White Paper
Intel® Xeon® Processor- and
Intel Atom™ Processor-Based Servers
Big Data Analytics

Perform Predictive Analytics and Interactive Queries on Big Data

Abstract

Recent innovations in data warehousing and business analytics dramatically increase the capability and potential value of today's massive, diverse, and often fast-moving data flows. Companies now perform interactive queries and predictive analytics using all available data, including operational data and the huge amounts of poly-structured data available from logs, social networks, sensors, and many other sources. In this white paper, we define a practical, cost-effective infrastructure for supporting data-driven decision-making on an enterprise scale. These solutions build on the hybrid IT infrastructure introduced in an earlier Intel paper, "Extract, Transform, and Load Big Data with Apache Hadoop.*"

Overview

The torrents of data flowing into businesses today can fuel new insights and actions in near-real-time. Organizations ready for that opportunity stand to gain a substantial competitive advantage. According to a recent article in Harvard Business Review by Andrew McAfee and Eric Brynjolfsson, "The more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results. In particular, companies in the top third of their industry in the use of data-driven decision making were, on average, 5 percent more productive and 6 percent more profitable than their competitors."[1]

Most IT decision-makers are well aware of the value big data analytics offers to their business. In an Intel survey of 200 IT professionals, 46 percent considered improving analytics capabilities to be one of their top priorities, and 90 percent rated the importance of improving analytics either 5 or 4, on a scale of 1 to 5, where 5 is "most important."[2]

The traditional approach to analysis is to copy data from operational systems into a data warehouse for queries and analysis. While that model still applies to many scenarios, companies now integrate more data—and more diverse data—into their analytics environments, including documents, multimedia files, operational logs, social networking posts, Internet click-streams, sensor measurements, and more. Businesses combine historical data with fresh operational data and high-volume streaming data to achieve a 360 degree view of their customers and their internal operations.

Many companies also are shortening the time it takes to analyze large data sets, so they can integrate analytics into time-sensitive business processes. All of this can be accomplished today using proven methods, but the overwhelming volumes, continuous generation, and poly-structured nature of big data requires upgrades and additions to traditional infrastructure.

Parviz Peiravi,
Principal Architect,
Intel Enterprise Solution Sales

Ajay Chandramouly,
Big Data Domain Owner,
Intel IT

Chandhu Yalla,
Big Data Engineering Manager,
Intel IT

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

Table of Contents

In the first paper of this series, "Extract, Transform, and Load Big Data with Apache Hadoop,*" we introduced Apache Hadoop* software as a solution for ingesting and preparing large volumes of poly-structured data for analysis. In this paper, we add capabilities to support advanced analysis, such as interactive queries and predictive analysis. We also discuss methods to support additional requirements, such as complex event processing (CEP), which typically requires the rapid analysis of high-volume streaming data.

The result is a cost-effective infrastructure for data-driven decision-making that delivers high value across many use cases and can be adapted and extended as needs grow.

The Analytics Continuum

As companies move toward the pervasive use of data for business decision making, analytics capabilities tend to evolve along a predictable trajectory (Figure 1).

- **Descriptive analytics** act on historical data to help decision-makers understand what happened and why it happened, so they can make better decisions going forward. This is the traditional data warehousing model. IT organizations combine data from many sources and develop data models and reports that provide insight into key business issues and performance metrics.
- **Interactive queries** allow users to engage directly with data, so they can iterate their way to desired insights more creatively. Analysts slice and dice, mine, query, and visualize data to unearth new insights. Many data warehouse and business intelligence (BI) software vendors offer self-service, interactive BI tools for business users as well as for data analysts.
- **Predictive analytics** uses data to predict future events. The data can be fresh, historical, structured, unstructured, or all of the above. Once you develop the predictive models, you can extend them to provide prescriptive recommendations for achieving desirable outcomes. Although predictive analytics can aid in

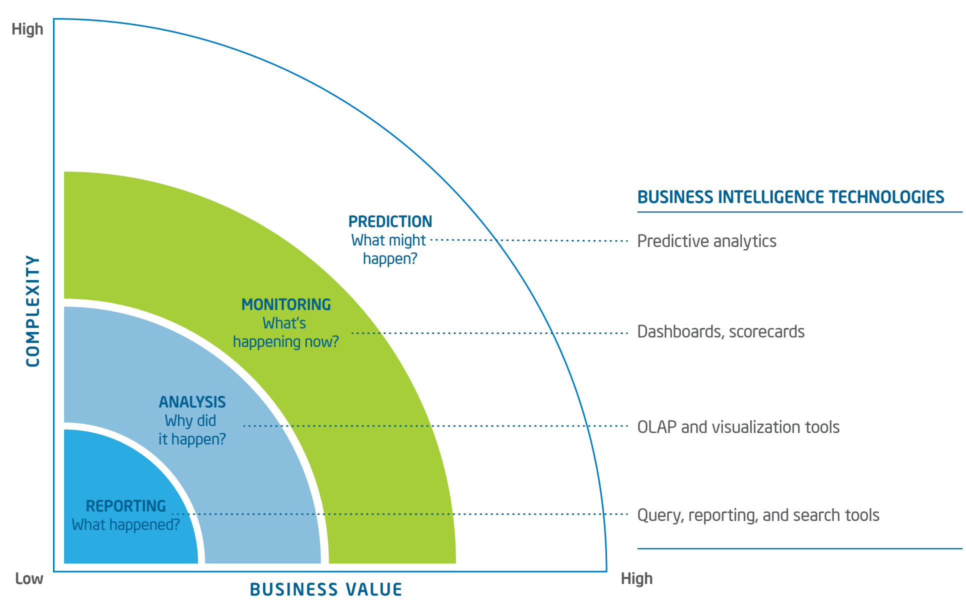

Figure 1. Businesses can achieve higher value by moving up the analytics continuum to support more advanced analytical models, such as interactive queries and predictive analytics.
Source: Business Analytics 3.0 blog, http://practicalanalytics.wordpress.com/predictive-analytics-101/

strategic business planning, its greatest value may come from tactical guidance at the point of decision, and operational guidance at the point of execution—both of which require fast execution to support real-time business processes. Retailers use predictive analytics to deliver more compelling offers to individual customers, healthcare organizations use it to select best-fit treatment protocols, and financial services organizations use it to increase investment returns and reduce risk.

Foundational Elements of a Big Data Analytics Platform

A modern platform for big data analysis can be built using just two key elements: an enterprise data warehouse (EDW) for structured data analysis, and an Apache Hadoop cluster for storing and processing unstructured data (Figure 2). By linking the two with a high-speed data loader, you can quickly move large amounts of data from one to the other to address diverse use case requirements.

Each platform has its strengths. The EDW is ideal for complex analyses that act on structured data and cannot readily be broken down into smaller tasks. The Hadoop cluster is ideal for executing relatively simple data processing functions that can be broken into parallel tasks and executed across distributed data sets of any size and type.

Note: Depending on the size of your business and the complexity of your analytic requirements, your big data platform may require more than two key elements. For example, Intel's big data environment includes multiple data warehouses optimized for different data sets and analytic models. It also includes an array of line of business (LOB) data marts focused on more localized needs. The two-element platform defined in this white paper addresses the needs of many businesses and provides a clear conceptual framework for understanding the basic requirements of a big data platform.

For more information about Intel's big data strategy, see the IT@Intel white paper, "Using a Multiple Data Warehouse Strategy to Improve BI Analytics."

The Enterprise Data Warehouse (EDW)

The EDW includes a relational database management system (RDBMS) and built-in tools for managing and analyzing data. Although enhanced in many ways, a modern EDW is not fundamentally different from legacy systems, and can be readily integrated into existing environments. However, many modern systems offer a number of attributes that make them better suited for today's requirements.

- **A modular, massively-parallel processing (MPP) architecture** provides the scalability to support large data sets. MPP data warehouse platforms are offered as appliances by many vendors. Processing, memory, I/O, and storage are highly optimized for performance and scalability and software is pre-integrated to simplify deployment. You can implement these systems much faster and with less risk than a more traditional, build-your-own data warehouse.
- **Integrated data management and advanced analytic tools** provide more and better options for extracting value from data. Choose solutions that fit with your existing BI and analytics environment, but also look for new capabilities, such as support for R, a statistical computing and graphics software environment that is popular among data scientists. Your development teams can use R to extend and complement traditional analytic models based on the Structured Query Language (SQL) and other models.

If your use cases have extreme performance requirements, you may want to consider deploying an analytics appliance with an in-memory database. By storing all data in main memory, these appliances practically eliminate data access latencies to improve query performance by up to 100 times versus traditional systems. They are ideal for complex event processing (CEP) and other real-time, big data applications. However, such solutions entail significant costs, due to the requirement that all data be stored in memory. Before deploying an in-memory database, be sure you have specific usage models in mind that both require the performance and justify the expense.

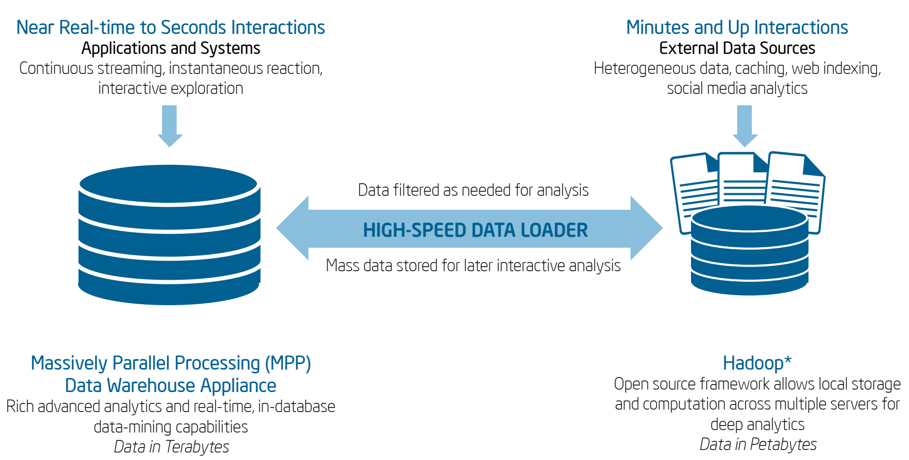

Figure 2. An MPP data warehouse appliance and a cluster of servers running Apache Hadoop* provide a flexible and extendable foundation for big data analytics.

MPP appliances with advanced analytic capabilities are available from numerous vendors, and some support in-memory databases. These systems often combine proprietary data filtering technologies with Intel® Xeon® processor family-based blades and commodity storage drives to deliver big data performance at relatively low cost, and with little configuration or maintenance. You can add blades to scale performance and capacity. Each blade is connected directly to multiple solid-state drives (SSDs) for high-speed data access through parallel streaming. This architecture eliminates the latencies associated with retrieving data from directly-attached spinning disks or a separate, shared storage system.

Server blades need to provide maximum performance, I/O throughput, and scalability. The Intel® Xeon® processor E7 family is a good fit for these requirements, and is used by many MPP appliance vendors to deliver high-end capabilities on a cost-effective platform. These processors provide up to 10 cores and 20 threads and up to 30 MB of last-level cache. They support up to 32 GB DIMMS, so they are ideal for data-intensive analytics applications. They also provide advanced reliability, availability, and serviceability (RAS) features to improve data integrity and support high levels of system uptime—both of which are essential in mission-critical analytics environments.

In choosing an MPP appliance, be aware that it does not have to satisfy all of your structured analytics needs. Ideally, it should fit into your existing data analytics and business intelligence environments, so you can continue to make use of your legacy data warehouses and data marts. As you develop new use cases, you can run queries on the system or systems that meet requirements at the lowest cost.

Your data volumes will grow over time—probably very rapidly—so make sure the appliance you choose can scale to meet your projected needs. Performance can vary appreciably depending on workloads and architectural features, so before purchasing, test your prospective appliance using a data set that closely mirrors your production environment. Also make sure that the new warehouse supports your preferred tools and methods, or be prepared to train staff as needed.

Storage performance is critical to support fast query response times and also for quick data loads and backups. A number of advanced storage technologies are now supported as mainstream features in storage systems based on Intel Xeon processors. These technologies reduce the amount of physical data that is required to store logical copies of the data as seen by the file system of the data warehouse.

Look for the following storage features in the storage solution of your MPP appliance or in any shared storage system you implement.

- **Data de-duplication** was initially introduced for backup systems, because businesses were storing many copies of the same or nearly the same data for recovery purposes. This technology is now extending to inline production data, and is set to become a standard feature on storage controllers.
- **Data compression** was originally an offline process designed to reduce the amount of disk space required to store a fixed amount of data. It is now being extended to online production disk storage systems and is also set to become a standard feature in many storage controllers. The trend is toward real-time compression that does not compromise performance.
- **Intelligent tiering,** also known as automated tiering, moves "hot" data to faster storage devices, such as SSDs, and moves cold data to slower performing and higher capacity drives. Tiered storage can be completely transparent to applications and users, and provides a cost-effective way to improve performance using a relatively small number of faster but more costly drives.
- **Thin provisioning** gives the appearance of more physical capacity than is actually available, which enables administrators to automatically expand capacity on demand. It saves time and reduces the typical cost of having to overprovision available capacity well before you need it.

The Apache Hadoop* Cluster

You can handle the unstructured portion of your data using a cluster of servers running Apache Hadoop software. With its distributed, parallel processing capabilities, a Hadoop cluster can rapidly ingest, store, and process petabytes of poly-structured data by coordinating local storage and computation across tens, hundreds, or even thousands of servers. Each server stores and processes a subset of the data. Because applications execute in parallel, performance and capacity scale with each server you add to the cluster.

The Hadoop framework includes a variety of components for managing data and applications, including:

- **HBase,*** a Not Only SQL (NoSQL) database that allows for low-latency, quick lookups in Hadoop.
- **Hive,*** a data warehouse system for Hadoop that facilitates data summarization, ad-hoc queries, and the analysis of large data sets. Hive provides a mechanism for projecting structure onto stored data and for querying the data using a SQL-like language (HiveQL).
- **Pig,*** an interactive scripting environment for processing data.
- **Mahout,*** a data-mining library that provides algorithms for clustering, collaborative filtering, regression testing, and statistical modeling.
- **Sqoop,*** an import/export utility designed for exchanging data with RDBMSs.
- **Oozie,*** a workflow environment for coordinating complex data processing operations.

By itself, a Hadoop cluster provides a fast, massively-scalable solution for preparing and loading data into the EDW and other analytics environments. Analytics also can be performed directly on the poly-structured data stored within the Hadoop cluster, although query capabilities are not as advanced or comprehensive as in a traditional SQL-based EDW.

Hadoop applications tend to run most efficiently on clustered dual-socket servers based on the Intel® Xeon® processor E5 family, configured with local storage drives, and linked together in a 10 Gigabit Ethernet network. However, workloads vary and many factors can influence performance, scalability, and cost. For more detailed information on designing a best-fit infrastructure for Hadoop, see the Intel technical white paper, "Extract, Transform, and Load (ETL) on Apache Hadoop.*"

Extending Your Big Data Analytics Platform

Performing Interactive Queries on Large, Poly-Structured Data Sets

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

You can use Hive and HBase to add structure to your data, so you can perform more 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 EDW for interactive querying.

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

Implementing Predictive Analytics to Improve Real-Time Outcomes

In some cases, you may want to perform analytics directly on poly-structured data in Hadoop. There are some limitations you will need to address. Hadoop is a framework for distributed data processing and storage. It is not a database, so it lacks functionality that is necessary in many analytics scenarios. Fortunately, there are many options available for extending Hadoop to support complex analytics, including real-time predictive models.

As one example, Intel IT used its Hadoop cluster in combination with Cassandra,* a Not Only SQL (NoSQL) database, to provide the foundation for a real-time recommendation system. The BI team developed the predictive algorithms using Hadoop's Mahout data mining library. The algorithms act on historical data stored in Hadoop, and then transfer the results into Cassandra, which provides the fast, low-latency data retrieval needed to support real-time scenarios. During an online interaction, the appropriate results are retrieved from Cassandra and combined with contextual data, such as user input and location, to provide best-fit recommendations in real time.

Intel IT uses a number of different NoSQL databases to extend the functionality of its Hadoop cluster, including Cassandra, HBase, and MongoDB*. There are more than a hundred different NoSQL databases you can use with or without Hadoop to address diverse requirements. Many are available as open-source downloads, as vendor-supported software distributions, or both. In general, a NoSQL database relaxes the rigid requirements of a traditional RDBMS so you can implement new usage models that require lower latency and higher levels of scalability. NoSQL databases come in four major types.

- **Key/Value Object Stores** are the simplest type of database and can store any kind of digital object. Each object is assigned a "key" and is retrieved using the same key. Since you can only retrieve an object using its associated key, there is no way to search or analyze data until you have first retrieved it. Data storage and retrieval are extremely fast and scalable, but analytic capabilities are limited. Examples of key/value object stores include Redi and Riak.
- **Document Stores** are a type of key/value store in which documents are stored in recognizable formats and are accompanied by metadata. Because of the consistent formats and the metadata, you can perform search and analysis without first retrieving the documents. Examples of document stores include Couchbase and MongoDB.
- **Columnar Databases** provide varying degrees of row and column structure, but without the full constraints of a traditional RDBMS. You can use a columnar database to perform more advanced queries on big data. Examples of columnar databases include Cassandra and HBase.
- **Graph Databases** store networks of objects that are linked using relationship attributes. For example, the objects could be people in a social network who are related as friends, colleagues or strangers. You can use a graph database to map and quickly analyze very complex networks. Examples of graph databases include Allegro and Neo4J.

With the right NoSQL database, you can implement analytics use cases that would be impossible on a Hadoop cluster and either impossible or cost-prohibitive on a traditional EDW. There are many excellent NoSQL products available today to address a wide range of requirements at relatively low cost. However, features and capabilities vary greatly. Spend the time and effort early to be sure your chosen database can address the requirements of your use case.

Analyzing High-Volume, Fast-Moving Data Streams

Your business may generate large volumes of data from sensors, call centers, Internet sites, and other sources. You can gain valuable insights by ingesting and then analyzing that data. In some cases, you can achieve even higher value by filtering and analyzing the data as it is generated, without taking the time to load it into your EDW, or even into your Hadoop cluster. You may even want to integrate the results into your real-time predictive models, either as triggering events, or to provide real-time context to improve predictions and recommendations.

Solutions are available to address this need. For example, Twitter developed an application called Storm[3] that it uses to analyze hundreds of millions of tweets per day[4] in real time to identify the "trending topics" that are of most interest to its customers. Although it might be easier to ingest and then analyze this data, the resulting delays would make the trending feature far less compelling to users.

Storm is now available as an open-source application for processing both static and streaming data in real time to support complex event processing (CEP). You can use it to identify, filter, and process targeted information in a "fire-hose" of streaming data. Storm provides guaranteed message passing, and developers can use any software language that has an adapter for connecting to Storm's architecture. Once the initial insights have been gleaned, you can load all or part of the data into Hadoop or your EDW for more extensive analysis.

Curating Data for Usability and Risk Mitigation

Most businesses with advanced analytics capabilities already have mature data curation solutions integrated into their EDW and data mart environments. Big data does not change the requirements for data curation in any fundamental way, but the scope and complexity of big data, if not carefully managed, can lead to unacceptable risks. Data is generated or obtained from far more sources and may go through multiple transformations as it is ingested, stored, analyzed, and used. Maintaining data quality and usability is more challenging in such a complex data environment, as is maintaining legal and regulatory compliance.

Security and Compliance

A primary goal of most businesses is to learn more about their customers, so big data repositories are often replete with personally identifiable information (PII), such as credit card numbers, purchase histories, and medical information. You will need solutions for anonymizing PII at ingestion and for encrypting sensitive data, both at rest and in motion. Access controls also are important, including user authentication and the ability to authorize access to systems and data at appropriate levels of granularity. Also pay attention to infrastructure security, including methods for securely managing systems and applications and for securing communications between systems and among the nodes within each system.

API Management

Application programming interfaces (APIs) play an increasingly vital role in connected, data-driven businesses. Insight leads to value when applied to business decision-making, and APIs provide powerful and flexible tools for integrating insights into business processes and sharing them with partners and customers. As with the data itself, manage and secure your APIs centrally to provide flexible but controlled access to information and resources for users and developers. Caching also may be necessary to deliver fast response times for geographically distributed individuals and companies. Appropriate caching policies can speed-up global delivery and reduce the load on your data center.

Products and solutions for addressing data curation requirements in big data environments are evolving. Some vendors are extending existing solutions to help address escalating requirements. Others are developing new solutions that are purpose-built for big data.

Intel offers several solutions that can help (Figure 3).

- **A Service Gateway for Big Data Analytics.** The Intel® Expressway Service Gateway (Intel® ESG) is an enterprise-class software-appliance that can help you manage user access, data movement, and APIs more flexibly and securely. It provides interoperability with other infrastructure, mobile clients, and partners to unlock data for value added analytics, and it includes support for integration, routing, and legacy-to-mobile service enablement. It also provides support for authentication, authorization and accounting (AAA) security, so you can protect your assets more effectively as you expand your use cases and connect with more data, users, and applications.
- **Security-Optimized Hadoop Software.** The Intel Distribution for Apache Hadoop software provides integrated support for advanced security, including access controls and hardware-accelerated encryption using Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI). In combination with Intel ESG, you can selectively encrypt sensitive files, so they remain secure whether they are at rest or in motion within your Hadoop cluster. The Intel Distribution is highly optimized for manageability and performance on Intel® architecture. It includes an automated

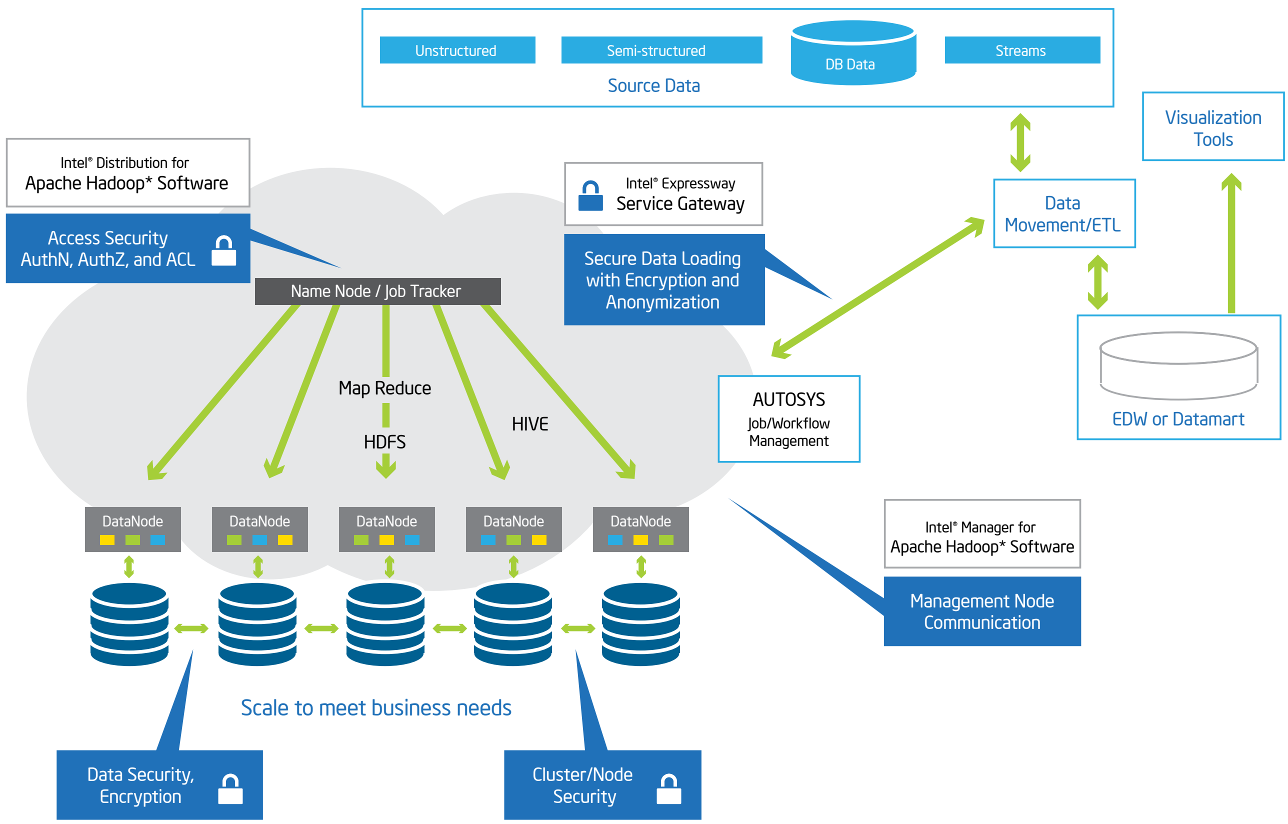

Figure 3. The Intel® Expressway Service Gateway and the Intel® Distribution for Apache Hadoop* software can be used together to support enterprise-class security in big data environments.

tuning engine that optimizes application performance quickly and with very little effort from IT personnel.

- **Cloud-based API Management.** Mashery, an Intel subsidiary, delivers enterprise solutions for advanced API management. Mashery solutions run on-premise, in the cloud, or on hybrid infrastructure that combines the advantages of both. Global support with advanced caching provides responsive performance from almost anywhere in the world.
- **API Management for the Telecommunications Industry.** Aepona, another Intel subsidiary, delivers comprehensive, cloud-based API support specifically tailored to the requirements of communications service providers (CSPs). The Aepona software platform helps CSPs extend the value of big data analytics more effectively to monetize their assets in new and more productive ways.

Getting Started

As discussed in this paper, there are many options for ingesting, storing, analyzing, and using big data. To improve your chance of success:

1. Start by identifying one or more high-value use cases for your business.
2. Locate the data you will need to support your use cases, and determine the kinds of flows, transformations, and analyses you will need to achieve your business goals.
   - Consider whether you need to support advanced analytics capabilities, such as interactive queries and predictive analytics.
3. Evaluate your infrastructure needs based on your defined requirements.
   - Can you simply extend your existing EDW with a Hadoop cluster for big data ETL?
   - Do you need to move up to an MPP appliance or even an in-memory database to support real-time queries on large structured data sets?
   - Do you need to add specialized components, such as a NoSQL database for interactive queries or low-latency lookups on large volumes of poly-structured data?
   - Do you need a CEP solution, such as Storm, for processing high-volume, fast-moving data streams?

In each case, let your specific business, data, and analytics requirements guide your decisions. The solutions you develop will most likely be extendable to more use cases, but focusing on the demands of each individual use case improves your chances of success and can help you achieve larger and faster returns on your investment.

4. As you develop your first use cases, take the time to explore and understand your data curation requirements. Laying a proper foundation for your first use cases will protect your business from potentially serious consequences. It will also allow you to extend your analytics capability more efficiently in the future, and without undue risk.

Conclusion

Today's growing volumes of poly-structured data can fuel new insights and decisive action in near-real-time. Those businesses ready to take advantage of that stand to gain a substantial competitive advantage. A hybrid, big data analysis infrastructure using an EDW appliance and clustered servers running Apache Hadoop software provides a flexible foundation for traditional descriptive analytics using big data, and also for interactive queries and predictive analytics to support advanced use cases that deliver higher value. Complex analytics can be applied to all data types, results can be integrated into real-time processes, and data analysts and business users can engage directly with data to uncover new insights.

Infrastructure designs should be based on the notion that one size does not fit all. Intel offers cost-effective, high-performing infrastructure support across the full range of big data requirements—along with reference architectures that can help you deploy new solutions more easily and with less risk.

Depending on specific data sets and analytic workloads, different capabilities are required and can be added relatively easily to a well-designed, hybrid infrastructure based on an MPP data warehouse appliance and a Hadoop cluster. Above all, information security is paramount, so look for solutions that let you implement strong security today, and that you can extend and adapt to accommodate a rapidly growing analytics environment.

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

Reference Architecture: Predictive Analytics and Interactive Queries on Big Data

[1] *Harvard Business Review:* "Big Data: The Management Revolution," by Andrew McAfee and Erik Brynjolfsson, October, 2012. http://hbr.org/2012/10/big-data-the-management-revolution

[2] Source: Peer Research, Big Data Analytics, August 2012, Intel Corporation. http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf

[3] Source: "Twitter Storm: Open Source Real-time Hadoop," by Bienvenido David, September 26, 2011. http://www.infoq.com/news/2011/09/twitter-storm-real-time-hadoop

[4] Source: "Twitter hits 400 million tweets per day, mostly mobile," by Dan Farber, CBS editor, June 6, 2012. http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/

Printed in USA 1013/DF/HBD/PDF Please Recycle 329777-001US