IT@Intel White Paper
Intel IT
IT Best Practices
Big Data and IT Innovation
February 2013

Integrating Apache Hadoop* into Intel's Big Data Environment

In one proof of concept, the new platform enabled us to perform root cause analysis and automated incident prevention, with a potential to reduce the number of incidents by 30 percent.

Executive Overview

Intel IT compared the Intel® Distribution for Apache Hadoop software (Intel® Distribution) to two other Apache Hadoop* distributions using a well-defined set of evaluation criteria. The evaluation criteria included aspects of platform architecture; administration, operations, and support; and the unique value proposition of each distribution. We tested, validated, and graded each distribution against these criteria, and gave each distribution a score.

Our evaluation confirmed that using Intel Distribution offers significant advantages over other distributions.

- Platform architecture that supports seamless integration with existing infrastructure, high availability, and multi-tenancy support
- Ease of administration, operation, and support through streamlined setup, management, security, and troubleshooting
- Improved performance through optimization for Intel® architecture, along with enhanced encryption using Intel® Advanced Encryption Standard New Instructions
- A close relationship with the open source community and alignment to the open source roadmap

Our Hadoop platform will complement other business intelligence platforms, such as our cost-efficient, highly scalable Intel architecture-based enterprise data warehouse, in-memory database, and custom data warehouses. During our evaluation and deployment, we developed a number of best practices that will help us support business groups' adoption of the new platform.

Our initial investment in planning has resulted in a platform capable of supporting structured and multi-structured analytic data use cases, and scalable to meet evolving needs. In one proof of concept, the new platform enabled us to perform root cause analysis and automated incident prevention, with a potential to reduce the number of incidents by 30 percent. In the current IT economic environment, this represents a significant cost savings and cost avoidance, and helps to increase employee productivity.

Assaf Araki
Big Data Analytics Engineer, Intel IT

Ajay Chandramouly
Big Data Industry Engagement Manager, Intel IT

Nghia Ngo
Big Data Capability Engineer, Intel IT

Sonja Sandeen
Big Data Project Manager, Intel IT

Darin Watson
Big Data Platform Engineer, Intel IT

Chandhu Yalla
Big Data Engineering Manager, Intel IT

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

The proliferation of digital technologies and digital storage has created an explosion of data that is beyond the processing capabilities of traditional data platforms. Examples of user-generated and machine-generated data include Web logs, radio-frequency identification, sensor networks, social networks, Internet text, security logs, and video archives.

Intel's currently deployed business intelligence (BI) platforms do not support multi-structured data, and these platforms cannot accommodate the big data analysis required for deeper insights and faster, better decision making. The fundamental change in the type and amount of data in the enterprise demands a similar change in our vision, strategy, and platforms for handling data analysis.

To address this issue, we have assessed the available data in use at Intel and developed a strategy for managing that data.[1] To deliver business value through a wide range of data and analysis needs, we have developed a strategy that envisions multiple BI platforms that support storage and analysis of data with different characteristics. Examples include our cost-efficient, highly scalable Intel® architecture-based enterprise data warehouse, in-memory database, and custom data warehouses. Our strategy includes choosing the most appropriate BI platform for each use case.[2] This proactive strategy is a key step in realizing value from rapidly growing and diverse data sets.

High-velocity, High-volume Analysis is Becoming Increasingly Important

Traditionally, we have focused on the analysis of structured data in relational databases. But today, most of the data is unstructured and is accumulating at a high rate from the Web, networks, sensors, and other sources. The ability to rapidly perform high-volume analysis is becoming more important. To maintain Intel's competitive advantage, supported by quick, well-informed decision making, we need to take advantage of many more data sources than we have in the past. However, because our current relational data warehouses are not designed for analyzing this type of data, we investigated other BI platforms based on Not Only SQL (NoSQL). These platforms are better suited for the storage and processing needs of large volumes of unstructured data in a timely manner.

The leading and most widely used NoSQL platform is the open source Apache Hadoop* project, which includes the Hadoop Distributed File System (HDFS*) and HBase*, a non-relational, distributed database. Other NoSQL solutions—open source or proprietary—are not as mature as Hadoop and HBase. In addition to its maturity, a Hadoop-based platform can enable us to maintain a flexible IT ecosystem that evolves with our needs.

Several Hadoop-based solutions are available, including the pure open-source code and third-party distributions. Currently, the pure open-source version of Hadoop is designed for batch processing, and HBase is not optimized for high-velocity processing. We decided that a third-party Hadoop distribution would best meet our analysis needs.

[1] For more information, see "Enabling Big Data Solutions with Centralized Data Management," January 2013.

[2] For more information, see "Improving BI Analytics at Intel with Multiple Data Warehouses", release date Spring 2013.

Apache Hadoop* Platform Poses Specific Challenges

Although business groups at Intel see the value of using Hadoop-based solutions, several factors make adopting those solutions challenging.

- Most Hadoop-based platforms are built on open source technology. The dynamics of managing open source development and support is a new concept to most of Intel's development community and potentially impacts many areas of development such as product and capacity management, storage and migration services, and governance.
- Application developers need to develop new skills. For example, they must change from the familiar SQL language to writing MapReduce* code in Java*. Also, the distributed algorithms, which are less intuitive and common than traditional sequential algorithms, require a different way of thinking.
- Big data—especially multi-structured big data—is a relatively new field. Keeping up with the constantly changing array of available tools, hardware, and software solutions requires a significant investment in education and continuous improvement.

To help address these challenges, we were chartered to evaluate multiple Hadoop distributions and deliver a fully integrated production platform. Our task involved functioning like a big data service provider, improving the ease of platform adoption and making it easier for business groups to derive business value from big data.

CHOOSING AN APACHE HADOOP DISTRIBUTION

Our strategy involved comparing the Intel® Distribution for Apache Hadoop* software (Intel® Distribution) to two other Hadoop distributions, then choosing the one that best met Intel's requirements and integrating this new technology with our existing infrastructure and other BI platforms. This approach maximizes the benefits of each technology deployment and enables business groups to use the appropriate BI platform for their particular use case.

Strategic Objectives

During the evaluation process, we kept the following strategic objectives in mind:

- Increase and deliver high-performance, high-velocity analysis and reduce storage costs by using compute and storage clusters with cost-effective but powerful servers based on the Intel® Xeon® processor E5 family.
- Reduce the support, administration, and management burden by using a single dashboard.
- Improve the performance of processing large data sets and another layer of security by utilizing a high-performance network featuring a 480 gigabit-per-second cluster fabric bandwidth.
- Increase the speed of data availability, reduce business unit adoption times, and arrive at business value faster by delivering big data as a comprehensive service and tightly integrating the platform with existing security infrastructure, data warehouses, and tools.
- Simplify the configuration and management process by using enterprise access management, role-based security integration, and directory services.

Evaluation Methodology

First we developed a list of evaluation criteria. Each criteria was qualified and a weighting factor from one to five was assigned to ensure a comprehensive result based on the priorities for selection. Then we tested, validated, and graded each distribution against these criteria, and gave each distribution a score. We confirmed that the Intel Distribution best satisfied our criteria—including reducing the obstacles to adoption by business groups—and demonstrated the lowest total cost of ownership (TCO) in terms of initial cost and ongoing support.

As shown in Figure 1, Intel Distribution is a comprehensive solution that contains the full distribution from the Apache Hadoop open source project, along with MapReduce, HDFS*, and related components such as the Hive* data warehouse infrastructure and Pig* data flow language. Intel Distribution also supports Apache Mahout* and the Intel® Graph Builder for Apache Hadoop software. Solution elements are pre-integrated to simplify management and deployment and enable faster time to market, which helps minimize training and financial investments.

After selecting the Intel Distribution, we designed and implemented the entire platform in just five weeks. The strong cross-organizational partnership between Intel IT and the Intel Software and Services Group made this accomplishment possible.

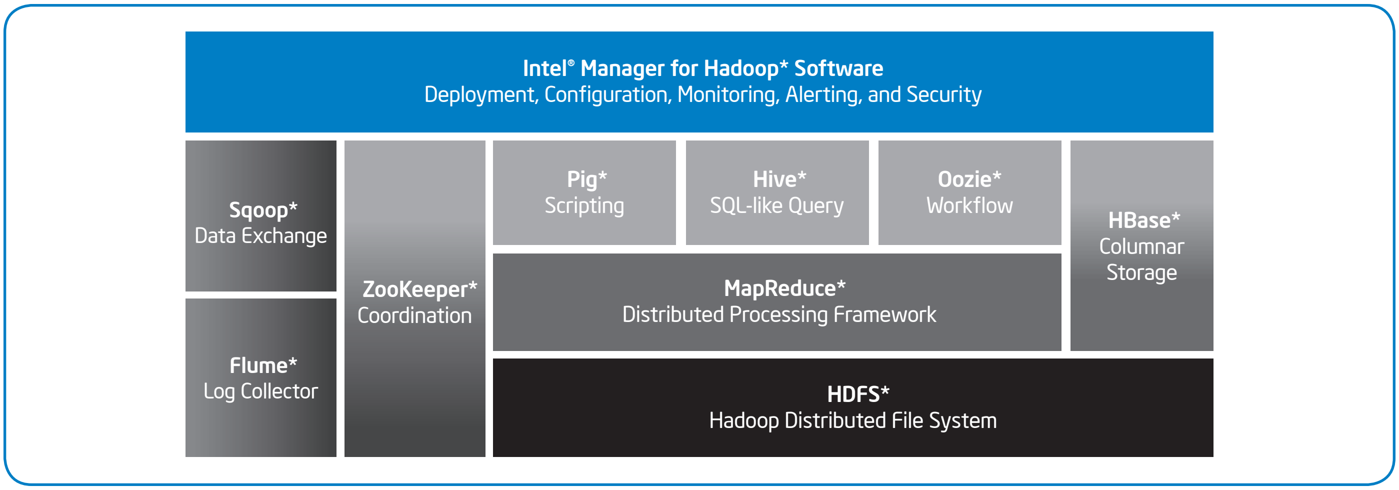
Intel® Manager for Hadoop* Software
Deployment, Configuration, Monitoring, Alerting, and Security
Sqoop*
Data Exchange
Flume*
Log Collector
ZooKeeper*
Coordination
Pig*
Scripting
Hive*
SQL-like Query
Oozie*
Workflow
HBase*
Columnar Storage
MapReduce*
Distributed Processing Framework
HDFS*
Hadoop Distributed File System

Figure 1. The software components of Intel® Distribution for Apache Hadoop* software provide a comprehensive big data solution.

Evaluation Criteria

We divided our evaluation criteria into several categories. The following subsections describe how the Intel Distribution best met our requirements in each of these categories.

PLATFORM ARCHITECTURE

The Hadoop distribution needed to meet several high-level enterprise requirements.

- **Security integration.** We have an extensive information security infrastructure in place to protect one of Intel's greatest assets–intellectual property.
- **High availability.** Intel's business runs 24 hours a day, 7 days a week, and 365 days a year and cannot afford downtime.
- **Multi-tenancy support.** In a distributed system such as a Hadoop cluster, we must be able to prioritize and allocate compute resources to a particular job.
- **Integration with existing BI platforms and analytical tools.** We have made significant investments in tools and technologies that support data management and analysis. We wanted a distribution that provided open APIs that allow other tools to be easily integrated and used.

In our evaluation, we found that the Intel Distribution best met these requirements.

Integration Details

The Intel Distribution supports integration with all of the following:

- Existing data warehouses and massive parallel processing systems
- BI reporting tools and analytics engines
- Data tools, such as various extract, load, and transfer tools
- Enterprise scheduling and access management tools
- Advanced analytics tools such as Mahout, a machine learning library that includes MapReduce algorithms and integration with the open source R statistical programming language

ADMINISTRATION, OPERATIONS, AND SUPPORT

While capabilities are important, the ease of administration, operation, and support is also a factor in determining TCO. In particular, for each Hadoop distribution we examined the upgrade, provision, and configuration management features, its ease of use, and the learning curve.

The information we gained enabled us to discuss with upper management both the amount of effort required to form support teams and the amount of supplier-based training and consulting services that would be required.

In our evaluation, we found the evaluation team didn't require any specific training on the Intel Distribution, underscoring its ease of adoption and integration capabilities. Although the Intel Distribution is easy to learn, formal training is also available, which our engineering team received from the Intel Distribution product team, after the evaluation was complete and prior to the production implementation. This training helped us resolve technical issues and gain the knowledge and confidence for prompt delivery and implementation.

Intel Distribution includes Intel® Manager for Hadoop Software (Intel® Manager). Intel

Manager is a web-based management console designed to install, configure, manage, monitor, and administer the Hadoop cluster. It uses Nagios* and Ganglia* to monitor resources and configure alerts in the cluster. We found that with little or no training, teams can use Intel Manager to streamline setup, management, security, and troubleshooting for Hadoop clusters.

Intel Manager also supports secure authentication and authorization using Kerberos and built-in access control rules. With this powerful, easy-to-use tool, we can focus critical expertise on driving business value from the Hadoop environment instead of having to worry about the details of cluster management.

UNIQUE VALUE PROPOSITION

During the evaluating of each distribution, we explored what unique value the third-party supplier provided, compared to other distributions or to the pure open source code. We also evaluated how well each distribution was connected to the open source community and aligned with the Hadoop release roadmap.

We concluded that the Intel Distribution provided the most value compared to the other two distributions, offering the following benefits:

- Optimized for Intel® architecture
- Ability to take advantage of Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI)
- Fully aligned to the open source community, which avoids unnecessary delays in the future adoption of new capabilities

Optimization for Intel® Architecture

While the other distributions we evaluated were focused only on software, Intel's software teams have optimized the open source Hadoop stack to take full advantage of the high-density, cost-effective, and easily scalable Intel Xeon processor E5 and Intel Xeon processor E7 families, dramatically reducing the time it takes to analyze data. Because Hadoop code is highly distributed, coding efficiencies are multiplied across the infrastructure, improving performance and reducing energy consumption and capacity requirements for servers and storage controllers.

Intel Distribution provides sub-second queries and analytics over large data sets stored in HBase. As shown in Figure 2, our internal measurements showed a 5x performance increase with the fully optimized Intel Distribution on Intel architecture, compared to the same jobs run on an unoptimized open source stack. This optimization and resulting velocity enable the agile decision making that provides optimal business value.

Security Features Embedded in the Solution Stack

The Intel Distribution was the only distribution in our evaluation that takes advantage of Intel AES-NI for HDFS and MapReduce encryption, providing both file- and cell-level accelerated encryption. The Intel Distribution also allows fine-grained access control lists using directory services, creating further value.

Intel Xeon processors feature hardware-assisted security technologies that cover the entire compute platform—from the data center to the desktop or mobile client. Together, these technologies support faster response time when accessing encrypted data, stronger authentication to help protect sensitive information, and enhanced protection against security breaches. Intel Distribution also includes many data security features, such as enhanced authentication and provisioning for more secure data movement.

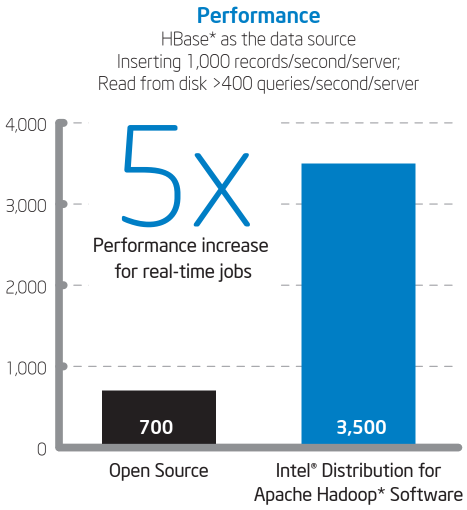

Figure 2. Optimized for Intel® hardware, the Intel® Distribution for Apache Hadoop* software provides a 5x performance increase over non-optimized open source code. Intel internal measurements.

Alignment to Open Source Standards and Roadmap

The technical features available must be balanced against the investment required to ensure future alignment to open source community releases. Reduced TCO can be achieved only if a solution has value-added technology, is easy to deploy and maintain, and can scale with future technology enhancements.

Intel is a well-known contributor to open source software, with significant contributions to the Hadoop framework and related projects. Intel's Hadoop contributions focus on enabling the open source community and users to fully utilize underlying hardware, storage, and networking technology for the next-generation data center. In our evaluation, we concluded that the Intel Distribution was as closely aligned to the open source roadmap as any other third-party distribution. Also, most Intel Distribution components can take advantage of open source code.

Intel® Distribution Platform Components

We realized that big data use cases at Intel will continue to grow, and needs will change over time. Therefore we wanted to build a cost-effective and scalable platform that was able to support both compute- and storage-intensive use cases.

Our initial investment in planning has resulted in a platform that

- Is capable of supporting unstructured and multi-structured analytic data use cases
- Features a platform design and architecture right-sized for today and the immediate future
- Is scalable and expandable to meet evolving needs

Our Hadoop platform contains the following basic components:

- Servers based on the Intel Xeon processor E5 family, with a total of 192 cores, composed of 16 nodes that use a total of 96 terabytes of storage space
- The latest generation of the Intel Distribution, which is based on the 1.x release of Hadoop

Apache Hadoop Use Cases at Intel

We have identified three initial use cases where the unstructured and multi-structured data analysis capabilities of the Intel Distribution can add significant business value.

- Incident Predictability. We wanted to proactively understand and monitor client incidents and use big data to automate root cause analysis and incident prevention. We conducted a proof of concept (PoC) that demonstrated the potential to reduce the number of incidents by 30 percent, or about 4,000 incidents per week.
- Recommendation Engine. We wanted to deliver a better mobile experience through contextually aware applications. By utilizing Mahout on top of Hadoop we are able to deliver a highly scalable recommendation service that various solutions can utilize.
- Customer Insight. This use case processes raw web data and then integrates this data with internal consumption data available in a different data warehouse. By using the Web metrics, the Intel supply chain can improve product availability and maintain appropriate inventory levels for specific regions.

We are actively engaging with other Intel business groups to identify and deploy additional use cases.

Best Practices for Designing and Deploying a Hadoop Platform

As we evaluated, designed, and deployed our Hadoop platform using the Intel Distribution, we identified several best practices that enabled us to swiftly implement the platform and derive business value from it.

Table 1, on the next page, describes these best practices.

NEXT STEPS

We are actively working with three customer projects. In 2012, we demonstrated a small team of five people skilled in BI can deliver in just six months up to USD 10 million in returns. We are now exploring use cases for big data analysis where we can expect returns of five to 10 times higher.

These use cases will further demonstrate the business value of the Intel Distribution. We plan to scale and expand the platform and its capabilities as use case demand increases. We also plan to continue studying our first set of use cases to further learn about and evolve the platform and ensure delivery of a solid platform. This approach will help us quickly acquire business value and best-known methods that we can apply to the new set of use cases. We anticipate much of this business value will be derived from predictive analytics. While data mining in big data sets is valuable, using the power of the Hadoop platform to identify future trends is even more valuable.

Our team will include this new BI platform as part of a fully integrated BI service, incorporating the current set of BI platforms and associated IT processes. This service will include providing prescriptive guidance for development and architecture and standardized processes and tools.

Table 1. Best Practices for Designing and Deploying a Hadoop* Platform

Take Advantage of Internal and External Resources	
Invest in training as early as possible	Open source solutions require different application development skills, and traditional development methodologies may not be sufficient in an open source environment. Realizing that developers needed to change the way they think, work, and respond, we provided training that would help them develop the necessary new skills.
Develop small, localized expert teams to become the domain experts	We formed a fully dedicated team of architects, engineers, and developers and gave them decision making authority. We found this approach enabled us to deliver solutions quickly and increased the rate of adoption of big data technologies at Intel.
Understand and consistently take advantage of open source community resources	We found that becoming familiar with available open source projects, existing open source code that can be reused, and solutions from the industry can help avoid having to rerun proofs of concept in situations where the internal use case and the existing project or solution is the same—ultimately saving time and effort.
Consult internal and external subject matter experts	We found these resources can help get a Hadoop project off to a good start and can minimize the technical team's learning curve. The experts made recommendations about approaches and tools that are best suited for a particular project and helped demystify big data and Hadoop.
Start Small to Reduce Rework and Redesign	
Develop the core framework and the critical factors for business requirements	By starting small in terms of teams of experts, platform, and projects, we were able to deliver a full production platform in five weeks. We continue to use this approach as we pilot two to three projects at a time.
Use virtual machines when possible	Virtualization can support functional testing without over-subscribing system resources. Because provisioning can be done quickly, virtualization also enables rebuilds that may be required to validate system integrity.
Prioritize platform and application integration	We found that focusing on integrating existing tools and platforms with the core Hadoop framework makes it easier to scale the platform and add complementary Hadoop components as needs evolve.
Use Agile Methodology	
Prioritize and deliver with agility and flexibility	Instead of over-engineering solutions and aiming for perfection, we prefer to quickly deliver a solution that can immediately provide results. This approach enables us to deliver business value in a short amount of time.
Publish a reasonable, standard product offering that meets current needs	The big data solution ecosystem is volatile—new suppliers and new integration tools appear constantly. It can be challenging to keep up with update cycles measured in days and weeks, not months and years. We found it wasn't feasible to spend six months designing and deciding on technology, because in those six months Hadoop technology would have changed by several generations.
Frame project design around big value requirements and deliver in smaller blocks of time	Our team breaks down deliverables into four- to six-week intervals. We follow an agile methodology, which we applied to the evaluation; we continue to apply it to ongoing projects.
Invest in Automation and Standardization	
Automate to enhance support, maintenance, and delivery	Although it was time-consuming to develop automated scripts during the engineering build and test phases, we found that in the end, the time was well spent. Automation enabled significant time savings for future projects.
Standardize and create re-usable templates and scripts	We encouraged our small team of experts to be decisive about standardization. In some cases the decisions were not ideal, but still provided a solid foundation upon which to build as more customers begin to use the platform.
Enforce development and control standards	Implementing a multi-tenancy cluster creates a certain amount of lack of control over compute resources. We established working procedures and control processes to better assign and prioritize compute resources, commensurate with job priority.
Address Training Requirements	
Prepare for a steep learning curve	Because our technical team was new to open source technology, we scheduled time for the team to become familiar with the technology in a lab environment, through hands-on analysis and by taking Hadoop developer and administration training courses. For example, they needed to learn how to write MapReduce* code in Java* and develop distributed algorithms.
Provide cross-functional training	We learned that big data skills and training could transcend job roles such as application developer, engineer, or analyst. For example, an engineer may need some training in data analytics to fully comprehend the end usage of the data.
Engage in Proactive Customer Interaction	
Actively manage new engagements to guide customers to the right solutions	Interest level in the Hadoop platform was high, and we were inundated with requests before we even had a production platform. To deal with this situation, we found we needed to educate customers about what big data really is and which BI platform, such as in-memory database, Hadoop, or enterprise data warehouse, would be best suited for their particular data set. Guiding the customer to the appropriate solution helps ensure each project's success.
Prioritize projects to optimize use of limited resources	Once we identified projects well-suited to the Hadoop platform, we decided we needed more information to prioritize these projects given the limited resources available to us. We asked our customers to answer a short survey, which we then graded on a weighted scale. We used the results to identify projects with high business value and solid customer commitment, and have found this to be a valuable tool for prioritization.
Collaborate through knowledge sharing	Customers rely on our engineering and solutions team to guide them and answer their questions. We have found educating the customer and providing an overview of big data is a necessity. We encourage customers who are new to big data to create a workgroup that can share knowledge throughout their team. We guide them to knowledge resources including online group discussion forums, blogs, newsletters, and technical or business audience forums.

CONCLUSION

After comparing two other Hadoop distributions to the Intel Distribution, we confirmed the Intel Distribution best meets our needs. In particular, it offers significant value compared to the other distributions we evaluated. The Intel Distribution supports seamless integration with our existing security, management, and analysis tools; features a highly available platform architecture that supports multi-tenancy; includes Intel Manager for Hadoop Software; and is optimized for Intel architecture. Other advantages of the Intel Distribution include extensive support services and training, and a close relationship with the open source community.

We have already deployed a production version of the Intel Distribution and have identified three use cases that demonstrate the business value of this addition to our BI portfolio. In one PoC, the new platform enabled us to perform root cause analysis and automated incident prevention, with a potential to reduce the number of incidents by 30 percent.

During our evaluation and deployment, we developed a number of best practices that have helped us build a cost-effective, flexible, and scalable big data platform right-sized for Intel's needs today and that can expand to meet evolving needs. In 2012 we made significant progress toward meeting Intel IT's big data goals; we will continue to build on those successes in 2013 and beyond.

RELATED INFORMATION

- Intel Big Data Resources: www.intel.com/bigdata
- "Enabling Big Data Solutions with Centralized Data Management," January 2013
- "Improving BI Analytics at Intel with Multiple Data Warehouses", release date Spring 2013

CONTRIBUTORS

Moty Fania, Intel IT

ACRONYMS

BI	business intelligence
HDFS	Hadoop Distributed File System
NoSQL	Not Only SQL
PoC	proof of concept
TCO	total cost of ownership

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

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/performance/resources/benchmark_limitations.htm or call (U.S.) 1-800-628-8686 or 1-916-356-3104.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. For more information go to: www.intel.com/content/www/us/en/processors/processor-numbers.html

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Printed in USA ♻ Please Recycle 0213/ACHA/KC/PDF 328640-001US