Planning Guide

Getting Started with Big Data
How to Move Forward with a Successful Deployment

Why You Should Read This Document
This planning guide provides background information and practical steps for IT managers who want to plan and implement big data analytics initiatives, including:

• Today’s IT landscape for big data and the challenges and opportunities associated with this disruptive force

• Big data technologies that comprise a flexible big data platform, with a focus on the Apache Hadoop* framework and in-memory analytics

• The importance of putting the right infrastructure in place for various use cases and for an optimal big data deployment

• Three basic “next steps” and a checklist to help IT managers move forward with planning and implementing their own big data project
Contents

3  Today's IT Landscape for Big Data Analytics
4  Understanding Big Data Technologies
6  Deploying Big Data Solutions
12 Get Started with Big Data Analytics:
   Three Basic Steps
15 Intel Resources for Learning More
Today’s IT Landscape for Big Data Analytics

The new buzz about big data has shifted from hype to a more considered conversation. Still-maturing technologies, skills shortages, and shifts in the way IT and the business work together are the new reality. Exploiting big data is not simple. Yet the business case for big data remains compelling. Despite a backlash from skeptics who have become disillusioned with the economic promises of big data, organizations are forging ahead. For example, IDG’s latest big data enterprise survey found that nearly half (49 percent) of those surveyed have either deployed or are planning to deploy big data–related projects. Maintaining the status quo represents too great a risk of being left behind by competitors.

Today, vendors offer a growing number of comprehensive, enterprise-ready platforms and solutions that build on technology innovations. The conversation has moved beyond “Is there value in big data?” to “How can I use it to create value for my organization?” and “How can I accelerate time to insight to gain competitive advantage?”

What’s Next for Big Data: Predictive Analytics, Real Time, and the Internet of Things

Big data derives most of its value from the insights it produces when analyzed—helping organizations to discover patterns, find meaning, make decisions, and ultimately respond to the world with intelligence. As technology matures and the conversation continues to evolve, organizations will develop new ways to gain insight by operationalizing approaches to big data that generally have been out of reach for mainstream business.

For example, organizations are turning to predictive analytics to help them deepen engagement with customers, optimize processes, and reduce operational costs. The combination of real-time data streams and predictive analytics—sometimes referred to as processing that never stops—has the potential to deliver significant competitive advantage for business. For an overview of predictive analytics, including why it matters and how businesses can operationalize it, see Predictive Analytics 101: Next-Generation Big Data Intelligence.

What Exactly Is “Big Data?”

Big data refers to huge data sets that are orders of magnitude larger (volume); more diverse, including structured, semistructured, and unstructured data (variety); and arriving faster (velocity) than you or your organization has had to deal with before. This flood of data is generated by connected devices—from PCs and smart phones to sensors such as RFID readers and traffic cams. Plus, it’s heterogeneous and comes in many formats, such as text, documents, images, videos, weblogs, and transactions.

The Internet of Things (IoT)—Internet-enabled devices that network and communicate with one another and the cloud—is also driving big data analytics innovation. Gartner estimates that the IoT will have an installed base of over 33 billion “things” in 2020, generating massive amounts of data in a fast-moving data stream. Most of this data will be generated by machines via embedded sensors and actuators linked through wired and wireless networks that communicate using the same protocol that connects to the Internet. Human-generated data from devices such as mobile phones and tablets also will be part of the mix. This data can be used to unlock correlations between events, automate intelligent systems, and provide the insight to solve new and more complex business and social problems. Find out more about Intel’s perspective on machine-generated data in the Internet of Things Video: IoT Explained.

Relieving the Pressure on IT

With so much at stake for the business, big data initiatives can’t happen in a vacuum. IT must forge a strong partnership with business leaders to identify big data opportunities and move forward with needed support. Big data also requires new business, technical, and analytical skill sets to help model complex business problems and discover insights, integrate systems, build out massive databases, and administer distributed software frameworks.
Full adoption of big data analytics can be implemented in three basic, high-level steps. The order is important, although activities will overlap as you move forward:

1. **Understand how big data will impact your organization culturally.** Work with business leaders to determine the cultural boundaries of your implementation—internally and externally.

2. **Hire the skills you need.** Acquire the business, technology, and analytics skill sets you need, such as data scientists, system architects, and data engineers.

3. **Implement your big data solution.** Identify technology requirements and implement the solution stack.

Although all three steps are critical to your success, this guide focuses on step 3: implementing big data solutions. The International Institute for Analytics is an excellent source for resources on the first two steps.

---

**Understanding Big Data Technologies**

Traditional tools and infrastructure are not efficient for working with larger, varied, and rapidly generated data sets. For organizations to realize the full potential of big data, they must find a new approach to capturing, storing, and analyzing data.

Big data technologies use the power of a distributed grid of computing resources and “shared nothing architecture,” distributed processing frameworks, and nonrelational databases to redefine the way data is managed and analyzed. Server innovations and scale-up in-memory analytics solutions make it possible to optimize compute power, scalability, reliability, and total cost of ownership for the most demanding analytics workloads, including ingesting and storing streaming data and real-time analytics.

**The Big Data Solution Stack**

Depending on the use case, the big data solution stack includes a high-performance infrastructure powering some combination of distributed processing frameworks such as Apache Hadoop software, nonrelational analytics databases, and analytics applications.

From a functional standpoint, these technologies complement each other and work together as a flexible big data platform that can also take advantage of existing data management architecture. For example, Hadoop historical analysis can be moved into nontraditional relational database management systems (RDBMSs) or NoSQL databases to allow for ingestion and analytics on a wide variety of unstructured data sources. Or structured data from traditional enterprise data warehouses (EDWs) can be combined with unstructured data for further analysis.

**Apache Hadoop® Software**

Apache Hadoop software is a complete open-source framework for big data and has emerged as one of the best approaches to processing large and varied data sets. The Hadoop framework provides a simple programming model for distributed processing of large data sets. It includes the Apache Hadoop Distributed File System (HDFS), a framework for job scheduling called Apache Hadoop YARN, and a parallel processing framework called Apache Hadoop MapReduce. Several additional components support specific capabilities: for example, the ingestion of data (the Apache Flume service and Apache Spark software), queries and analysis (Apache Pig, Apache Hive, and Apache HBase software), and coordination of workflows (Apache Oozie), as well as management and monitoring of the underlying server cluster (Apache Ambari software). Combined, the Apache software components are a powerful framework for processing and analyzing distributed data in batch for historical analysis.

**From Batch to Near-Real Time**

The Apache Hadoop framework has been best known for batch processing, but Apache Spark software is quickly evolving the platform to near-real-time speed. Spark is an engine for rapid, large-scale data processing that runs in-memory or against disk storage. It includes a number of high-level tools that can be combined seamlessly in the same application such as machine-learning algorithms, structured data processing, graph data computations, and stream processing.
The Hadoop framework is available through the open-source community or as a packaged distribution from vendors who include value-added software and services (such as management, training, and support). Many of these distributions can integrate with EDWs, RDBMSs, and other data management systems so that data can move between Hadoop clusters and other environments to expand the pool of data to process or query.

**In-Memory Analytics**

In-memory computing is a game changer for business, with the potential to significantly impact the power and speed of big data analytics. With in-memory computing, real-time, data-driven decision making becomes possible—and affordable for mainstream businesses.³

In-memory computing eliminates one of the major limitations to many big data solutions—the high latency and I/O bottlenecks caused by accessing data from disk-based storage. In-memory computing keeps all relevant data in the main memory of the computer system. Data can be accessed orders of magnitude faster, making it available for immediate analysis—and business insights are available almost instantly. With in-memory computing, whole data marts and data warehouses can be moved into DRAM for rapid analysis of the entire data set.

In-memory analytics integrates analytics applications and in-memory databases on dedicated servers. This is ideal for analytics scenarios with heavy compute requirements and real-time data processing. Examples of in-memory database solutions include the SAP HANA* platform (developed jointly by Intel and SAP), Oracle* Database In-Memory Option for Oracle 12c, IBM* in-memory systems with BLU Acceleration, SAS* In-Memory Analytics, and Apache Spark (see sidebar “From Batch to Near Real-Time” above).

For more about current in-memory vendor solutions and how in-memory computing innovation is changing the way businesses analyze big data, read the white paper Changing the Way Businesses Compute and Compete with Analytics.

---

**Intel and Cloudera Join Forces**

In March 2014, Intel announced a substantial equity ($740 million) and intellectual property investment in Cloudera, provider of the most popular Apache Hadoop* software version on the market. Intel also announced it would exit the big data analytics software market with its own distribution and work with Cloudera to integrate optimizations from the Intel* Distribution for Apache Hadoop software (also referred to as the Intel Data Platform) into Cloudera’s distribution including Apache Hadoop (CDH).

Together, Intel and Cloudera continue to drive innovation through open-source technologies, with a focus on security, performance, and management to accelerate Hadoop adoption in the mainstream enterprise. Cloudera also is working closely with Intel to make sure its products make the best use of Intel data center technologies. The Intel and Cloudera technology collaboration also combines efforts on foundational Hadoop* technologies to help move the software framework forward and encourage open-source developers to innovate in and on top of the platform.

For more information about CDH, visit cloudera.com.

---

**More Power and Speed**

In-memory computing has been around for a while in the form of distributed data grids and large, expensive installations. However, today’s in-memory systems are faster, more powerful, and more cost-effective. Why? As predicted by Moore’s Law, the cost of memory continues to fall—the costs for DRAM and NAND flash memory have dropped dramatically—while at the same time, the number of processors per chip is rising. With the Intel® Xeon® processor E7 v3 family, a four-socket server can be configured with up to 6 terabytes (TB) of memory and an eight-socket server with up to 12 TB, enough to hold many of today’s largest databases within the memory of a single server. With third-party node controllers, the Intel Xeon processor E7 v3 family can scale up to more than 32-socket systems for even more storage capacity.

This, combined with server innovations such as performance enhancements (for example, built-in technology for transaction speed acceleration and faster scan operations⁴) and the maturing of analytics software platforms, makes scale-up in-memory database architectures more affordable.
NoSQL Databases
These nonrelational databases come in four different types of stores—key value, columnar, graph, or document—and provide high-performing, high-availability storage at web scale. They are useful for handling massive streams of data and flexible schema and data types with fast response times. NoSQL databases use a distributed and fault-tolerant architecture, which provides system dependability and scalability. Examples of NoSQL databases include Apache HBase, Apache Cassandra®, MarkLogic®, MongoDB®, and Apache CouchDB® software.

Columnar Analytics Databases
These grid-based databases store data using columns instead of rows, reducing the number of data elements to be read while processing queries and providing fast performance for running a large number of simultaneous queries. Columnar analytics databases are read-only environments that deliver price-for-price performance and scalability advantages over conventional RDBMS systems. They are used for EDWs and other query-intensive applications and optimized for storage and retrieval of advanced analytics. SAP® Sybase® IQ, ParAccel® Analytic Platform, and HP® Vertica® Analytics Platform rely on columnar analytics databases.

Graph Databases and Analytics Tools
Graph databases are a type of NoSQL database that is increasing in importance. These databases are particularly useful for highly connected data in which the relationships are more numerous or more important than the individual entities. Graph data structures are flexible and make it easy to connect and model data. They are faster to query, and more intuitive to model and visualize. Much of the growth in big data is graph in nature.

Graph databases work alone or in conjunction with other graph tools, such as graph visualizations, graph analytics, and machine learning. For example, with machine learning, graph databases can be used to mine and predict relationships to solve a range of problems.

Deploying Big Data Solutions

Big data deployments can have large infrastructure requirements. The hardware and software choices made at design time can have a significant impact on performance and total cost of ownership. IT can get the most from a big data deployment by ensuring that the right infrastructure is in place for your particular use case and that Hadoop and analytics software are optimized and tuned for best performance.

Build the Capabilities Your Business Needs
With a flexible, extensible big data platform, IT can build the capabilities the business needs while choosing the most cost-effective systems to handle each use case. The following three usage models build on each other to deliver increasing value.

Extract, Transform, and Load (ETL)
Extract, transform, and load (ETL) aggregates, preprocesses, and stores data, but traditional ETL solutions can’t handle the volume, speed, and variety that characterize big data. Because the Hadoop platform stores and processes data in a distributed environment, Hadoop breaks up incoming data into pieces and handles the processing of large volumes in parallel. The inherent scalability of Hadoop accelerates ETL jobs so that time to analysis is significantly reduced. Find out more about ETL using Hadoop software in the white paper Extract, Transform, and Load Big Data with Apache Hadoop®.

Interactive Queries
Combining the Hadoop framework with a modern EDW based on massively parallel processing (MPP) architecture extends your big data platform to handle interactive queries and more advanced analytics. Hadoop can ingest and process large volumes of diverse streaming data and load it into the EDW for ad hoc Structured Query Language (SQL) queries, analysis, and reporting. Because Hadoop processes a wide variety of data types, the EDW is enriched with data that is not generally feasible to store in traditional EDWs. Plus, data stored in the Hadoop infrastructure can persist over a much longer duration, enabling you to provide more granular, detailed data through the EDW for high-fidelity analysis.
Predictive Analytics
Predictive analytics extracts greater value from data by using historical data to predict what may happen in the future. Intel IT recommends combining an EDW based on MPP architecture that can perform complex predictive analytics quickly with a Hadoop cluster for fast, scalable, affordable ETL. The Hadoop cluster can also be extended with tools and other components to perform additional data processing and analytics functions. For more details about this usage model, read the white paper Perform Predictive Analytics and Interactive Queries on Big Data.

Intel IT’s Big Data Platform
As Intel IT has worked with the business to develop big data use cases, they have combined elements of the first two usage models with predictive analytics for a flexible hybrid analytics infrastructure.

Intel IT uses Hadoop to offload the ingestion, transformation, and integration of unstructured data from social media, web traffic, and sensor logs into an EDW based on MPP architecture. The advantage here is that by adding structure to the heterogeneous data in Hadoop during extract and transform, and then loading it into the EDW, users can apply traditional business intelligence (BI) and analytics tools for interactive queries and other advanced analysis.

Intel IT deploys Hadoop software running on the Intel® Xeon® processor E5 family for heterogeneous data ingestion, caching, web indexing, and social media analytics. The Hadoop software filters the data as needed for analysis and moves it to a data warehouse appliance.

The data warehouse appliance is based on MPP architecture and is used to rapidly perform complex predictive analytics and interactive data exploration with near-real-time results. The data warehouse appliance is a third-party solution built on the Intel Xeon processor E7 family to deliver high performance and availability at relatively low cost. The appliance integrates with existing BI solutions and provides support for advanced analytics tools such as the R statistical package.

Intel IT further extended its big data platform by developing a predictive analytics engine in-house to provide ongoing predictive services. A first use case is the implementation of a real-time recommendation service. For this service, the BI team developed predictive algorithms using the Apache Mahout® data mining library. These algorithms act on historical data stored in Hadoop, and then transfer the results into the NoSQL Cassandra® database. Cassandra software provides the fast, low-latency data retrieval required for real-time use scenarios. During a user interaction online, results are retrieved from the Cassandra database and combined with contextual data (user input and location, for example) to provide real-time best-fit recommendations.

To deliver the extreme query responsiveness required for real-time analysis of high-volume data sets, Intel IT conducted tests to determine the optimal platform for a cost-effective, high-performance in-memory BI solution. See Configuring an In-Memory BI Platform for Extreme Performance for best practices related to combining server speed, number of processor cores, cache size, and memory for industry-standard Intel Xeon processor-based servers.
Infrastructure for the Hadoop* Framework

**Servers**
The Hadoop framework works on the principle of moving computing closer to where the data resides, and the framework typically runs on large server clusters built using standard hardware. The framework easily scales on servers based on Intel Xeon processors. The combination of the Hadoop framework with standard server platforms provides the foundation for a cost-efficient and high-performance analytics platform for parallel applications.

From a cost-benefit perspective, two-socket servers based on the latest Intel Xeon processor E5 family are the optimal choice for most Apache Hadoop workloads.

These servers are generally more efficient for distributed computing environments than large-scale multiprocessor platforms. They deliver exceptional performance and provide greater efficiencies in load balancing and parallel throughput compared with smaller, single-socket servers. Technologies for accelerating encryption, reducing latency, and increasing bandwidth are built into the processors.

Intel Xeon processors support error-correcting code (ECC) memory, which automatically detects and corrects memory errors, a common source of data corruption and server downtime. An Apache Hadoop cluster has a lot of memory (typically about 64 gigabytes [GB] or more per server), making ECC memory a critical feature.

Intel IT has created a flexible big data platform that delivers insights by using batch and real-time processing and by incorporating multiple technologies and solutions.
Intel® Processors for the Apache Hadoop® Framework

Two-socket Intel® Xeon® E5 family-based servers are an excellent choice for Apache Hadoop® deployments.

| Intel Xeon processor E5 v3 family | • Built on Ivy Bridge microarchitecture using Intel’s industry-leading 22 nanometer (nm) 3-D Tri-Gate transistor technology for superior performance and energy efficiency
|  | • Designed for compute, storage, and networking to work better together
|  | • Elastic scaling to adapt to fluctuating workload and increasing network and storage demands with Intel Turbo Boost Technology
|  | • Performance and I/O enhancements to improve and balance overall system performance and increase server efficiency
|  | • Improves performance and reduces latency of cloud applications like Memcached with the advanced traffic steering capability of Intel Ethernet Flow Director
|  | • Enables rapid encryption and decryption to encourage pervasive data protection
|  | • Error-correcting code (ECC) memory
|  | • Support for integrated 10 and 40 gigabit Ethernet (GbE) networking and simplified data center infrastructure
|  | • Server-and data-center-level telemetry for power monitoring and management that can optimize power consumption
|  | • Exceptional performance for single-and multi-threaded applications, including high-performance computing

Networking and Storage

Big data server platforms are benefited by dramatic improvements in mainstream compute and storage resources and are complemented by 10 gigabit Ethernet (10 GbE) solutions for a balanced system. The increased bandwidth associated with 10 GbE is critical to importing and replicating large data sets across servers. Intel 10 gigabit Ethernet solutions provide high-throughput connections, and Intel Solid-State Drives (SSDs) are high-performance, high-throughput hard drives for raw storage.

To enhance efficiency, storage needs to support advanced capabilities such as compression, encryption, automated tiering of data, data deduplication, erasure coding, and thin provisioning—all of which are supported with the Intel Xeon processor E5 family today.

Intel has done considerable testing using Intel Xeon processor E5 family-based servers as the baseline server platform for Hadoop clusters. A team of Intel big data, network, and storage experts measured Apache Hadoop performance results for various combinations of networking and storage components. Balancing the compute, storage, and networking resources provided a significant performance advantage—according to TeraSort benchmarks, processing time was reduced from 4 hours to 12 minutes—near-real-time results. For more information on high-performance Hadoop clusters on Intel technologies, see the white paper Big Data Technologies for Near-Real-Time Results.
Infrastructure for In-Memory Analytics Solutions

Servers
While Hadoop software can ingest and prepare large sets of heterogeneous data, advanced analytics (monitoring, interactive queries, predictive analysis) requires more powerful infrastructure. Modular MPP data warehouse platforms are offered as an appliance by a number of vendors. These appliances deliver with preintegrated software for simplified deployment and feature highly optimized processing, memory, I/O, and storage. Integrated data management and advanced analytics tools provide new ways to work with your data. Many solutions may also be compatible with your existing BI and analytics environment.

Extreme performance requirements may require an in-memory analytics appliance that combines database and analytics in a dedicated system. These systems are ideal for complex event processing (CEP) and other real-time applications, including streaming data. Servers based on the Intel Xeon processor E7 v3 family of products provide the memory, execution resources, and reliability needed for real-time, business-critical services in healthcare, energy, financial trading, and logistics by scaling memory to optimize resources to workload. Each generation measures significant performance boosts over the previous generation, and a large memory footprint and flexible configuration meet the requirements for high-capacity, massive streaming workloads.

In addition, because in-memory systems typically handle larger data sets and more scalable workloads per server than traditional solutions, they provide the data integrity and high availability necessary to support mission-critical processes.

Intel® Processors for In-Memory Analytics
The Intel® Xeon® processor E7 v3 family accelerates analytics for data-intensive scale-up workloads and in-memory databases.

| Intel Xeon processor E7 v3 family | • Built on Intel's industry-leading 22 nanometer (nm) 3-D Tri-Gate transistor technology for superior performance and energy efficiency |
| | • Large memory capacity: |
| | - Four- and eight-socket configurations with up to 1.5 terabytes (TB) of memory per socket for up to 6 TB or 12 TB of memory per server |
| | - Scalability up to 32-socket systems by using a third-party (OEM) node controller |
| | • Support for both DDR4 and DDR3 synchronous dynamic random-access memory (SDRAM) for fast transfer rates to help run expanding workloads |
| | • Exceptional performance for transactional workloads with Intel Transactional Synchronization Extensions 4 over previous-generation processors |
| | • Accelerated scan operations with Intel Advanced Vector Extensions 2 (Intel AVX2) 4 |
| | • Performance for single- and multi-threaded applications, including scale-up high-performance computing and technical applications |
| | • Extra capacity and flexibility for storage and networking connections with integrated PCI Express® (PCIe®) 3.0 ports, which improves bandwidth and supports PCIe-based solid-state drives |
| | • Forty advanced reliability, availability, and serviceability (RAS) features designed for 99.999 percent (five nines) uptime and improved data integrity for mission-critical analytics workloads, including Intel Run Sure Technology 4 |
| | • Broad, open, industry-standards-based ecosystem of leading partners that offer solutions optimized to run on Intel architecture, including in-memory databases for advanced analytics, and transaction-intensive workloads such as enterprise resource planning (ERP), online transaction processing (OLTP), and customer relationship management |
Networking and Storage

In-memory environments are extremely data intensive, requiring 10 gigabit Ethernet solutions for the speed and bandwidth necessary to avoid bottlenecks and support low-latency communications.

While in-memory systems rely on memory, persistent storage is needed to maintain SQL compliance for database transactions. Transactions must be atomic, consistent, isolated, and durable (ACID). Intel SSDs provide low-latency, high-bandwidth persistent storage that is standards based and cost-effective to support ACID compliance with real-time performance.

Optimizing and Tuning for Best Performance

Intel is a major contributor to open-source initiatives such as Linux*, OpenStack*, KVM, and Xen* software. Intel has also devoted resources to Hadoop analysis, testing, and performance characterizations, both internally and with systems and solutions providers, such as Cloudera. Through these technical efforts, Intel has observed many practical trade-offs in hardware, software, and system settings that have implications in the data center. Designing the solution stack to maximize productivity, limit energy consumption, and reduce total cost of ownership can help optimize resource utilization while minimizing operational costs.

The settings for the Hadoop environment are a key factor in getting the full benefit from the rest of the hardware and software solutions. Based on extensive benchmark testing in the lab and at customer sites using Intel processor-based architecture, Intel’s optimization and tuning recommendations for the Hadoop system can help you configure and manage your Hadoop environment for both performance and cost.

Getting the settings right requires significant up-front time, because requirements for each enterprise Hadoop system will vary depending on the job or workload. The time spent optimizing for your specific workloads will pay off not only in better performance, but also in a lower total cost of ownership for the Hadoop environment. See Optimizing Hadoop* Deployments for specific settings.

Benchmark Performance

Benchmarking is the quantitative foundation for measuring the efficiency of any computer system. Intel developed the HiBench suite as a comprehensive set of benchmark tests for Hadoop environments.10 Individual measures represent 10 important Hadoop workloads with a mix of hardware usage characteristics. HiBench includes microbenchmarks as well as real-world Hadoop applications representative of a wider range of data analytics such as search indexing, machine learning, and queries. HiBench 3.0 is now available as open-source software under Apache License 2.0. You can download the software, learn more about specific workloads, and find out how to get started at https://github.com/intel-hadoop/HiBench.
Get Started with Big Data Analytics: Three Basic Steps

If you’ve read this far, you now have a good understanding of the IT landscape for big data, its potential value to organizations, and the technologies that can help you get insights out of structured, semistructured, and unstructured data resources. Plus, you have a good overview of the basics for getting the right infrastructure in place and running smoothly to support your big data initiatives.

You can get started with your big data analytics project by following the three basic steps we described in the early pages of this guide. Although this guide has focused on technology and step 3, you can use the following checklist to help you work through the critical activities in all three steps.

**Step 1: Understand how big data will impact your organization culturally.**

- Develop an understanding of the value that big data analytics can bring to your organization.
  - Talk with your peers in IT and the business.
  - Take advantage of [Intel IT Center resources for big data](#) to get up to speed on the technologies.
  - Understand vendor offerings.
  - Take tutorials and examine user documentation offered by Apache.

- Collaborate with business leadership on a big data strategy and approach. Develop:
  - **The business case for big data** – How will big data analytics drive value for your business? What are the key business challenges it will address?
  - **Short-, mid-, and long-term objectives** – What are the key phases to achieving your big data goals?
  - **Current and future state of your IT infrastructure** – Can your data center support the big data platform? Assess your current data center technology and describe, if necessary, your plan to upgrade computing, storage, and networking resources.
  - **Data sources and data quality** – What are the primary sources of data internally? What additional data might you purchase? How will you ensure quality?
  - **Big data platform and tools** – What platform will you use to build your solution? What software and tools are needed to achieve your purpose?
  - **Metrics for measuring success** – How will you measure system performance? Base your success on how many jobs are submitted, parallel processed, and completed efficiently.
• **Work with your business users to articulate the big opportunities.**

  ✓ Identify and collaborate with business users (analysts, data scientists, marketing professionals, and so on) to find the best business opportunities for big data analytics in your organization. For example, consider an existing business problem—especially one that is difficult, expensive, or impossible to accomplish with your current data sources and analytics systems. Or consider a problem that has never been addressed before because the data sources are new and unstructured.

  ✓ Prioritize your opportunity list and select a project with a discernible return on investment. To determine the best project, consider your answers to these questions:

    - What am I trying to accomplish?
    - Does this project align with strategic business goals?
    - Can I get management support for the project?
    - Does big data analytics hold a unique promise for insight over more traditional analytics?
    - What actions can I take based on the results of my project?
    - What is the potential return on investment to my business?
    - Can I deliver this project with a 6- to 12-month time to value?
    - Is the data that I need available? What do I own? What do I need to buy?
    - Is the data collected in real time, or is it historical data?

**Step 2: Hire the skills you need.**

• **Understand and plan for the skills you need in the business and IT.**

  ✓ What skills do you need to successfully accomplish the initiative? Are those resources in-house?

  ✓ Will you build skills from within the company? Hire new talent? Outsource?

  ✓ Where will these individuals reside in the business? In IT?

**Step 3: Implement your big data solution.**

• **Develop a use case(s) for your project.**

  ✓ Identify the use cases required to carry out your project.

  ✓ Map out data flows to help define what technology and big data capabilities are required to solve the business problem.

  ✓ Decide what data to include and what to leave out. Identify only the strategic data that will lead to meaningful insight.

  ✓ Determine how data interrelates and the complexity of the business rules.

  ✓ Identify the analytical queries and algorithms required to generate the desired outputs.

  ✓ Consider whether you need to support advanced analytics such as interactive queries or predictive analytics, or support real-time data streams.
• **Identify the gaps between current- and future-state capabilities.**

  ✔ What additional data quality requirements will you have for collecting, cleansing, and aggregating data into usable formats?

  ✔ What data governance policies will need to be in place for classifying data; defining its relevance; and storing, analyzing, and accessing it?

  ✔ What infrastructure capabilities will need to be in place to ensure scalability, low latency, and performance, including computing, storage, and network capabilities?

  ✔ Do you need to add specialized components like a NoSQL database for low-latency lookups on large volumes of heterogeneous data?

  ✔ If you plan to process a steady stream of real-time data, what additional infrastructure and memory capabilities will you need? Will you require an MPP in-memory analytics appliance? A CEP solution?

  ✔ Are you considering cloud computing for your delivery model? What type of cloud environment will you use? Private, hybrid, public?

  ✔ How will data be presented to users? Findings need to be delivered in an easy-to-understand way to a variety of business users, from senior executives to information professionals.

• **Develop a test environment for a production version.**

  ✔ Adapt reference architectures to your enterprise. Intel is working with leading partners to develop reference architectures that can help as part of the Intel Cloud Builders program around big data use cases.

  ✔ Define the presentation layer, analytics application layer, data warehousing, and if applicable, private- or public-based cloud data management.

  ✔ Determine the tools users require to present results in a meaningful way. User adoption of tools will significantly influence the overall success of your project.
Intel Resources for Learning More

In addition to the resources already cited in this paper, check the following for further interesting content.

Web Sites

For additional resources about:
- Big data: intel.com/bigdata
- The Internet of Things: intel.com/iot
- The New Center of Possibility: intel.com/centerofpossibility
- Intel Xeon processor E5 family: intel.com/xeone5
- Intel Xeon processor E5 v3 family software solutions: http://transformingbusiness.intel.com
- Intel Xeon processor E7 family: intel.com/xeone7

About Big Data Platforms

Accelerate Big Data Analysis with Intel® Technologies
This paper highlights technologies available from Intel that enterprises can use to scale up Apache Hadoop clusters to handle the increasing volume, variety, and velocity of data. By using fewer, more powerful servers, enterprises can significantly reduce operational costs. intel.com/content/www/us/en/big-data/big-data-analysis-intel-technologies-paper.html

Maximizing Marketing Insight through Big Data Analytics
This white paper describes a collaboration between Intel's Corporate Marketing Group and Intel IT to deploy a marketing analytics program to help Intel allocate marketing spend more effectively within and across media channels. intel.com/content/www/us/en/it-management/intel-it-best-practices/maximizing-marketing-insight-through-big-data-analytics-paper.html

Big Data Mining in the Enterprise for Better Business Intelligence
This white paper from Intel IT describes how Intel is putting in place the systems and skills for analyzing big data to drive operational efficiencies and competitive advantage. Intel IT, in partnership with Intel business groups, is deploying several proofs of concept for a big data platform, including malware detection, chip design validation, market intelligence, and a recommendation system. intel.com/content/www/us/en/it-management/intel-it-best-practices/mining-big-data-in-the-enterprise-for-better-business-intelligence.html

Evaluating Apache Hadoop* Software for Big Data ETL Functions
Read the white paper on how Intel IT evaluated Hadoop for ETL functions and the circumstances in which they gained the most benefit. intel.com/content/www/us/en/it-management/intel-it-best-practices/evaluating-apache-hadoop-software-for-big-data-etr-functions-paper.html

How Intel IT Successfully Migrated to Cloudera Apache Hadoop*
This white paper from Intel IT describes the benefits and best practices for migrating to the Cloudera Distribution for Apache Hadoop. intel.com/content/www/us/en/it-management/how-intel-it-successfully-migrated-to-cloudera-apache-hadoop-paper.html

Leading Advances in the Utilization of Big Data in the Healthcare Industry
This white paper explains the current state of big data adoption in the Japanese healthcare industry and examines solutions offered by Microsoft and Intel that support underlying big data technologies. intel.com/content/www/us/en/healthcare-it/big-data-healthcare-tokyo-paper.html
One Big Data Strategy, Three Massively Parallel Processing Platforms (MPP)
Learn about Intel IT's multiple platform strategy for big data in this Intel IT Business Review that describes the enterprise data warehouse (EDW), Apache Hadoop, and massively parallel processing (MPP) platforms.

Predictive Analytics: Use All Your Data to Compete and Win
How you analyze big data is as important as the data itself. This solution brief describes how organizations can cost-effectively implement an extensible big data platform for descriptive analytics, interactive queries, and predictive analytics.

Turn Big Data into Big Value: A Practical Strategy
Intel innovations in silicon, systems, and software can help you to deploy three usage models (ETL using Apache Hadoop software, interactive queries, and predictive analytics on the Hadoop platform) and other big data solutions with optimal performance, cost, and energy efficiency. software.intel.com/sites/default/files/article/402150/turn-big-data-into-big-value.pdf

About Intel Partner Solutions
For additional solution-oriented content from our partners, visit the New Center of Possibility page on the IT Center at: intel.com/centerofpossibility.

Amazon Web Services*
This solution brief describes how to develop a cost-effective big data engine in the cloud using the Amazon Web Services platform running on servers based on the Intel Xeon processors. intel.com/content/www/us/en/big-data/cloud-analytics-aws-product-brief.html

Dell* In-Memory Appliance for Cloudera* Enterprise
This solution brief describes an easy-to-deploy in-memory appliance solution from Dell, running on servers based on the Intel Xeon processor E7 family and using Cloudera Enterprise and the Apache Spark service for real-time analytics.

SAP HANA* Platform
Find out how to enable real-time analytics by reducing latency with the SAP HANA in-memory database running on servers based on the Intel Xeon processor E7 family. intel.com/content/www/us/en/big-data/real-time-analysis-sap-product-brief.html
Endnotes


2. The IoT will grow to 26 billion units installed in 2020. In addition, the number of smart phones, tablets, and PCs in use will reach about 7.3 billion units. Together, this means more than 33 billion connected devices will be in use by 2020. Source: “Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020.” Gartner (December 12, 2013). [gartner.com/newsroom/id/2636073](http://gartner.com/newsroom/id/2636073)


4. Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com.

5. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests such as SYSmark* and MobileMark* are measured using specific computer systems, components, software, operations, and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

6. Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

7. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.


9. Cluster configuration: one head node (name node, job tracker), 10 workers (data nodes, task trackers), Cisco Nexus* 5020 10 gigabit switch.

Baseline worker node: Supermicro* SYS-1026T-URF 1U servers with two Intel Xeon processors X5690 series at 3.47 gigahertz (GHz), 48 GB RAM, 700 GB 7,200 RPM SATA hard drives, Intel Ethernet Server Adapter I350-T2, Apache Hadoop 1.0.3 software, Red Hat* Enterprise Linux 6.3 operating system, Oracle Java* 1.7.0_05 platform.

Upgraded processor and base system in worker node: Dell* PowerEdge* R720 2U servers with two Intel Xeon processors E5-2690 product family at 2.90 GHz, 128 GB RAM.

Upgraded storage in worker node: Intel Solid-State Drive 520 Series.

Upgraded software in worker node: Intel Distribution for Apache Hadoop software 2.1.1.

10. Huang, Shengsheng, Jie Huang, Jinquan Dai, Tao Xie, Bo Huang. The HiBench Benchmark Suite: Characterization of the MapReduce-Based Data Analysis. IEEE (March 2010).
More from the Intel IT Center

Planning Guide: Getting Started with Big Data is brought to you by the Intel IT Center. The Intel IT Center is designed to provide IT professionals with straightforward, fluff-free information to help them implement strategic projects on their agenda, including virtualization, data center design, big data, cloud, and client and infrastructure security.

Visit the Intel IT Center for:

- Thought leadership on data center and business client trends and perspectives on innovations to watch
- Planning guides, case studies, and solution spotlights to help you implement key projects
- Information on how Intel’s own IT organization is implementing cloud, virtualization, security, and other strategic initiatives
- Information on events where you can hear from Intel product experts as well as from Intel’s own IT professionals

Learn more at intel.com/ITCenter.

Share with Colleagues

This paper is for informational purposes only. THIS DOCUMENT IS PROVIDED "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION, OR SAMPLE. Intel disclaims all liability, including liability for infringement of any property rights, relating to use of this information. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

Copyright © 2015 Intel Corporation. All rights reserved. Intel, the Intel logo, the Experience What’s Inside logo, and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.