IT@Intel White Paper

Intel IT
Big Data and Business Intelligence
October 2013

Intel IT Best Practices for Implementing Apache Hadoop* Software

In just five weeks, we implemented a low-cost, fully realized big data platform based on the Intel® Distribution for Apache Hadoop* software, which is delivering BI results worth millions of dollars to Intel.

Ajay Chandramouly
Big Data Domain Owner, Intel IT

Sonja Sandeen
Big Data Product Manager, Intel IT

Chandhu Yalla
Big Data Engineering Manager, Intel IT

Yatish Goel
BI Technical Integrator, Intel IT

Nghia Ngo
Big Data Capability Engineer, Intel IT

Darin Watson
Platform Engineer, Intel IT

Executive Overview

Recognizing the value that big data can contribute to business intelligence (BI), in 2011 Intel began researching and developing a big data platform based on the Apache Hadoop* open source software framework. After a paper analysis and technical evaluation of products from multiple vendors, we implemented a low-cost, fully realized big data platform based on the Intel® Distribution for Apache Hadoop* software (Intel® Distribution) in just five weeks. This platform currently supports three use cases, with more in development, delivering BI results worth millions of dollars to Intel.

We found that hardware, networking, and software selections for a Hadoop implementation all significantly influence performance, total cost of ownership, and return on investment. Maximizing value requires combining a cost-effective infrastructure with a Hadoop distribution optimized for performance. Maximizing performance requires implementing the most effective data transfer and integration methods for the existing enterprise BI platforms and the selected Hadoop platform, as well as processes to ensure optimal use in a multitenancy environment.

Our research and optimization efforts enabled us to achieve the following benefits with our selected Hadoop implementation:

- Performance and value through a solution designed to run on the latest Intel® architecture
- Full integration with our existing BI environment, as well as our security, management, and analysis tools
- Development of reusable "how to" guidelines for adding more cost-effective, flexible, and scalable big data platforms in the future

This paper describes how Intel IT quickly implemented its first big data Hadoop platform, why we made certain decisions and optimizations, and what we achieved in our first three use cases. In the future, we anticipate that our use of Hadoop platforms will increase to meet the demands of new use cases and the need to improve Intel's operational efficiency, market reach, and business results.

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

In an age when organizations such as Intel are rich in data, the true value of this data lies in the ability to collect, sort, and analyze it to derive actionable business intelligence (BI). Recognizing the need to add big data capabilities to our BI efforts, Intel IT formed a team to evaluate several Apache Hadoop* distributions and consider implementation options. Our goal was to deliver a production platform in 10 weeks or less.

Intel's big data platform, as shown in Figure 1, consists of three components:

- **A massively parallel processing (MPP) platform.** Unlike traditional business-analytics solutions, which process online transactions, the MPP platform was built for large-scale analytics. The MPP platform uses blade servers based on the Intel® Xeon® processor E7 family. We use the MPP platform where high performance at a lower cost—compared to using an enterprise data warehouse—is needed.
- **The Intel® Distribution for Apache Hadoop* software (Intel® Distribution).** We chose the Intel Distribution for its ability to process large volumes of variable, multidimensional, unstructured data where join conditions are unknown and where the goal is to discern patterns. The Intel Distribution is built using industry-standard servers based on the Intel® Xeon® processor E5-2600 product family.
- **A predictive analytics engine.** We developed this engine to deliver information and insight using real-time, ongoing predictive analytics.

To ensure a satisfactory return on investment (ROI) from the start, we targeted three use cases that required a big data platform: contextual recommendation, incident prediction, and web analytics. Our strategy was to start small, minimizing costs while improving our solution. Having three use cases to consider helped ensure that we would design a platform capable of handling a variety of data types, making it more adaptable to additional use cases in the future.

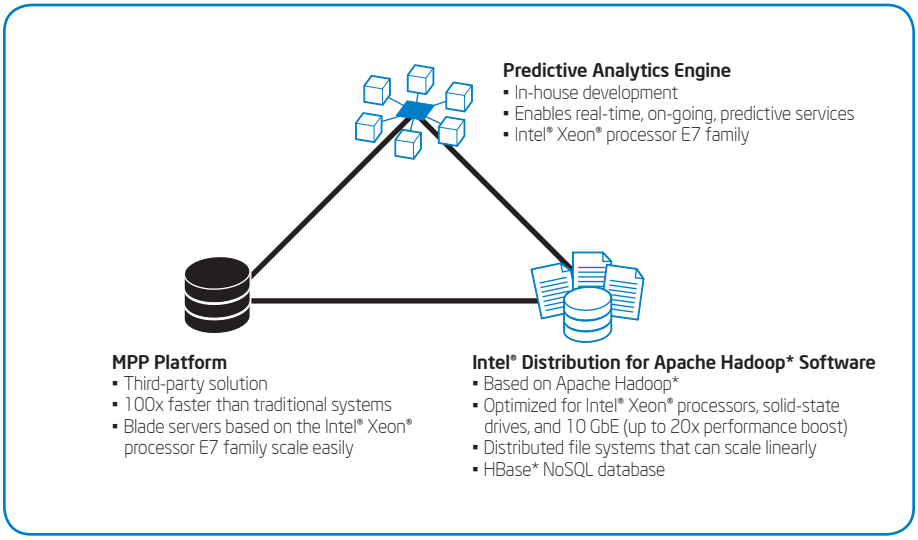


Figure 1. Intel's big data platform comprises three components: a massively parallel processing (MPP) platform, the Intel® Distribution for Apache Hadoop* software, and a predictive analytics engine.

Gaining proficiency in managing open source development and support was challenging early in our process. Intel IT application developers, who were accustomed to using relational database language and traditional sequential algorithms, had to learn to write MapReduce* code in Java* and use distributed algorithms. While not the subject of this paper, such skills are an important part of implementing a big data platform and were performed in parallel with our discovery process and implementation.

While many leading Internet companies have pioneered the use of big data platforms and derived excellent value from them, handling big data is still new to many organizations. As interest in big data platforms has grown, so have the complexity and the number of product choices available for implementing them. We believe IT organizations need to evaluate Hadoop distributions, infrastructure needs, and data integration techniques to ensure delivery of a platform that delivers maximum ROI.

Hadoop Distribution Considerations

Hadoop is a top-level open source project of the Apache Software Foundation. Several suppliers, including Intel, offer their own commercial Hadoop distributions, packaging the basic software stack with other Hadoop software projects such as Apache Hive*, Apache Pig*, and Apache Sqoop*. These distributions must integrate with data warehouses, databases, and other data management products so data can move among Hadoop clusters and other environments to expand the data pool to process or query.

To find the best distribution for our purposes, the BI big data project team selected three Hadoop distributions to evaluate, comparing the Intel Distribution to two other Hadoop

Intel® Distribution for Apache Hadoop* Software

The settings for the Hadoop environment are critical for deriving the full benefit from the rest of the hardware and software. The Intel® Distribution for Apache Hadoop* software (Intel® Distribution) includes Apache Hadoop* and other software components optimized by Intel to take advantage of hardware-enhanced performance and security capabilities.

The Intel Distribution is an open source software product designed to enable a wide range of data analytics on Apache Hadoop. It is optimized for Apache Hive* queries, provides connectors for open source R* statistical programming language, and enables graph analytics using Intel® Graph Builder for Apache Hadoop* software—a library to construct large data sets into graphics to help visualize relationships between data. Intel® Manager for Apache Hadoop* software, included in the Intel Distribution, provides a management console that simplifies the deployment, configuration, and monitoring of a Hadoop deployment.

The Intel Distribution is available worldwide for evaluation.

Key Features:

- Boost in Hadoop performance through optimizations for Intel® Xeon® processors and Intel® 10 GbE Server Adapters
- Data confidentiality through encryption and decryption performed without a performance penalty in the storage layer—Hadoop Distributed File System* (HDFS)—taking full advantage of enhancements provided by Intel® Advanced Encryption Standard New Instructions
- Role-based access control with cell-level granularity available through HBase*, an open source, nonrelational distributed database that runs on top of HDFS
- Multisite scalability and adaptive data replication enabled through HBase and HDFS
- Up to a 3.5x improvement in Hive query performance
- Support for statistical analysis with the R programming language connector
- Graph analytics enabled through the Intel Graph Builder for Apache Hadoop software

distributions using a well-defined set of evaluation criteria that included the following:

- Overall platform architecture, including security integration, high availability, and multitenancy support
- Application architecture and capabilities, including integration with extract-transform-load (ETL) tools, the machine learning and data mining library Mahout*, and the R-based data management and analysis collection (RHadoop*)
- Ability to optimize platform hardware capabilities
- Administration and operations, including upgrade, provision, and configuration management
- Supplier support

Hadoop Infrastructure Considerations

A Hadoop framework typically performs parallel processing on large server clusters built using standard hardware to provide a cost-efficient, high-performance analytics platform. To achieve a high ROI, Intel IT sought a server-and-rack design that would perform cost-effectively for the Hadoop distribution we selected, as well as provide hardware-assisted security technologies. We also wanted to select and implement networking and storage solutions capable of handling the structured and multistructured data of our BI use cases, while also providing excellent data transfer speeds, scalability, and security.

Data Integration Considerations

To feed data from our three use cases into the big data platform and integrate the platform with our existing BI environments, we needed to define a data integration methodology. In the multistructured world of big data, the goal is to extract data from the source and load this raw data quickly and efficiently into the big data container. The distributed compute and storage power of the big data engine is used to perform data transformations and run algorithms that condense large volumes of data into useful aggregated results. An example of such a data-reducing algorithm is MapReduce.

ETL tools move data from sources to targets. Since no single ETL tool meets all of Intel's diverse business and project requirements, Intel IT performed a comprehensive paper study to shortlist the tools to consider, and then we evaluated the selected tools in test case scenarios. Our considerations for ETL tools included cost, performance, integration within our environment, ease of use, big data capabilities, metadata management, code migrations, support, and other factors.

SOLUTION

Intel IT's "start small" strategy enabled us to take an iterative, agile methodology-like approach. We worked with Intel IT BI teams and other groups to design and implement a 16-server, 192-core Hadoop platform, including all software and data integration solutions, in just five weeks.

To deliver the best ROI, the team was chartered to deliver a platform design and architecture that was right-sized for our current needs as well as the foreseeable future. The 5-week time frame included implementing and testing our platform design, bringing it into the enterprise, optimizing it, and putting it online. Choosing to build a solution sized specifically for our three intended use cases allowed us to move fast and minimize monetary risks. By making it scalable and expandable to meet evolving needs, we delivered a platform adaptable to a variety of other use cases in the future.

The order of the steps we took for selecting our Hadoop distribution, infrastructure, and data integration methods, and then implementing this solution (see Figure 2), has provided valuable "how to" procedural guidelines for future efforts.

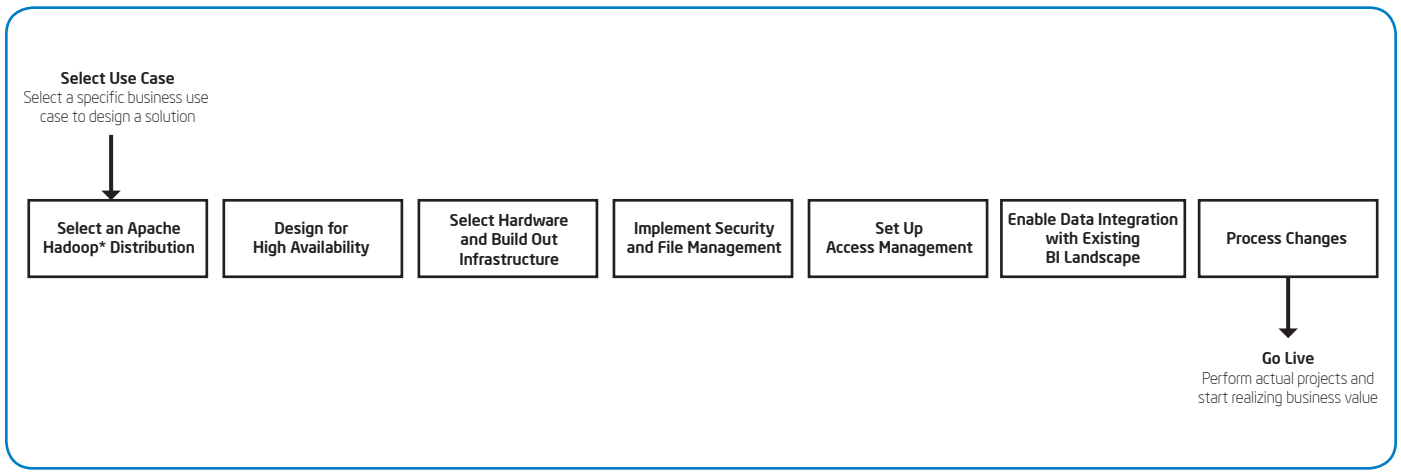


Figure 2. The steps Intel IT took to specify and implement our Apache Hadoop* distribution provide valuable "how to" procedural guidelines for future Hadoop implementations.

Selecting a Hadoop Distribution

We tested three Hadoop distributions. Only the Intel Distribution, which was designed to run on Intel® architecture-based servers, met all requirements with no major issues. In addition, it offers continuous innovation opportunities on an open platform designed to take advantage of the latest hardware-enhanced capabilities. Implemented on our hardware platform, the Intel Distribution provides the following:

- Performance boost over legacy infrastructure through Hadoop optimizations for Intel® Xeon® processors and 10 GbE networking
- Data confidentiality without performance penalty—testing at Intel demonstrated that the Intel® Advanced Encryption Standard New Instructions,[1] available in select Intel Xeon processors, accelerate encryption and decryption in Hadoop up to 19x
- Future-proof design with integration capabilities for next-generation analytics, visualization, and hardware solutions
- Enterprise-grade support and services available from Intel's suppliers

As shown in Figure 3, the Intel Distribution is a comprehensive solution that contains the full distribution from the Apache Hadoop open source project, along with MapReduce, Hadoop Distributed File System* (HDFS), and related components such as the Hive data warehouse infrastructure and Pig data flow language (see Table 1). The Intel Distribution also supports Apache Mahout, a machine learning library with MapReduce algorithms and the Intel® Graph Builder for Apache Hadoop* software. Solution elements are pre-integrated to simplify deployment and management, as well as enable faster time to market. These elements help minimize training and financial investments.

[1] No computer system can provide absolute security under all conditions. Built-in security features available on select Intel® Core™ processors may require additional software, hardware, services and/or an Internet connection. Results may vary depending upon configuration. Consult your PC manufacturer for more details.

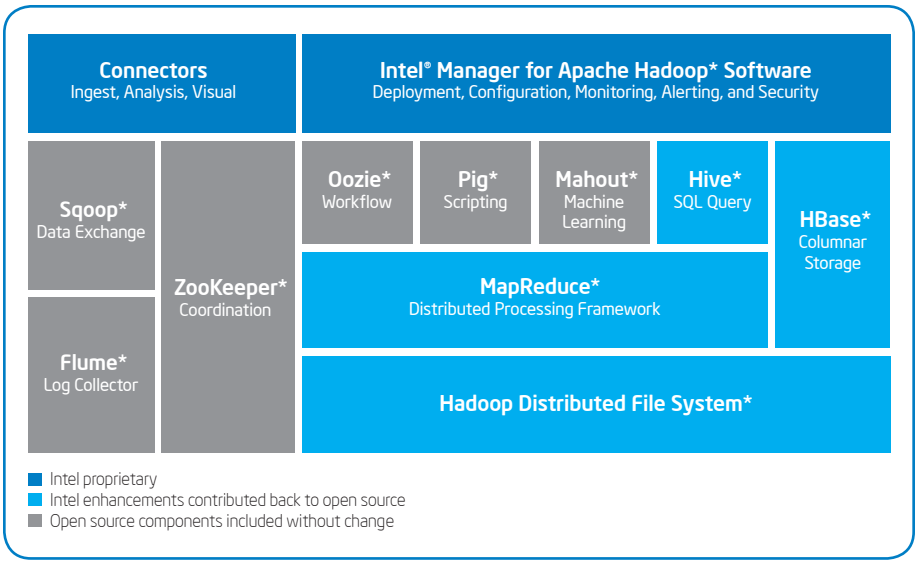


Figure 3. Intel® Distribution for Apache Hadoop* software provides a comprehensive solution for deploying, configuring, managing, and securing a big data platform.

Table 1. Functions of Big Data Platform Software Design Components

Component	Function
Intel® Manager for Apache Hadoop* Software	A management console that simplifies the deployment, configuration, and monitoring of an Apache Hadoop deployment. It automates the configuration of alerts and responses to unexpected events and failures within the Intel® Distribution for Apache Hadoop* software.
Hadoop Distributed File System* (HDFS)	A distributed, scalable, Java*-based file system providing a storage layer for large volumes of unstructured data.
MapReduce*	A software framework providing the Hadoop compute layer for the Map function that divides a query into multiple parts and processes data at the node level. This framework also provides the Reduce function for aggregating Map results to determine the answer to a query.
HBase*	A nonrelational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts, and deletes.
Hive*	A Hadoop-based data warehousing-like framework that allows users to write queries in a SQL-like language called HiveQL, which are then converted to MapReduce.
Oozie*	Oozie is a workflow scheduler system to manage Hadoop workflow jobs with actions that run MapReduce and Pig jobs.
Pig*	A Hadoop-based language that is relatively easy to learn and adept at extremely deep and long data pipelines, surmounting a limitation of SQL.
Mahout*	A data mining library that takes the most popular data mining algorithms for performing clustering, regression testing, and statistical modeling and implements them using the MapReduce model.
Flume*	A framework for populating Hadoop with data.
Sqoop*	A connectivity tool for moving data from non-Hadoop data stores, such as relational databases and data warehouses, into Hadoop.
ZooKeeper*	A centralized service for maintaining configuration information and naming, as well as providing distributed synchronization and group services.

Table 2. Basic Platform Design Components

Component	Implementation	Benefit
Server	16 two-socket servers based on the Intel® Xeon® processor E5-2600 product family (6-core)	Provides the best combination of performance, energy-efficiency, built-in capabilities, and cost-effectiveness, including Intel® Integrated I/O to help prevent data bottlenecks
RAM	96 GB per data node	Supports coexistence of HBase* and MapReduce*
Drives	25 TB HDFS* raw storage per data node	Fulfills the deep storage requirements of a big data platform
Network Adapters	10 GbE converged network adapters	Supplies the increased bandwidth critical to importing and replicating large data sets across servers
Switches	2x 48-port 10 GbE	Enables high-bandwidth connectivity for enterprise-class performance

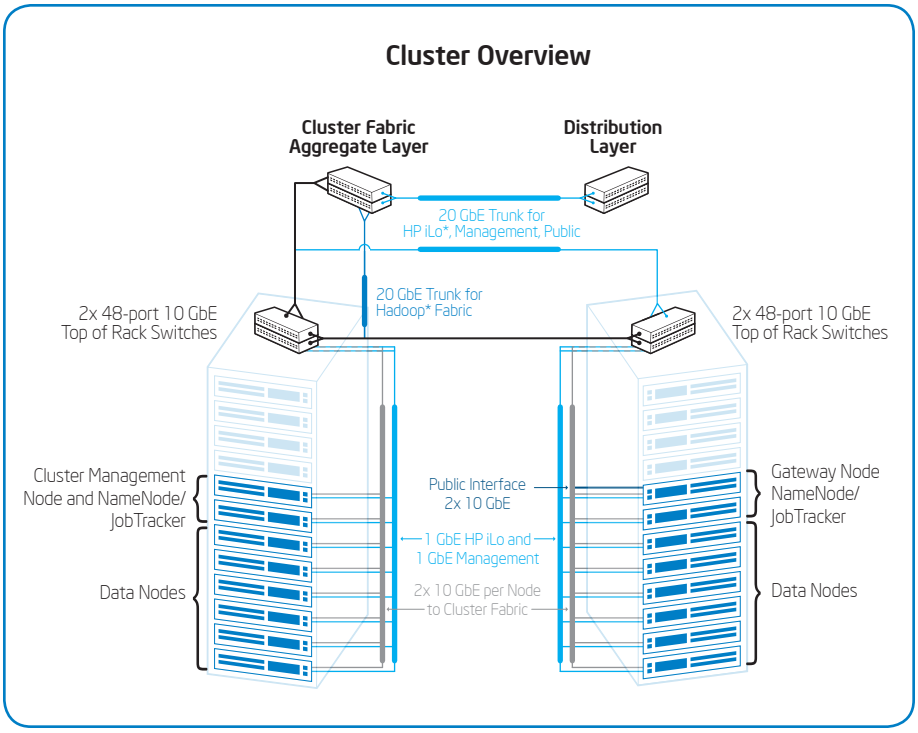


Figure 4. Intel IT's big data platform hardware network and rack design occupies two racks in our production center and enables three-way replication with approximately 100 terabytes of targeted usable storage space.

Designing for High Availability

We implemented a high-availability design for the network and NameNode, the critical centerpiece of the HDFS that keeps the directory tree of all files. Implementing high availability at the NameNode level is essential to strengthening the platform against failure.

For our Hadoop installation, which spans multiple racks, we wanted to ensure that replicas of data exist on different racks. This design protects against the loss of a switch, which would render portions of the data unavailable due to all the replicas being underneath it. In our design, HDFS components are rack-aware and we use a three-way replication of HDFS storage. This high-availability design provides important protection against unexpected downtime if NameNode or part of the network goes down.

Selecting and Building Out the Infrastructure (Servers, Network, and Rack Design)

To help select a Hadoop distribution, the team collaborated and designed the hardware infrastructure and network. They also sized the environment. For cost-effectiveness and scalability in compute- and storage-intensive applications, they selected a structure of clusters with multiple nodes using servers based on the Intel® Xeon® processor E5 family (see Table 2). These servers are connected with 480 Gbps cluster fabric bandwidth. The servers reside in two racks in our production center and are equipped to provide 300 terabytes (TB) of HDFS storage capacity, enabling three-way replication with approximately 100 TB of targeted usable storage space (see Figure 4).

PLATFORM ENTERPRISE MANAGEMENT DESIGN

To run, manage, and monitor Intel's internal big data cluster, Intel IT uses CentOS* 6.3, an enterprise-class Linux* distribution derived from sources freely provided to the public. Important management elements include third-party enterprise configuration management software, open source monitoring and performance management tools (Nagios* and Ganglia*), third-party management utilities, and third-party authentication services.

In addition, we use a Kernel-based Virtual Machine (KVM). This open source, full virtualization solution for Linux on x86 hardware contains valuable virtualization extensions such as Intel® Virtualization Technology[2] (Intel® VT). We use KVM on the gateway to perform process isolation between enterprise software components and core Hadoop ecosystem components to increase the reliability and manageability of the platform.

Implementing Security and File Management

Security is a major requirement for our environment since some of our data consists of customer records and other confidential data. Our security plan uses a secure cluster reference design calling for highly secure third-party 10 Gb rack switches to isolate Hadoop traffic from our shared access and distribution layer, as well as provide flexibility in network design.

A key element of this design is a dedicated gateway server through which all connections to and interactions with our cluster must pass. Since the cluster environment is behind our firewall, we expose only this one gateway server, making it the main interface. All users and applications accessing the Hadoop environment have to connect through the gateway server.

This gateway server has other purposes as well, such as job launching and hosting the KVM. To address potential bandwidth limitations of this design and improve the resiliency of the network, we combine a Multi-Chassis Link Aggregation Group (M-LAG) at the aggregate and top of rack switches with dual 10 GbE bonded network interface cards and network isolation for the cluster fabric.[3] We also enable load balancing across multiple switches for capacity expansion.

Since our Hadoop platform is a multitenancy environment, we also developed a system to restrict the access of data from one project to the other. This prevents one user from accidentally deleting another user's data, as well as viewing data without authorization. To restrict access to single projects, we designed and set up dedicated folders for each project on the local gateway server and HDFS. We control access by granting permission to a project group to project folders. The project group is assigned to an enterprise access management (EAM) entitlement through our access management system.

Setting Up Access Management

For access management to operate smoothly in conjunction with other applications running on the CentOS operating system, we needed a way to reduce the maintenance overhead of managing users, groups, passwords, permissions, request processing, tracking, and user auditing. Our solution was to manage and integrate users with Microsoft Active Directory* and an existing EAM tool developed by Intel IT. This solution allows us to authenticate the user through Active Directory and manage the user's access through the EAM tool. The solution enables users to access the system through the same IDs and passwords as their Active Directory accounts.

Our access management system automatically creates a user through an impersonation of the user's Active Directory account. Setting up this convenient feature required integration with Active Directory using a third-party tool and integration with the EAM through the Intel IT tool.

Enabling Data Integration with Existing Business Intelligence Landscape

Our comprehensive paper study and PoC comparison of shortlisted tools revealed the need for multiple data integration tools. We selected a group of ETL tools and provided prescriptive guidance on which tool is best for which job. Each tool is defined and supported to address a specific business need (see Figure 5).

[2] Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, and virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit www.intel.com/go/virtualization

[3] A LAG is a method of multiplexing over multiple Ethernet links to increase bandwidth and provide redundancy. An M-LAG is a type of LAG with constituent ports that terminate on separate chassis.

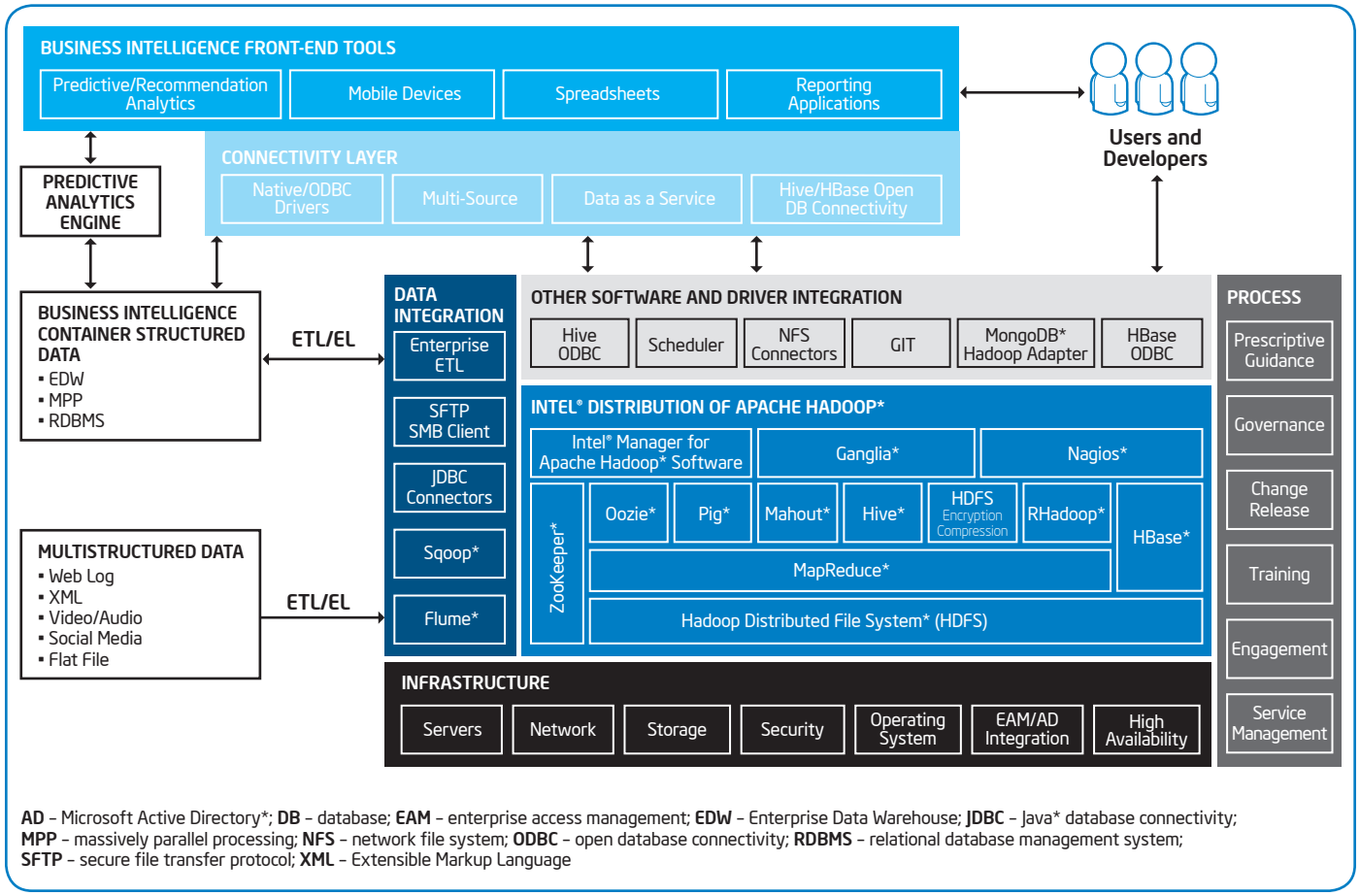
Having clear positioning on each tool and defining the tool decision flow helps our BI and big data project teams make the right decisions.

- **Flume.** Specified for scenarios requiring the collection, aggregation, and writing to HDFS of streaming log data from event logs, system logs, web clicks, and similar sources.
- **Sqoop.** Recommended for relational database management-to-HDFS data movement and vice versa. Decision boxes are created in the decision flow to avoid its limitations with certain data types.
- **Command-line HDFS "put file."** Employed for use in simple, straight file loads where data is already being delivered to the big data environment and requires no joins or transformations. The put file is supported through a script/command-line interface only.
- **Enterprise ETL tool.** Used for complex use cases that other data ingestion ecosystems, such as Sqoop and Flume, do not support. We fully support such ETL tools for scenarios such as business groups needing complex processing and complicated transformation work up front. An enterprise ETL tool can prepare a data file that can be used to load the data into HDFS using the put file command.
- **Capacity scheduler.** Deployed to prevent the impact of capacity on our multitenancy Hadoop environment. We use a capacity scheduler to manage the workload, allocating certain map and reduce slots for each project. Running batch is also a critical solution piece. Through integration with our enterprise scheduler tool, we enable users to trigger a job based on a time, an event, or a combination of both.

Making Process Changes

Implementing a multitenancy cluster compromises control over compute resources to some degree. To counter this potential issue, we established working procedures and control processes that allow us to better assign and prioritize computer resources commensurate with job priority. We also instituted a review process to determine when a project needs our Hadoop platform or is better suited to another of our BI analytics solutions.

For support, issue escalation, and service requests, we aligned with our current IT direction, using a software-as-a-service application that provides service management for issue escalation and request service. In addition, we are currently working on guidance documentation for developers and project managers, training documentation and classes,



AD – Microsoft Active Directory*; **DB** – database; **EAM** – enterprise access management; **EDW** – Enterprise Data Warehouse; **JDBC** – Java* database connectivity; **MPP** – massively parallel processing; **NFS** – network file system; **ODBC** – open database connectivity; **RDBMS** – relational database management system; **SFTP** – secure file transfer protocol; **XML** – Extensible Markup Language

Figure 5. For data integration with existing business intelligence tools, Intel IT's big data platform uses a variety of tools, each supporting a specific business need.

change and release processes for migration to the platform, and a governance process for code reviews and monitoring adherence to our standards and best practices.

RESULTS

Intel's first internal big data compute-intensive production platform with the Intel Distribution of Hadoop launched at the end of 2012 (see Figure 6). This platform is already delivering value in our first three use cases, helping us to identify new opportunities as well as reduce IT costs and enabling new product offerings.

Using our Hadoop implementation, Intel IT customers can now use the power of big data on a robust and scalable enterprise-ready platform based on the Intel Distribution that is integrated across our BI capability landscape. Our three use cases are fully operational, and a number of new uses cases are in various stages of development.

Our internal big data platform expands our BI data container strategy, enabling the following:

- Structured and multistructured analytic data use cases
- Platform design and architecture right-sized for today and the immediate future
- Scalable and expandable design capable of meeting evolving needs

Three Big Data Use Cases and Their Estimated Value

By designing the platform for three specific use cases, Intel IT delivered nearly immediate value to the organization. Each use case is delivering or has the potential to deliver BI results worth millions of dollars to Intel.

CONTEXTUAL RECOMMENDATION ENGINE

Our big data platform enables a generic, reusable context-aware recommendation engine and analytic capabilities for a mobile location-based service. This service combines new, intelligent context-aware capabilities—including an algorithm for collaborative filtering—to help users find products, information, and services with map management technologies. The recommendation engine design is already being used for additional uses cases. In sales, we are using it to help decide what products should be offered to which resellers to maximize their sales and ours. Our recommendation engine may be offered in the future as a paid service.

LOG INFORMATION ANALYTICS FOR INCIDENT PREDICTION

Our big data platform is helping to identify and correlate potential issues opened with IT. We are tracking the event log data that precedes incident data and using linear regression and time-series analysis to predict future behavior and impact. Such predictive analytics help reduce incidents and the impact on users and IT, decreasing IT operations, support costs, and time-to-resolution with proactive and predictive issue and symptom analysis.

We estimate that using our big data platform for incident prediction will provide a 10- to 30-percent reduction in new incidents at an estimated IT cost avoidance of USD 4 million over two years.

WEB ANALYTICS FOR CUSTOMER INSIGHT

We are landing and ingesting web data in Hadoop and integrating this external data with internal customer data to provide customer and network usage analytics for Intel.com and customer advertising. These web analytics give our sales and marketing groups the ability to perform deep analysis of web usage data for marketing or content navigation purposes. These analytics also provide the means to predict and adjust product positioning and pricing based on response to marketing campaigns, as well as improve the efficiency of the Intel supply chain.

Intel sales and marketing groups estimate the ROI for web analytics in demand generation will be USD 10 million by 2014. Intel predicts that the smart analytics applied to the Intel supply chain will deliver up to USD 20 million in value in 2013 by improving availability of the right products at the right time and helping maintain the proper inventory levels for each region.

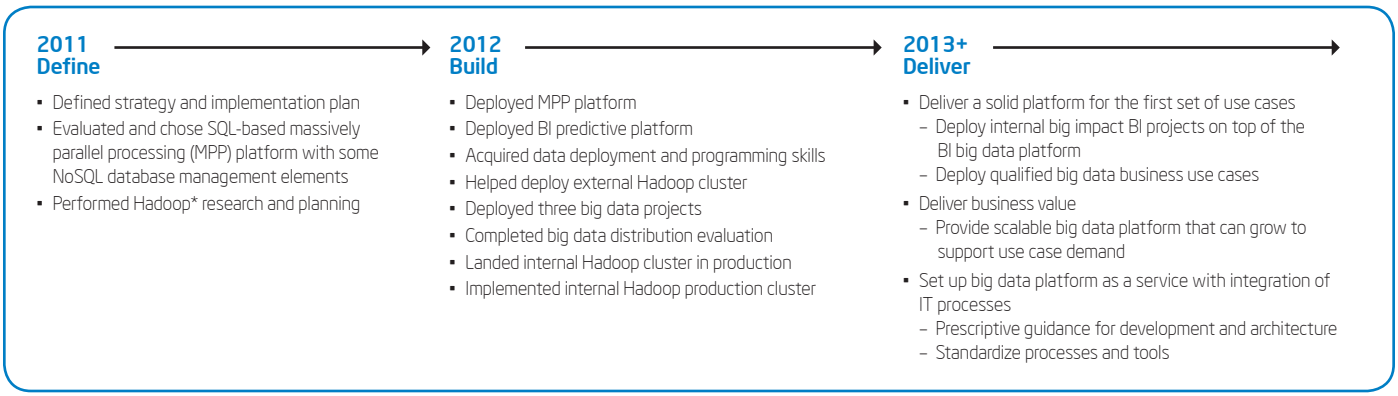


Figure 6. This timeline provides an overview of the progress Intel IT has made in bringing a big data platform into the organization's business intelligence (BI) capabilities.

CONCLUSION

Based on our experience, Intel IT believes obtaining optimal results from a Hadoop implementation begins with carefully choosing the most advantageous hardware and software. Fine-tuning the environment to achieve the highest ROI calls for an in-depth analysis of available Hadoop distributions, cost-efficient infrastructure choices, and careful integration with the existing BI environment to ensure efficient data transfer, strong security, and high availability.

Our strategy of starting small and using an iterative approach enabled us to cost-effectively develop a big data platform for our three targeted use cases that can be ultimately scaled to handle a diverse range of additional use cases. Through this process, we have also been able to develop a number of best practices and how-to procedural guidelines that will continue to guide us as we expand our big data platform and build additional big data platforms in the future.

FOR MORE INFORMATION

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

- "Enabling Big Data Solutions with Centralized Data Management"
- "Integrating Apache Hadoop* into Intel's Big Data Environment"
- "Mining Big Data in the Enterprise for Better Business Intelligence"
- "Using a Multiple Data Warehouse Strategy to Improve BI Analytics"

CONTRIBUTOR

Moty Fania

ACRONYMS

BI	business intelligence
EAM	enterprise access management
EL	extract-load
ETL	extract-transform-load
HDFS	Hadoop Distributed File System
KVM	Kernel-based Virtual Machine
MPP	massively parallel processing
M-LAG	Multi-Chassis Link Aggregation Group
NoSQL	not-only SQL
ROI	return on investment
SQL	structured query language
TB	terabyte

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