order to detect abnormal activity.

To identify these anomalies over the time-scale required, Intel is using big data technologies to collect the raw, unstructured data and structure it, and then use statistical models, such as predictive analytics, to detect anomalous patterns of activity.

With this PoC, we are aiming for real-time identification of these behaviors, so that malware can be quickly identified and contained. Being able to collect and analyze data amassed over months or even years will help us better predict the sources and nature of compromises, enabling us to implement better preventative measures and systems.

### Chip Design Validation

Chip design involves an extensive amount of testing before a design is ever realized in silicon. This testing continues well into the various stages of silicon implementation where hundreds of sensors collect data at sample rates, thousands of times per second. Such extensive testing generates enormous amounts of data.

In this PoC, Intel IT is studying how a big data platform can be used to optimize the

### The Challenges and Rewards of Big Data

The term *big data* broadly describes information problems that challenge traditional relational database approaches, either by sheer volume, or by variety of sources and types. This variety of data, including text, audio, video, click streams, log files, and more, is big data's other most important characteristic, and why even a few terabytes of unstructured data can be considered big data.

Internet search companies such as Google and Yahoo were among the first to develop big data tools, which were needed to index the World Wide Web. Other Internet companies soon followed, developing other components to handle orders and recommendations, Facebook-type messages, and other problems at Internet-scale. Now, enterprise IT is applying these same tools to high-value business problems that have been difficult to analyze and solve.

Using these new techniques isn't always easy or straightforward. There are significant challenges involved in integrating, deploying, and maintaining these new tools, most of which are still maturing, and require new IT skills in Linux* and Java*. Developing and optimizing a big data solution requires rethinking the problem in terms of parallel computing constructs, such as MapReduce, and not all problems parallelize as well as web indexing. Moreover, solving problems at-scale may require the relaxation of the ACID semantics—atomicity, consistency, isolation, and durability—that database programmers have come to depend on, and trading off low latency for high throughput.

Where current data systems suffice, there may be no point in switching. But for all those problems that have been out of reach, big data solutions may be the answer. Even where a clear use case has not yet been defined, companies can take advantage of the low cost of big data storage to capture and store virtually everything possible and extract its latent value later.

validation process, analyzing billions of rows of both structured and unstructured data to help speed up the design process and time to production—and ultimately time to market.

A good example for this usage model is something we call coverage. In the world of post-silicon validation, there are no clear rules as to when a chip is ready to be launched. On the one hand, releasing a chip with bugs can result in great damage to the company's reputation. On the other hand, excessive testing can delay the introduction of the chip, resulting in the company losing millions of dollars in sales. The concept of coverage aims to avoid these extreme cases. By collecting data on the logical and physical states in which the processor has been tested—or covered—we can better understand how the tests and testing tools are doing, and determine whether the chip is ready to be released into the market.

Big data analytics can also assist in the debug process by being able to automatically cluster and sort identified defects, as well as perform root cause analysis on large volumes of historical test. Through extensive analysis of the large masses of data collected—not just samples—we can achieve a much more comprehensive picture of the progress at each stage and discover ways to improve and streamline the design process, as well as ultimately improve the product itself.

## Market Intelligence

For a company such as Intel that has worldwide sales and a global supply chain, it's critically important to be able to foresee changing market conditions and make accurate forecasts about what might happen next month, six months from now, and even five to 10 years from now. Global companies must sort through enormous amounts of data, including weather trends, global economic data, discussion forums, news sites, social networks, wikis, tweets, and blogs. From this data, companies can make accurate projections, plan sales strategies, assess competitor threats, anticipate changes in consumer behavior, strengthen supply chains, and improve business continuity plans.

For this PoC, we are working with Intel business groups to analyze data from a wide variety of disparate sources in anticipation of achieving the following:

- Improve our projections on potential sales in various parts of the world, fine-tune production levels, and provide more accurate forecast to our shareholders
- Build and test scenarios based on potential global events, to determine their effect on our markets, our supply chains, and our ability to respond to market demand and competitive challenges
- Uncover new users and new uses for our products

## Recommendation System

With the amount of content growing exponentially, users often need assistance finding the information that best matches their inquiries and interests. For this reason, the demand for recommendation-based services is growing across Intel for both internal and external applications. Recommendation systems, similar to those provided by Amazon and Netflix to their customers, assist users by reducing search and navigation time and enabling more personalized and targeted results. This improves productivity, credibility, and the overall user experience.

Implementing a scalable recommendation system requires predictive analytics and big data expertise because it involves executing complex resource-intensive algorithms over large volumes of historical data.

This PoC is focused on building a generic, reusable recommendation engine that includes a two-layer—offline and online—architecture on top of our big data platform. The offline component is a batch-oriented process that executes the core of the recommendation algorithm. It guarantees that our models can scale up by performing the big data crunch on a scalable environment. The online component acts as a service layer for any service request. It loads the relevant intermediate calculation that was calculated during the offline phase, and performs the last step in the algorithm for generating the recommendation itself. It also applies context-configured logic to filter and adjust the final recommendation according to the request context.

The scalability of the solution is achieved by implementing the core of the algorithms using Mahout. Mahout is an open source data mining library written in Java on top of Hadoop. Mahout takes advantage of the Hadoop architecture by executing parallel jobs inside a cluster of commodity hardware in a shared-nothing environment. All intermediate results are written into the MPP RDBMS for fast retrieval by the online component.

Deploying this recommendation service will be a key enabler in providing just-in-time personalized content. This will enable us to increase employee productivity when using Intel internal applications and to also help provide a competitive advantage that can improve external customer selection of our products and thus contribute to our revenue. We could apply the experience and knowledge we gain from providing a complex predictive analysis over large data volumes to delivering similar solutions in the future.

## CONCLUSION

**Intel IT is taking a systematic approach to adding big data analytics to its overall BI efforts, starting with several proofs of concept in 2012. By adding the ability to mine and analyze big data, Intel expects to evolve its BI capabilities from descriptive analytics to predictive and prescriptive analytics that will enable deeper and richer insights into business patterns and trends.**

We have completed the first step, which was to design and build a big data platform combining a third-party data warehouse appliance with Hadoop, an open source framework for processing massive volumes of data across multiple servers. This solution allows us to perform MPP on structured data and distributed processing of large data sets over industry-standard servers. We are also developing internally the necessary big data skill sets, expertise, and sophistication within our IT BI staff and business groups.

Upon successful completion of these proofs of concept, Intel expects to take its big data platform into production and use it to solve high-value business problems to achieve new operational efficiencies and to both increase and add new revenue sources. Over the next few years, we anticipate our big data analytics program will grow, providing Intel with BI we can use to achieve new competitive advantages in manufacturing, security, marketing, market development, and IT.

## FOR MORE INFORMATION

**Visit www.intel.com/it to find white papers on related topics:**

- "Roadmap for Transforming Intel's Business with Advanced Analytics"


## CONTRIBUTORS

**Jessica Brindle,** Business Intelligence Strategic Planner, Intel IT


## ACRONYMS

| | |
|---|---|
| BI | business intelligence |
| MPP | massively parallel processing |
| PB | petabyte |
| PoC | proof of concept |
| RDBMS | relational database management system |
| RFP | request for proposal |
| TB | terabyte |

**For more information on Intel IT best practices, visit www.intel.com/it.**