(intel)

Predictive Analytics and Interactive Queries on Big Data

WRITERS

Moty Fania, Principal Engineer,
Big Data/Advanced Analytics, Intel IT

Parviz Peiravi, Principal Architect,
Intel Enterprise Solution Sales

Ajay Chandramouly,
Big Data Domain Owner, Intel IT

Chandhu Yalla,
Big Data Engineering Manager, Intel IT

CONTRIBUTORS

Sonja Sandeen,
Big Data Product Manager, Intel IT

Yatish Goel,
BI Technical Integrator, Intel IT

Nghia Ngo,
Big Data Capability Engineer, Intel IT

Darin Watson,
Platform Engineer, Intel IT

Jeff Ruby,
Datacenter Software Planner,
Intel Software & Services

Dan Fineberg,
Datacenter Software Product Marketing,
Intel Software & Services

Abstract

The business potential of big data analysis is enormous across virtually every business sector. The Intel IT organization has implemented use cases delivering hundreds of millions of dollars in business value. This paper discusses a few of those use cases and the technologies and strategies that make them possible. It also defines the architecture we use for big data analysis and provides an in-depth look at one of the most important components—an Apache Hadoop* cluster for storing and managing large volumes of poly-structured data.

The Big Data Opportunity

Companies today store large volumes of diverse data from web logs, click streams, sensors, and many other sources. The insights hidden within this poly-structured "big data" hold tremendous business value. At Intel, we have come to think of the development model for big data analysis in terms of "5-6-10." That is, a team of five people skilled in big data analysis can deliver in six months up to USD 10 million or more in returns. From our experience, that is a conservative estimate. We have big data use cases in production today delivering as much as USD 100 million.

To extract value, you must often integrate unstructured data with structured data from your core business applications and then query or analyze the combined data set. Higher value can often be achieved using interactive business intelligence (BI) tools. When data analysts and business users have direct access to big data using BI tools that fit their different levels of expertise, they can work independently to answer questions and find hidden relationships. Hundreds of users with diverse needs and expertise can generate and share insights to supplement the more ambitious efforts of centralized BI development teams.

In our experience, the highest value is achieved through predictive analytics, which apply advanced techniques such as machine learning and real-time regression analysis to predict future events and drive decisions or actions, potentially in near real-time.

Table of Contents

Laying the Foundation for Big Data

Key Technologies

Interactive queries and predictive analytics can run on a conventional data warehouse (DW) built on a traditional relational database management system (RDBMS). However, conventional DWs are not designed to handle the volume, variety, or velocity of today's big data. A number of new technologies address that challenge.

- **Apache Hadoop* software** is a cost-effective, massively-scalable platform for analyzing poly-structured data. It can store and process petabytes of data, including all the data types that don't fit into your RDBMS. Hadoop runs on clusters of commodity servers and storage drives—up to thousands of servers—so you can scale performance and capacity at a cost-per-terabyte that is far below that of a traditional DW. Currently, Hadoop supports batch processing only, although much work is being done, by Intel and others, to enable interactive queries and real-time analysis directly on data stored in Hadoop.

- **Not only SQL (NoSQL) databases** relax the constraints of a traditional RDBMS to deliver higher performance and scalability. NoSQL databases can extend the capabilities of Hadoop clusters by providing low-latency object retrieval or other DW-like functionality. Like Hadoop, NoSQL databases scale cost-effectively on clusters of commodity servers and drives. They can run on top of a Hadoop cluster or on a separate cluster. More than a hundred different NoSQL databases are available, offering a wide variety of capabilities to address specific use case requirements.

- **Massively parallel-processing (MPP) appliances** extend the capabilities of RDBMS-based data warehouses. You can deploy a blade-based MPP appliance quickly, integrate it with your existing BI environment, and scale it as your needs grow by adding more blades. These systems can store and process petabytes of structured data.

- **In-memory databases** dramatically improve performance by eliminating most of the data access latencies associated with shuttling data back and forth between storage systems and server processors. In-memory databases are available as an option on some of today's MPP appliances to provide real-time performance for the most demanding applications. However, due to the requirement that all data be held in main memory, in-memory solutions tend to be costly, and are appropriate only for, such as complex event processing (CEP), that require the performance and justify the expense.

Data Warehouses Strategies

A high-value solution for big data analysis could include just two DWs: an MPP appliance for structured data and a Hadoop cluster for poly-structured data. Connect the two with a high-speed data loader and you can move data quickly between them to meet a wide range of business needs (Figure 1).

Businesses with more complex data and analytics requirements can use the technologies described above in different combinations to power a variety of DWs, each optimized for particular kinds of data and analyses. Intel IT uses five different DW "containers" to provide a dynamic range of interactive query and big predictive analytics capabilities (Table 1). This multi-DW strategy provides a great deal of flexibility. It is possible to combine data from many different sources and perform both traditional BI and big data analyses on the platforms that address requirements in the most cost-effective way.

In the following sections, we provide a deeper look into what you can do with big data. We begin by describing a few specific use cases, and then we explore the requirements for implementing advanced usage models, including interactive queries and predictive analytics. Finally, we describe the architectures of our most important big data platforms and provide links to additional resources that can help you get started.
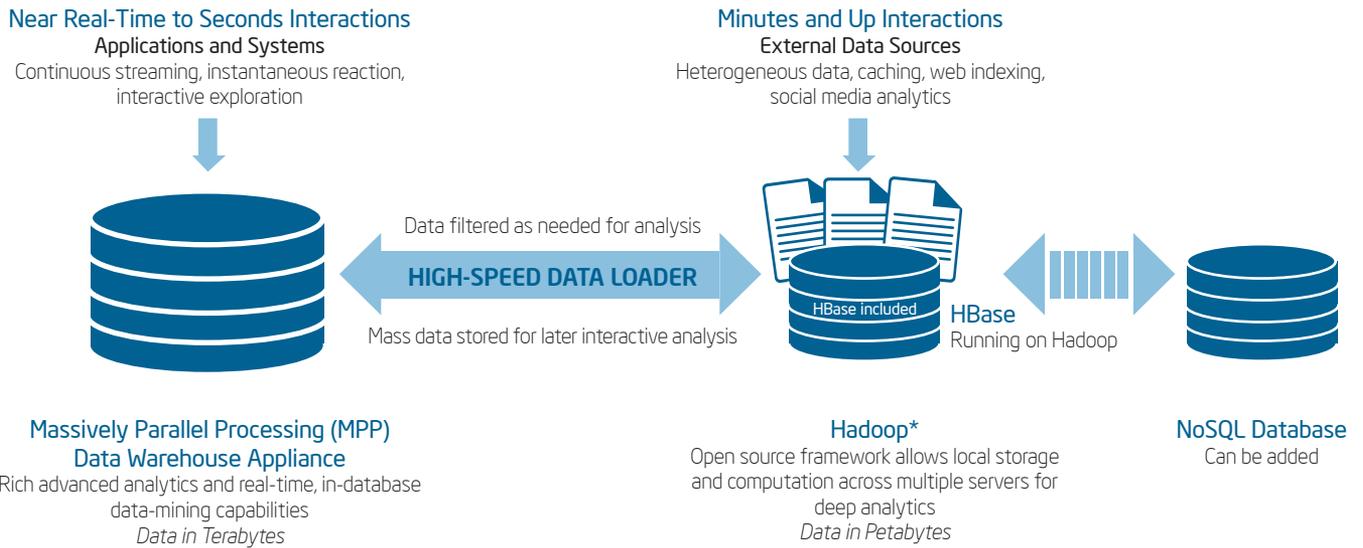
Figure 1. A flexible platform for big data analysis could include just two components: a massively-parallel processing data warehouse appliance and a cluster of servers running Apache Hadoop*. HBase and other NoSQL databases can be added to support specialized, low-latency requirements. (HBase is included with Hadoop and is often deployed on the same cluster.)

	ENTERPRISE DATA WAREHOUSE (EDW)	APACHE HADOOP* PLATFORM	MASSIVE PARALLEL PLATFORM	NO SQL CONTAINER	IN-MEMORY DW (IN PROGRESS)
Positioning and Intended Use	Traditional BI, highly-shared enterprise data requiring cross-organizational integration	LOB data, EDW archive, unstructured data	Advanced analytics, high-volume, LOB data; some shared data	Specialized container for unstructured, high concurrency and parallelism	Enterprise shared data based on combined OLTP/DW workloads
Relative Performance	Best for structured data	Best for unstructured data	Better	Very Good to Excellent (depends on use case)	Excellent
Agility Factor	Slower (highly governed)	Highly agile (low governance)	Agile (medium governance)	Depends on the platform	Depends on OLTP source
Data Type/Normalization	Structure, denormalized	Raw/structured, unstructured data	Structured	Unstructured, denormalized	Structured hybrid or column
Real-Time versus Batch Analytics	Batch/near real-time	Batch	Mix streaming/batch	Near real-time/batch	Real-time/batch
Historical Data Horizon	Long-term horizon; >5 years	Short to very long-term horizon; <3 years to >10 years	Mid-term horizon; 1 to 3 years	Short to very long-term horizon; <3 years to >10 years	Very short-term horizon; 6 months to 1 year
Supported Scalability	< 300 TB	> 1 PB	> 500 TB	> 1 PB	< 20 TB
Summarized Data	Raw and summarized	Mostly raw	Raw and summarized	Raw and summarized	Very little to high
BI Capabilities Usage	• Formatted reports • Ad hoc • Dashboards • OLAP/MOLAP	• Text parsing • Data mining • Temporary data • Web • Sensor • Archive • Predictive Analytics	• Data mining • Ad hoc • Predictive analytics	• Fast retrieval • Ad hoc analysis • Predictive analytics • Massive concurrency • Formatted reports	• Predictive analytics • Formatted reports • Dashboards • HOLAP/OLAP

Table 1. Intel IT uses a multi-warehouse strategy to provide more dynamic range of BI and big data analysis capabilities.
Key: LOB = Line of Business; OLTP = On-Line Transaction Processing; MOLAP = Multidimensional OLAP; HOLAP = Hybrid OLAP

Proven Big Data Use Cases

In developing our big data capability, the IT organization worked with Intel's business units to identify use cases that promised high returns with relatively light effort. Each use case described below provides an example of how you might apply big data analysis to achieve your business goals. The real-time recommendation system, in particular, offers potentially high value across virtually all industry sectors. It also has relatively sophisticated infrastructure requirements, so we provide more detail on that use case later in this paper.

Use Case A: Design Validation

Microprocessors are some of the most complex structures ever built. At Intel, products can take years to design, test, manufacture, and bring to market. Once we have completed a design and run through its first production trial, the post-silicon validation process can take several quarters of dedicated tests running in various labs by various teams and tools. All of this testing, analysis, and fine tuning must be done before we can send orders through our high-volume factories.

With billions of transistors on each chip, efficient tests are essential to provide effective bug handling and high-quality products, while avoiding unnecessary test redundancy. By applying interactive queries and predictive analytics to hundreds of millions of rows of structured and unstructured data, we now gain greater insight and make more accurate predictions to quantify how mature the process is, optimize the way bugs are handled, and determine whether more testing is necessary. For a recent product, we eliminated 36 percent of our test content without compromising completeness. As we apply this approach more broadly, we expect these efforts to reduce post-silicon validation time by 25 percent.

Use Case B: Market Intelligence

Like any global company, Intel gains huge strategic and tactical benefits when planners correctly anticipate global market demand and supply chain risks. Intel can now mine and analyze enormous data sets—including weather trends, global economic indicators, discussion forums, news sites, social networks, wikis, tweets, and blogs—to gain insights and improve the accuracy and granularity of forecasts one month, six months, and even years into the future. We also can test "what if?" scenarios to identify the likely impacts of possible events and to determine the best courses of action. As Intel fully deploys these capabilities, they will help the corporation plan more effectively and make better decisions based on deeper knowledge and more accurate predictions.

Use Case C: Real-Time Recommendation System

Real-time recommendation systems were pioneered by early Internet companies, such as Amazon, Google, and Netflix. This use case at Intel will aim to increase sales and improve the customer experience, and also to provide internal users with faster and more efficient access to information and resources.

A real-time recommendation system requires high-speed predictive analytics. The predictive algorithms provide recommendations based on explicit, implicit, and historical data, which might include data from enterprise resource planning (ERP) and customer relationship management (CRM) databases, Internet click-stream logs, and social networking posts. In some solutions implemented by Intel, variables such as location, time-of-day, season, weather, customer input, and very recent customer behavior are integrated into the analysis to provide highly personalized, context-aware recommendations.

Intel generates recommendations using a number of different models. One approach analyzes each individual's past behavior to predict future actions. Another matches the user with other users that have similar profiles, and assumes the user will behave in a similar manner. Yet another approach applies domain-expertise to generate a recommendation.

In general, more data delivers better results, and the best results often are achieved by combining multiple analytical models. For example, if an individual asks, "Why won't my car start?" the best response might take into account the make and model of the car, weather and road conditions, maintenance records, personal history, and the kinds of problems encountered by other individuals in similar situations. The recommendation system might also identify nearby service options and rate them based on convenience and customer reviews.

In any recommendation system, some of the data will already be stored in a DW. Other data will be generated in real-time by the user, by the interaction with the user, and perhaps by sensors or other applications. In most cases, time to results is a critical success factor.

Performing Advanced Analysis on Big Data

The Interactive Query Usage Model

Although Hadoop is fast and massively scalable, data processing is performed in batch mode, only. Source data and final results reside within the Hadoop Distributed File System (HDFS). Ingesting new source data and extracting results are separate steps that require initiating and completing additional jobs. Because of this limitation, Hadoop does not directly support interactive queries.

Hadoop software includes a module, Apache Hive*, designed to enable queries on big data. Hive facilitates querying large datasets in distributed storage by projecting structure onto the data and using an SQL-like language called HiveQL. Hive also allows map/reduce programmers to plug in custom mappers and reducers as an alternative to HiveQL. You can use Hive and Hadoop's NoSQL columnar database, HBase*, to perform complex queries, including ad hoc queries, in your Hadoop environment. Work is underway by Intel and others to deliver integrated support for true, interactive queries on Hadoop using standard Structured Query Language (SQL). Until these solutions become available, we recommend loading data into your enterprise DW (EDW) for interactive querying.

Hadoop excels as a high-speed, massively-scalable extract, transform, and load (ETL) solution, so it is relatively easy to add structure to your unstructured data as you move it into your EDW. You can then apply traditional BI and analytics tools to support interactive queries and other advanced needs. That approach also gives you the flexibility to merge diverse structured and unstructured data sets, so you can run queries against all relevant data.

The Predictive Analytics Usage Model

Intel Advanced Analytics developers use another Hadoop software component, the Apache Mahout* machine learning library, to create predictive algorithms that can run directly on Hadoop. However, since Hadoop processes data in batch mode, additional components are needed to integrate predictive analytics into real-time business processes.

A real-time recommendation engine, for example, uses predictive analytics to provide recommendations based on both structured and unstructured data. The extreme requirements of this use case—deep, predictive analytics with sub-second response times—are met using a two-layer, offline and online architecture. The offline component acts on historical data using batch-oriented processing. The intermediate results from batch analyses are integrated with real-time data from the user interaction and other sources to refine the recommendation based on context-specific data.

A typical workflow goes like this: A traditional RDBMS-based data warehouse, such as our EDW, performs data integration and pre-processing of the structured data as a batch process, and then transfers the data to the Hadoop cluster. The Hadoop cluster integrates the pre-processed data with relevant unstructured data, and applies predictive models, again as a batch job.

The results of the analyses are loaded into the online component of the recommendation system, a NoSQL database. The NoSQL database supports low-latency lookups, which allows results to be queried interactively. Intel uses a number of NoSQL databases in its big data environment, including Cassandra*, HBase, and MongoDB*. Like Hadoop, these are all open-source software applications that run on commodity servers and storage drives to provide high performance and massive scalability at relatively low cost.

For this particular implementation, we used Cassandra. Cassandra automatically replicates data across multiple servers, racks, and even data centers. Unlike Hadoop, it uses a gossip protocol rather than a centralized Name Node to manage transactions, so there is no single point of failure. Depending on how replication is configured, multiple servers, or even a whole data center, could fail without losing data or interrupting service. These features make it a good fit for the online component of a mission-critical recommendation system.

When a user engages with the system, the results of the relevant analysis are retrieved from Cassandra. A Java computation layer then integrates the real-time data, performs the final calculation, and responds to the user's request.

Reference Architecture

In the following sections, we describe the infrastructure for four of our five DW containers, with a particularly strong focus on the Hadoop cluster. We don't describe the independent DW and data marts, because those systems are relatively diverse, and they target more traditional requirements.

An EDW for Enterprise-Wide Structured Data

The EDW remains an essential component of Intel's analytics strategy for highly-shared enterprise data and for BI solutions that require cross-application views or business data integration. The EDW also operates as a data hub for downstream reporting and data mart solutions. The EDW supports traditional analytical models acting on structured data, including online analytical processing (OLAP), multidimensional analytical processing (MOLAP), data mining, and ad hoc queries. The BI development team uses the EDW to generate dashboards and executive reports, and to answer specific questions. Most queries run as batch jobs, but the system also supports interactive queries and near-real time analysis.

An EDW can be deployed on traditional symmetric multiprocessing (SMP) servers or on MPP architectures, depending on data volumes and analytic requirements. Intel IT uses a third-party MPP appliance built with a modular design that can scale up to petabytes of data. Individual modules are configured with the Intel® Xeon® processor E5 product family, which provides excellent performance for latency-sensitive analytics. The system is designed to accommodate future Intel Xeon processors, as they become available.

The current configuration can hold about 300 TB of data using a combination of solid-state drives (SSDs) and hard disk drives (HDDs). Data placement is automated so that frequently used data is stored on the faster SSDs and less frequently used

data is stored on the larger capacity HDDs. This automated tiering is transparent to applications, and provides substantial improvements in data throughput and latency compared with using only HDDs. The system delivers high availability, with a number of hot-standby nodes and a dual-InfiniBand network fabric for low-latency, high-bandwidth, fault-tolerant node-to-node communications.

An MPP Data Warehouse Appliance for High-Volume Structured Data

Intel IT uses another third-party MPP appliance for predictive analytics and operational reporting on large volumes of structured and semi-structured data. The system functions as a bridging platform between traditional analytic models and new big data models. It supports SQL-based queries, yet also integrates with Hadoop and supports applications based on MapReduce, which is part of the Hadoop framework. The data in this DW comes from many sources throughout the corporation. Because it does not store enterprise-wide shared data, governance is somewhat less strict, affording greater agility for developing and implementing new use cases.

The hardware infrastructure for this MPP DW combines blades based on multi-core Intel Xeon processors with proprietary data filtering components that help to speed data throughput. Each blade has multiple HDDs for fast, parallel data transmission. Altogether, the system includes a total of 96 cores (192 threads), 16 GB of RAM and 96 TB of disk storage. Blades are connected using a 10 Gigabit Ethernet (10GbE) network for fast data sharing within the warehouse. A separate 1GbE network provides access to the system for users and another supports data backups.

An In-Memory Database for Real-Time Analysis of Streaming Data Sets

Intel IT has performed extensive tests and believes in-memory analytics will ultimately become a valuable addition to Intel's BI

capabilities. This type of solution works especially well when business groups require extreme query performance and sub-second update latency. It is a useful platform for Complex Event Processing (CEP) and other use cases that require fast processing of streaming data.

Due to the unsurpassed speed of in-memory analysis, we believe this kind of DW may ultimately deliver the highest business value. However, cost can be a limiting factor, due mostly to the condition that all data must fit into main memory. It is important to have specific use cases that not only require the extreme performance, but also justify the expense. Based on typical data compression rates of about 5x, we currently recommend systems that support a maximum of about 10 to 20 terabytes.

A Hadoop Cluster for Massively-Scalable Big Data Analysis

Hadoop has become the de facto standard for handling massive amounts of poly-structured data, and it is a foundational component of Intel IT's big data strategy. Because most IT organizations are unfamiliar with Hadoop—and because Hadoop clusters are typically designed and built in-house—we provide more detail on configuring and deploying Hadoop in this section.

With its distributed, parallel processing capabilities, a Hadoop cluster can rapidly ingest, store, and process petabytes of poly-structured data. Hadoop software coordinates local storage and computation across tens, hundreds, or even thousands of servers. Each server stores and processes a subset of the data. Since applications execute in parallel, performance and capacity scale with each server added to the cluster.

Hadoop is available as free, open-source software and also as supported distributions from a number of value-added vendors. Intel IT initially evaluated three distributions, including the Intel® Distribution for Apache Hadoop* software (Intel® Distribution).

Issues to consider included performance, system and data security, manageability, high-availability, and the degree of support provided for advanced analytics.

The Intel Distribution was selected for a number of reasons. It includes the full distribution from the Apache Hadoop open source project, along with MapReduce, HDFS, and related components (Table 2), which helps us integrate Hadoop more flexibly into our larger analytics environment (Figure 2). Solution elements are pre-integrated to simplify management and deployment, which helps to reduce training and financial investments.

The main advantages of the Intel Distribution include:

- **Built-in security features.** The Intel Distribution provides integrated support for access controls, data encryption, and secure multi-tenancy. The Hadoop cluster supports multiple business units and contains sensitive business data, so these features are critical. We implemented role-based access controls with cell-level granularity using HBase, which runs on top of the Hadoop Distributed Filesystem (HDFS). The Intel Distribution also supports Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI), which provides hardware-assistance that improves encryption performance by up to 5.3x and decryption performance by up to 19.8x.[1] This will give Intel IT greater latitude for storing and analyzing sensitive data in the Hadoop environment.

- **High availability.** The Intel Distribution supports multi-site scalability and adaptive data replication using HBase and HDFS. Because Intel IT uses its big data platform to support important business functions across widely distributed business sites, these features are important now, and will become increasingly important as the Hadoop footprint expands.

- **Advanced analytics and data visualization.** A DW is only as good as the analytics tools it supports. The Intel Distribution includes Rhadoop, an integrated language connector for R, an advanced statistical language that has become widely popular among data scientists. It also includes the Mahout machine learning and data mining library and Intel® Graph Builder for Apache Hadoop software, a library for constructing graphics from large data sets. These tools make it easier to develop sophisticated applications to extract value from big data.

- **Simplified Management.** Intel® Manager for Apache Hadoop provides a management console that simplifies the deployment, configuration, tuning, monitoring, and security of the cluster. Managing a distributed architecture can be complex, and these tools help us reduce administrative costs and improve overall performance and availability.

COMPONENT	PURPOSE
Intel® Manager for Hadoop* Software	A management console that simplifies deployment, configuration, and monitoring. It also automates the configuration of alerts and responses to unexpected events and failures within the Hadoop cluster.
Hadoop Distributed File System (HDFS)	A distributed, scalable, Java*-based file system that provides a storage layer for large volumes of unstructured data.
MapReduce	A software framework that simplifies the development and execution of highly parallel applications. The Map function divides a query into multiple parts and processes data at the node level. The Reduce function aggregates Map results to determine the answer to the query.
Hive	A Hadoop-based data warehouse-like framework that allows users to write queries in a SQL-like language called HiveQL, from which they are converted to MapReduce.
Pig	A Hadoop-based language that is relatively easy to learn and adept at very deep, very long data pipelines, surmounting a limitation of SQL.
HBase	A non-relational database that allows for low-latency, quick lookups in Hadoop. HBase adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts, and deletes.
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.
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 Map Reduce model.
ZooKeeper	A centralized service for maintaining configuration information and naming, as well as providing distributed synchronization and group services.

Table 2. Intel® Distribution for Apache Hadoop* software—included components.

The hardware architecture for the Hadoop cluster (Figure 3, Table 3) is designed to deliver cost-effective performance and scalability for both compute-intensive and storage-intensive use cases. Two-socket servers based on the Intel Xeon processor E5 product family provide an efficient, high-performing server platform for these requirements. With up to 8-cores, 16-threads, and 30 MB of last level cache, these processors are well suited to the data-intensive, highly-parallel workloads of MapReduce applications.

A number of features built into the Intel Xeon processor E5 family are particularly advantageous for Hadoop. One feature is Intel® Integrated I/O, which reduces I/O latency by up to 32 percent and increases I/O bandwidth by as much as 2x.[2,3] Another is Intel® Data Direct I/O Technology (DDIO), which allows Intel® Ethernet adapters to communicate directly with processor cache, rather than only with main memory. This feature delivers more I/O bandwidth and lower latency, which is particularly beneficial when processing large data sets.

Each server is configured with 96 GB of RAM, which is enough that we can run MapReduce and HBase simultaneously on the same cluster. We provide 25 TB of storage capacity per server. This gives us sufficient capacity to support 3-way data replication for high availability and data protection, while still providing 100 TB of usable storage space.

The servers are housed in two racks and connected with a 480 Gigabit per second cluster fabric, using 10 Gigabit Intel® Ethernet Converged Network Adapters and two 48-port 10 Gigabit switches. Our data replication process is rack-aware, so data remains available in the event of a switch failure. As data volumes and performance needs grow, we can expand the cluster by adding more servers and switches.

Infrastructure note: Based on our experience, the infrastructure described above provides excellent performance for a wide range of Hadoop workloads acting on up to 100 terabytes of data. However, not all Hadoop workloads are alike. If you have exceptionally lightweight, I/O-intensive workloads, you might want to consider running them on a cluster of microservers based on the Intel® Atom™ processor C2000 system on a chip (SoC) to improve infrastructure efficiency. For more information about Hadoop workloads and Hadoop infrastructure considerations, see the Intel technical white paper: "Extract, Transform, and Load Big Data with Apache Hadoop*."

NoSQL Databases for Specialized Needs

NoSQL databases, such as HBase, Cassandra, and MongoDB, can run with Hadoop on the same infrastructure, or they can be deployed on a separate, similar cluster. Although infrastructure requirements will vary depending on the specific use cases, the same basic design considerations apply.
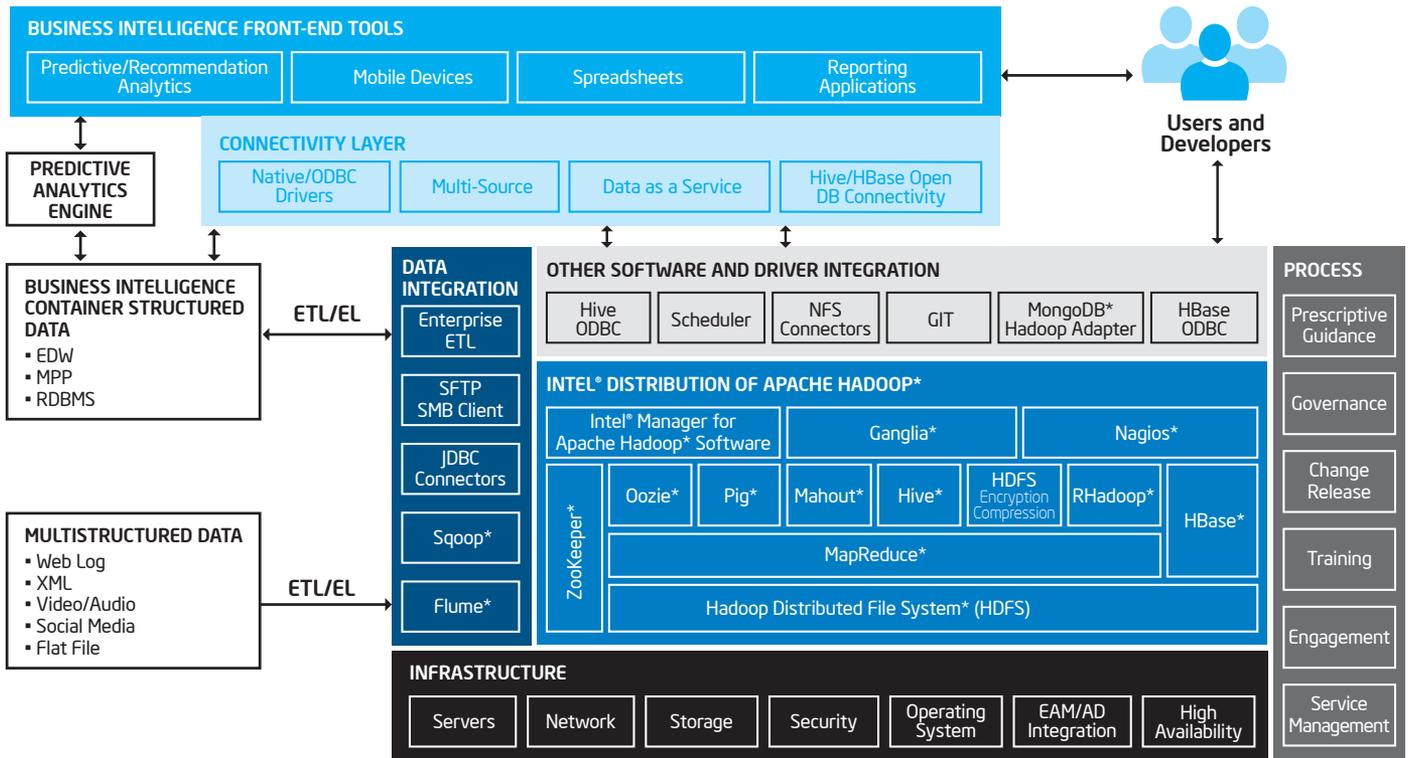
Figure 2. The Intel® Distribution includes the full distribution from Apache Hadoop*, which simplifies integration with other data analytics and BI infrastructure.

Key: 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

COMPONENT	SELECTED TECHNOLOGY	WHY SELECTED
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* on the same Hadoop cluster
Drives	25 TB HDFS raw storage per data node	Fulfills the deep storage requirements of a big data platform
Network Adapters	10Gb Converged Network Adapters	Provides the high network bandwidth needed when importing and replicating large data sets across servers
Switches	2x 48-port 10G	Enables high-bandwidth connectivity for enterprise-class performance

Table 3. Platform Design Components.

Cluster Overview

Table 3. A cluster of 16 two-socket servers based on the Intel® Xeon® processor E5 family provide excellent performance for a wide range of Hadoop workloads, acting on up to 100 TB of poly-structured data.

Managing Big Data

As you begin capturing, storing, and analyzing larger and larger amounts of data, you will eventually need to update your data management tools and strategies. Almost every business has diverse data sets residing in multiple transactional databases, DWs, data marts, and storage systems. It can be challenging to synchronize and unify these siloed data sources, to combine them as needed for analysis. In addition, legacy DWs and data marts continue to deliver business value, so emerging needs must be balanced against the disruption and potential risk associated with standardizing data formats and centralizing data management.

To begin addressing that need, Intel IT created a centralized Business Intelligence Data Management team. They make

high-quality enterprise data securely available to all parties, in consistent formats that simplify integration. The team maintains a data catalog and uses a customer engagement process to help business units identify and take advantage of big data opportunities. The team also works with business units to prioritize and coordinate new use cases.

Intel IT created two new roles to facilitate communication and decision making.

- **A service manager** supports each business group, such as Sales and Marketing, Supply Chain, and Design. Service managers understand big data capabilities and work directly with business units to understand business needs and data requirements, including time-criticality and the need for integration with existing LOB data marts.

- **A subject area product manager** ensures that data is available and usable for each subject area. These product managers work closely with business groups and IT support teams to determine the best way to structure the data so it can be used efficiently across the business.

Together, service managers and subject area product managers provide an efficient link between business units and application development teams. They have been instrumental in helping us define, prioritize, and implement successful big data use cases. A similar approach may be useful as you work to integrate increasing data types and data volumes into your analysis environment.

Conclusion

We are entering a new phase of the information era, one in which organizations query and analyze huge volumes of diverse data in real-time or near real-time to improve outcomes for their most critical business processes. In this paper, a variety of new technologies comprise a reference architecture to enable that transition. The key technologies include Apache Hadoop, NoSQL databases, MPP data warehouse architectures, and in-memory RDBMS.

There is no single technology that addresses the full range of big data requirements, but these and other technologies can be applied in various combinations—and combined with traditional systems and processes—to support high-value use cases through interactive query and predictive analytics capabilities. Intel IT uses multiple DWs and a centralized approach to data management to implement big data use cases that are now delivering hundreds of millions of dollars in business value. We believe this is a flexible and efficient approach that would work well for many businesses.

FOR MORE INFORMATION VISIT THESE LINKS ON intel.com:

The Intel IT Center

Big Data Intelligence

Private Cloud Solutions

Turn Big Data into Big Value

Extract, Transform, and Load Big Data with Apache Hadoop*

Big Data Mining in the Enterprise for Better Business Intelligence

IT@Intel: Data Center and Cloud Solutions