Intel IT Center

FEBRUARY 2013

Planning Guide

Getting Started with Big Data

Steps IT Managers Can Take to Move Forward with Apache Hadoop* Software

Why You Should Read This Document

This planning guide provides valuable information and practical steps for IT managers who want to plan and implement big data analytics initiatives and get started with Apache Hadoop* software, including:

- The IT landscape for big data and the challenges and opportunities associated with this disruptive force
- Introduction to Hadoop* software, the emerging standard for gaining insight from big data, including processing and analytic tools (Apache Hadoop MapReduce, Apache HBase* software)
- Guidance on how to get the most out of Hadoop software with a focus on areas where Intel can help, including infrastructure technology, optimizing, and tuning
- Five basic "next steps" and a checklist to help IT managers move forward with planning and implementing their own Hadoop project

Sponsors of Tomorrow.™ intel

Intel IT Center

FEBRUARY 2013

Planning Guide

Getting Started with Big Data

Steps IT Managers Can Take to Move Forward with Apache Hadoop* Software

Sponsors of Tomorrow.™ intel

Contents

The IT Landscape for Big Data Analytics

The buzz about big data analytics is growing louder.

Today, every organization across the globe is faced with an unprecedented growth in data. Imagine this: The digital universe of data was expected to expand to 2.7 zettabytes (ZB) by the end of 2012. Then it's predicted to double every two years, reaching 8 ZB of data by 2015.[1] It's hard to conceptualize this quantity of information, but here's one way: If the U.S. Library of Congress holds 462 terabytes (TB) of digital data, then 8 ZB is equivalent to almost 18 million Libraries of Congress.[2] That's really big data.

The Value of Big Data

What exactly is "big data," and where is it coming from?

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, including text, document, image, video, and more.

What about the 8 ZB of data projected for 2015? Nearly 15 billion connected devices—including 3 billion Internet users plus machine-to-machine connections—will contribute to this ocean of data.[3]

The real value of big data is in the insights it produces when analyzed—finding patterns, deriving meaning, making decisions, and ultimately responding to the world with intelligence.

Using Big Data Analytics to Win

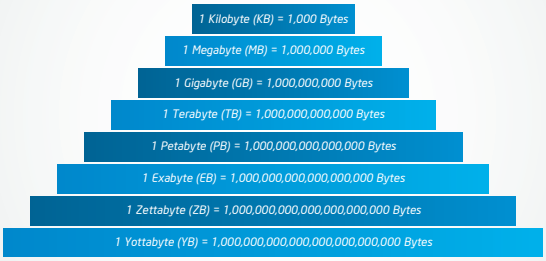
Big data is a disruptive force, presenting opportunities as well as challenges to IT organizations. A study by the McKinsey Global Institute established that data is as important to organizations as labor and capital.[4] The study concluded that if organizations can effectively capture, analyze, visualize, and apply big data insights to their business goals, they can differentiate themselves from their competitors and outperform them in terms of operational efficiency and the bottom line.

Big data analytics represents a significant challenge for IT organizations—and yet according to an Intel survey of 200 IT managers, 84 percent are already analyzing unstructured data, and 44 percent of those that aren't expect to do so by 2014.[5] The potential for big data is irresistible.

The three Vs characterize what big data is all about, and also help define the major issues that IT needs to address:

- **Volume.** The massive scale and growth of unstructured data outstrips traditional storage and analytical solutions.
- **Variety.** Traditional data management processes can't cope with the heterogeneity of big data—or "shadow" or "dark data," such as access traces and Web search histories.
- **Velocity.** Data is generated in real time, with demands for usable information to be served up immediately.

A Mountain of Data



Big data is measured in terabytes, petabytes, and even exabytes. Put it all in perspective with this handy conversion chart.

What Big Data Analytics Is (and Isn't)

Big data analytics is clearly a game changer, enabling organizations to gain insights from new sources of data that haven't been mined in the past. Here's more about what big data analytics is ... and isn't.

Big Data Analytics Is ...

- A technology-enabled strategy for gaining richer, deeper insights into customers, partners, and the business—and ultimately gaining competitive advantage.
- Working with data sets whose size and variety is beyond the ability of typical database software to capture, store, manage, and analyze.
- Processing a steady stream of real-time data in order to make time-sensitive decisions faster than ever before.
- Distributed in nature. Analytics processing goes to where the data is for greater speed and efficiency.
- A new paradigm in which IT collaborates with business users and "data scientists" to identify and implement analytics that will increase operational efficiency and solve new business problems.
- Moving decision making down in the organization and empowering people to make better, faster decisions in real time.

Big Data Analytics Isn't ...

- Just about technology. At the business level, it's about how to exploit the vastly enhanced sources of data to gain insight.
- Only about volume. It's also about variety and velocity. But perhaps most important, it's about value derived from the data.
- Generated or used only by huge online companies like Google or Amazon anymore. While Internet companies may have pioneered the use of big data at web scale, applications touch every industry.
- About "one-size-fits-all" traditional relational databases built on shared disk and memory architecture. Big data analytics uses a grid of computing resources for massively parallel processing (MPP).
- Meant to replace relational databases or the data warehouse. Structured data continues to be critically important to companies. However, traditional systems may not be suitable for the new sources and contexts of big data.

The Purpose of This Guide

The remainder of this guide will describe emerging technologies for managing and analyzing big data, with a focus on getting started with the Apache Hadoop* open-source software framework, which provides the framework for distributed processing of large data sets across clusters of computers. We'll also provide five practical steps you can take to begin planning your own big data analytics project using this technology.

Your New Best Friend: The Data Scientist

A new kind of professional is helping organizations make sense of the massive streams of digital information: the data scientist.

Data scientists are responsible for modeling complex business problems, discovering business insights, and identifying opportunities. They bring to the job:

- Skills for integrating and preparing large, varied data sets
- Advanced analytics and modeling skills to reveal and understand hidden relationships
- Business knowledge to apply context
- Communication skills to present results

Data science is an emerging field. Demand is high, and finding skilled personnel is one of the major challenges associated with big data analytics. A data scientist may reside in IT or the business—but either way, he or she is your new best friend and collaborator for planning and implementing big data analytics projects.

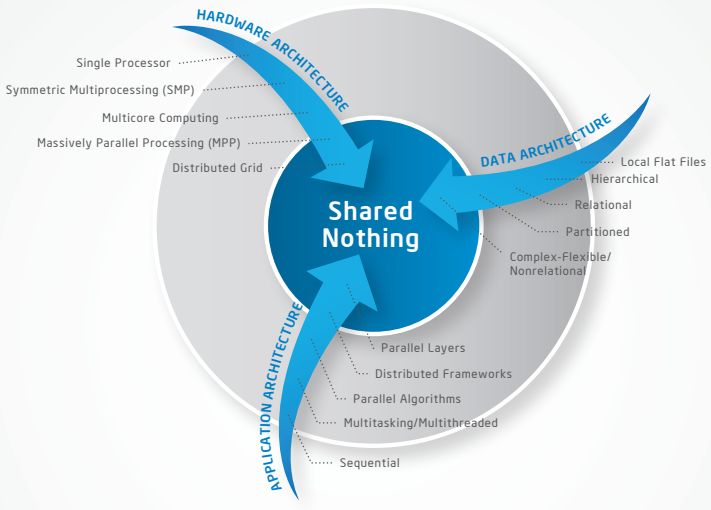
Emerging Technologies for Managing Big Data

For organizations to realize the full potential of big data, they must find a new approach to capturing, storing, and analyzing data. Traditional tools and infrastructure aren't as efficient working with larger and more varied data sets coming in at high velocity.

New technologies are emerging to make big data analytics scalable and cost-effective. One new approach uses 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.

Shared Nothing Architecture for Massively Scalable Systems

The new shared nothing architecture can scale with the huge volumes, variety, and speed requirements of big data by distributing the work across dozens, hundreds, or even thousands of commodity servers that process the data in parallel. First implemented by large community research projects such as SETI@home and online services such as Google* and Amazon*, each node is independent and stateless, so that shared nothing architecture scales easily—simply add another node—enabling systems to handle growing processing loads.



Shared nothing architecture is possible because of the convergence of advances in hardware, data management, and analytic applications technologies.

Source: "Data rEvolution." CSC Leading Edge Forum (2011).

Processing is pushed out to the nodes where the data resides. This is completely different from a traditional approach, which retrieves data for processing at a central point.

Ultimately, the data must be reintegrated to deliver meaningful results. Distributed processing software frameworks make the computing grid work by managing and pushing the data across machines, sending instructions to the networked servers to work in parallel, collecting individual results, and then reassembling them for the payoff.

Distributed Processing Frameworks and the Emergence of Apache Hadoop

Hadoop* is evolving as the best new approach to big data analytics. An outgrowth of the Apache Nutch* open-source Web search project,[6] Hadoop is a software framework that provides a simple programming model to enable distributed processing of large data sets on clusters of computers. The framework easily scales on hardware such as servers based on Intel® Xeon® processors.

Hadoop software is a complete open-source framework for big data analytics. It includes a distributed file system, a parallel processing framework called Apache Hadoop MapReduce, and several components that support the ingestion of data, coordination of workflows, management of jobs, and monitoring of the cluster. Hadoop is more cost-effective at handling large unstructured data sets than traditional approaches.

Hadoop offers several key advantages for big data analytics, including:

- **Store any data in its native format.** Because data does not require translation to a specific schema, no information is lost.
- **Scale for big data.** Hadoop is already proven to scale by companies like Facebook and Yahoo!, which run enormous implementations.
- **Deliver new insights.** Big data analytics is uncovering hidden relationships that have been difficult, time consuming, and expensive—or even impossible—to address using traditional data mining approaches.
- **Reduce costs.** Hadoop open-source software runs on standard servers and has a lower cost per terabyte for storage and processing. Storage can be added incrementally as needed, and hardware can be added or swapped in or out of a cluster.
- **Higher availability.** Hadoop recovers from hardware, software, and system failures by providing fault tolerance through replication of data and failover across compute nodes.
- **Lower risk.** The Hadoop community is active and diverse, with developers and users from many industries across the globe. Hadoop is a technology that will continue to advance.

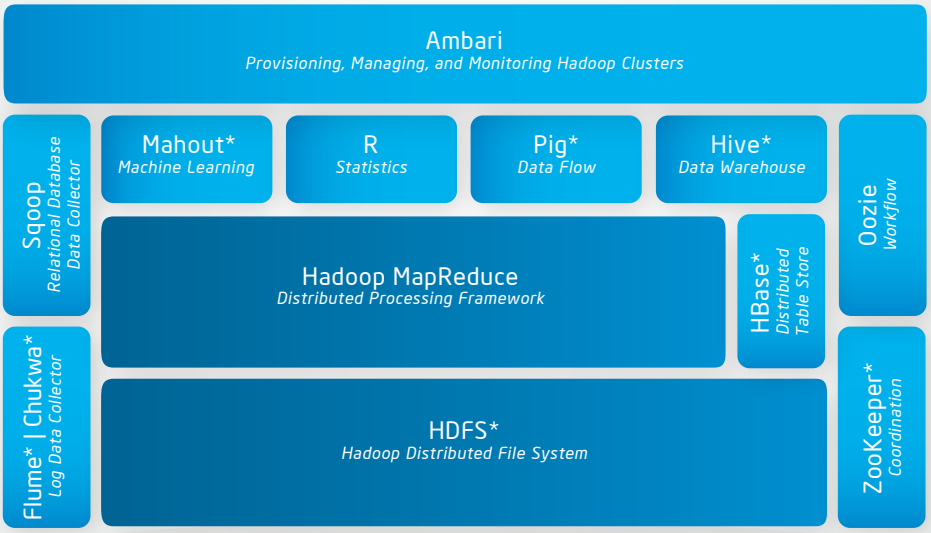
Big Data and the Cloud

Big data requires clusters of servers to support the tools that process large volumes, high velocity, and varied formats of big data. Clouds are already deployed on pools of servers and can scale up or down as needed for big data.

As a result, cloud computing offers a cost-effective way to support big data technologies and the advanced analytics applications that can drive business value.

Find out more about how big data can work in the cloud in *Big Data in the Cloud: Converging Technologies* at intel.com/content/www/us/en/big-data/big-data-cloud-technologies-brief.html.

Apache Hadoop* Stack



The Hadoop* software stack includes a number of components.

Does Hadoop* Software Replace My Existing Database Systems?

Hadoop* software is a massively scalable storage and data processing system—it's not a database. In fact, it supplements your existing systems by handling data that's typically a problem for them. Hadoop can simultaneously absorb and store any type of data from a variety of sources, aggregate and process it in arbitrary ways, and deliver it wherever it's needed—which could be serving up real-time transactional data or providing interactive business intelligence via your existing systems.

What about Apache Hadoop MapReduce?

MapReduce is the software programming framework in the Hadoop stack that simplifies processing of big data sets and gives programmers a common method for defining and orchestrating complex processing tasks across clusters of computers. MapReduce applications work like this: The map task splits a data set into independent chunks to be processed in parallel. The map outputs are sorted and then submitted to the reduce task. Both input and output are stored in Apache* Hadoop Distributed File System (HDFS*) or other storage such as Amazon S3, part of Amazon Web Services. Typically the data is processed and stored on the same node, making it more efficient to schedule tasks where data already resides and resulting in high aggregate bandwidth across the node.

MapReduce simplifies the application programmer's work by taking care of scheduling jobs, monitoring activity, and reexecuting failed tasks.

Apache Hadoop* at a Glance

Apache Hadoop* is a community-led effort that includes three key development subprojects as well as other related initiatives.

Key Development Subprojects	
Apache* Hadoop* Distributed File System (HDFS*)	The primary storage system that uses multiple replicas of data blocks, distributes them on nodes throughout a cluster, and provides high-throughput access to application data
Apache Hadoop MapReduce	A programming model and software framework for applications that perform distributed processing of large data sets on compute clusters
Apache Hadoop Common	Utilities that support the Hadoop framework, including FileSystem (an abstract base class for a generic file system), remote-procedure call (RPC), and serialization libraries
Other Related Hadoop Projects	
Apache Avro*	A data serialization system
Apache Cassandra*	A scalable, multimaster database with no single point of failure
Apache Chukwa*	A data collection system for monitoring large distributed systems built on HDFS and MapReduce; includes a toolkit for displaying, monitoring, and analyzing results
Apache HBase*	A scalable, distributed database that supports structured data storage for large tables; used for random, real-time read/write access to big data
Apache Hive*	A data warehouse infrastructure that provides data summarization, ad hoc querying, and the analysis of large data sets in Hadoop-compatible file systems
Apache Mahout*	A scalable machine learning and data mining library with implementations of a wide range of algorithms, including clustering, classification, collaborative filtering, and frequent-pattern mining
Apache Pig*	A high-level data-flow language and execution framework for expressing parallel data analytics
Apache ZooKeeper*	A high-performance, centralized coordination service that maintains configuration information and naming and provides distributed synchronization and group services for distributed applications

Source: Apache Hadoop, hadoop.apache.org.

Hear from Apache Hadoop* Experts

One way to learn about Apache Hadoop* software and its components is to hear directly from experts deeply engaged in the open-source community and its development work. Listen to the podcasts of interviews of community leaders for Apache Hadoop MapReduce, Apache* HDFS*, Apache Hive*, Apache Pig*, and Apache HCatalog, describing how each works, where it fits in the Hadoop* stack, and plans for continued development. PDFs accompany each podcast.

Hadoop* Adoption

As more and more enterprises recognize the value and advantages associated with big data insights, adoption of Hadoop software is growing. The Hadoop open-source technology stack includes an open-source implementation of MapReduce, HDFS, and the Apache HBase* distributed database that supports large, structured data tables.

After six years of refinements, Apache released the first full production version of Apache Hadoop 1.0 software in January 2012. Among the certified features supported in this version are HBase*, Kerberos security enhancements, and a representational state transfer (RESTful) API to access HDFS.[7]

Hadoop software can be downloaded from one of the Apache download sites. Because Hadoop software is an open-source, volunteer project, the Hadoop wiki provides information about getting help from the community as well as links to tutorials and user documentation for implementing, troubleshooting, and setting up a cluster.

Accelerating Big Data Analytics Standards

The Open Data Center Alliance (ODCA), an independent IT consortium comprised of global IT leaders from more than 300 companies, recently announced the formation of the Data Services Workgroup to document the most urgent requirements facing IT for data management. The workgroup will focus initially on creating usage model requirements that address security, manageability, and interoperability of emerging big data frameworks with traditional data management and data warehouse solutions. Based on the usage models, workgroup members will develop reference architectures and proofs of concept for commercial distribution with independent software vendors and OEM partners to test deployments and establish solutions for the enterprise market. The alliance will also collaborate with the open-source community to drive benchmarking suites. As technical advisor to ODCA, Intel will play a major role in the development of standards and best practices for big data analytics.

The Hadoop Ecosystem

The Hadoop ecosystem is a complex landscape of vendors and solutions that includes established players and several newcomers. Several vendors offer their own Hadoop distribution, packaging the basic stack with other Hadoop software projects such as Apache Hive*, Apache Pig*, and Apache Chukwa*. Some of these distributions can integrate with data warehouses, databases, and other data management products so that data can move between Hadoop clusters and other environments to expand the pool of data to process or query.

Other vendors provide Hadoop management software that simplifies administration and troubleshooting. A third group delivers products that help developers write Hadoop applications, provide search capabilities, or analyze data without using MapReduce. These products sit on top of platform software and include abstraction layers that marry a Structured Query Language (SQL) data warehouse to a Hadoop cluster as well as real-time processing and analytics. Finally, there's growing interest in offering subscription services via the cloud.

Apache Hadoop* Goes Commercial

Apache Hadoop*-related offerings are in market in several categories. The following vendors are a sample of the growing Hadoop* ecosystem.

Category	Vendor/Offering
Integrated Hadoop systems	▪ EMC* Greenplum* ▪ HP* Big Data Solutions ▪ IBM* InfoSphere* ▪ Microsoft* Big Data Solution ▪ Oracle* Big Data Appliance
Hadoop applications and analytical databases with Hadoop connectivity	▪ Datameer* Analytics Solution ▪ Hadapt* Adaptive Analytic Platform* ▪ HP Vertica* Analytics Platform ▪ Karmasphere* Analyst ▪ ParAccel* Analytic Platform ▪ Pentaho* Data Integration ▪ Splunk* Enterprise* ▪ Teradata* Aster* Solution
Hadoop distributions	▪ Cloudera's Distribution including Apache Hadoop (CDH) ▪ EMC Greenplum HD ▪ Hortonworks ▪ IBM InfoSphere BigInsights ▪ Intel® Distribution for Apache Hadoop Software ▪ MapR* M5 Edition ▪ Microsoft Big Data Solution ▪ Platform Computing* MapReduce
Cloud-based solutions	▪ Amazon* Web Services ▪ Google* BigQuery

See Big Data Vendor Spotlights for some of the Intel partners who offer big data solutions.

Note: The Hadoop ecosystem is emerging rapidly. This list is adapted from two sources: Dumbill, Edd. "Big Data Market Survey: Hadoop Solutions." *O'Reilly Radar* (January 19, 2012). http://radar.oreilly.com/2012/01/big-data-ecosystem.html and *Data rEvolution: CSC Leading Edge Forum*. CSC (2011). http://assets1.csc.com/lef/downloads/LEF_2011Data_rEvolution.pdf

Two Approaches to Using Hadoop Software for Big Data Analytics

Enterprises are taking two basic approaches to implementing Hadoop.

Hadoop-only deployments. Hadoop deployments are available as open-source software that can be downloaded free from Apache or as distributions from vendors that prepackage the Hadoop framework with certain components and management software to support system administration.

Hadoop-only deployments are ideal for building a big data management platform for unstructured data analytics and insight. Open-source tools also make it possible to query structured data using MapReduce applications, HBase, or Hive*.

Hadoop integrated with traditional databases. Organizations with traditional data warehousing and analytics in place can extend their existing platform to include an integrated Hadoop implementation. Connecting existing data management resources to Hadoop software provides an opportunity to tap both structured and unstructured data for insights. For example, analysis of complex call center transcripts can be married to structured data about buying behavior, such as specific SKUs, retail outlets, geographies, and so on. In this case, proprietary connectors are used to move data back and forth from Hadoop to traditional environments.

Intel® Distribution for Apache Hadoop* Software

Intel® Distribution for Apache Hadoop* software (Intel Distribution) includes Apache Hadoop and other software components optimized by Intel with hardware-enhanced performance and security capabilities. Designed to enable a wide range of data analytics on Apache Hadoop, Intel Distribution is optimized for Apache Hive* queries, provides connectors for R* for statistical processing, and enables graph analytics using Intel Graph Builder for Apache Hadoop software, a library to construct large data sets into graphs to help visualize relationships between data. Included in the Intel Distribution, Intel Manager for Apache Hadoop provides a management console that simplifies the deployment, configuration, and monitoring of a Hadoop* deployment.

Intel Distribution is available worldwide today for evaluation. Technical support is provided currently in the United States, China, and Singapore, with other geographies expected later in the year.

Find out more about the Intel Distribution.

Deploying Hadoop in Your Data Center

Big data analytics is a technology-enabled strategy that is much more than the hardware and software that support it. Nevertheless, as an IT manager, the responsibility for implementing big data initiatives in your data center will fall to you. Hadoop deployments can have very large infrastructure requirements, and hardware and software choices made at design time can have significant impact on performance and total cost of ownership. Data centers can get the most from their Hadoop deployments by ensuring that the right infrastructure is in place and that Hadoop software is optimized and tuned for best performance.

Put the Right Infrastructure in Place

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. This is where the data is stored and processed. The combination of Hadoop infrastructure with standard server platforms provides the foundation for a cost-efficient and high-performance analytics platform for parallel applications.

Setting Up Hadoop System Architecture

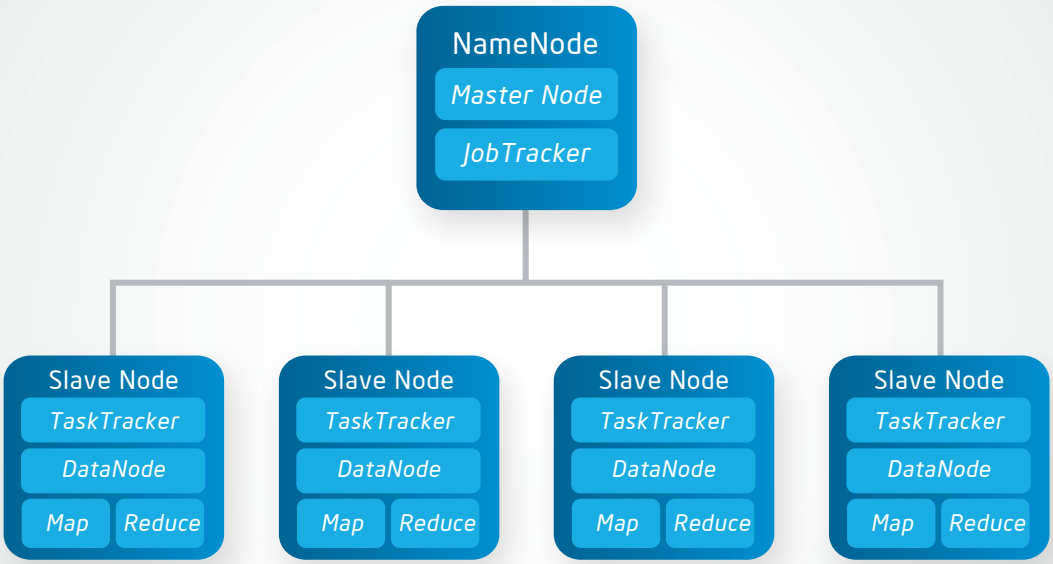
Each cluster has one "master node" with multiple slave nodes. The master node uses the NameNode and JobTracker functions to coordinate slave nodes to get the job done. The slaves use the TaskTracker function to manage the jobs scheduled by JobTracker, HDFS to store data, and map and reduce functions for data computation. The basic software stack includes Hive and Pig* for language and compilers, HBase for NoSQL database management, and Apache Sqoop and Apache Flume* for log collection. Apache ZooKeeper* provides centralized coordination for the stack.

The Cost of Big Data Analytics

A 2012 survey from *InformationWeek* tackles the question of big data economics, finding that budget constraints and other cost-related issues are top barriers for IT managers. Building your own Apache Hadoop* deployment and investing in storage and development resources or implementing a proprietary vendor solution can incur significant costs. While the cloud offers some potential relief, pricing models for public cloud providers may not offer enough. With storage and computing costs continuing to decline, deploying and managing your own Hadoop* clusters may provide the best economics over both public cloud and vendor systems—even adding in the cost of a skilled person to manage the hardware.

Source: Biddick, Michael. "The Big Data Management Challenge." *InformationWeek* (April 2012). http://reports.informationweek.com/abstract/81/8766/business-intelligence-and-information-management/research-the-big-data-management-challenge.html

Apache Hadoop* Deployment on a Cluster of Standard Server Nodes

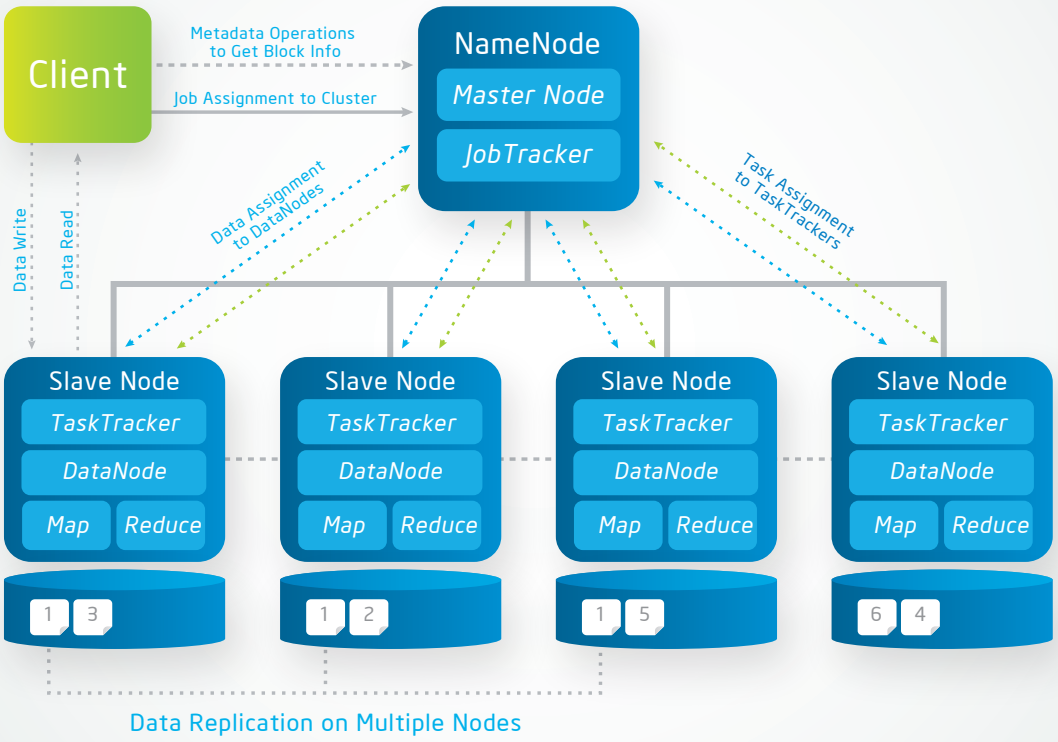


Source: *Intel® Cloud Builders Guide to Cloud Design and Deployment on Intel® Platforms: Apache Hadoop**. Intel (February 2012). intelcloudbuilders.com/docs/Intel_Cloud_Builders_Hadoop.pdf

Operating a Server Cluster

A client submits a job to the master node, which orchestrates with the slaves in the cluster. JobTracker controls the MapReduce job, reporting to TaskTracker. In the event of a failure, JobTracker reschedules the task on the same or a different slave node, whichever is most efficient. HDFS is location aware or rack aware and manages data within the cluster, replicating the data on various nodes for data reliability. If one of the data replicas on HDFS is corrupted, JobTracker, aware of where other replicas are located, can reschedule the task right where it resides, decreasing the need to move data back from one node to another. This saves network bandwidth and keeps performance and availability high. Once the job is mapped, the output is sorted and divided into several groups, which are distributed to reducers. Reducers may be located on the same node as the mappers or on another node.

Operation of an Apache Hadoop* Cluster



Jobs are orchestrated by the master node and processed on the slave nodes.

Hadoop Infrastructure: Big Data Storage and Networking

Hadoop clusters are enhanced by dramatic improvements in mainstream compute and storage resources and are complemented by 10 gigabit Ethernet (10 GbE) solutions. The increased bandwidth associated with 10 GbE is critical to importing and replicating large data sets across servers. Intel Ethernet 10 Gigabit Converged Network Adapters 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.

Get the guide to building balanced, cost-effective Hadoop clusters on 10 GbE.

Intel® Architecture: High-Performance Clusters

Using Intel® Xeon® processor E5 family-based servers as the baseline server platform for the cluster, a team of Intel big data, network, and storage experts measured Apache Hadoop* performance results for various combinations of networking and storage components. In general, the following define "good, better, and best performance" options for Intel-based infrastructure for your big data environment. (Note that certain variables could impact results for your data center.)

Performance	Server	Networking	Storage
Good	Intel Xeon processor E5 family	Gigabit Ethernet (GbE) or 10 GbE	Hard drives
Better	Intel Xeon processor E5 family	10 GbE	Hard drives and solid-state drives (SSDs) with tiered storage capabilities
Best	Intel Xeon processor E5 family	10 GbE	SSDs

Get more detail about the performance of each platform combination.

Optimize and Tune 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 fellow travelers such as HP, Super Micro, and 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 you 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 in a lower total cost of ownership for the Hadoop environment.

Benchmarking

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.[8] The individual measures represent 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 and machine learning. HiBench 2.1 is now available as open source under Apache License 2.0 at https://github.com/hibench/HiBench-2.1.

HiBench: The Details

Intel's HiBench suite looks at 10 workloads in four categories.

Category	Workload	Description
Microbenchmarks	Sort	▪ This workload sorts its binary input data, which is generated using the Apache Hadoop* RandomTextWriter example. ▪ Representative of real-world MapReduce jobs that transform data from one format to another.
	WordCount	▪ This workload counts the occurrence of each word in the input data, which is generated using Hadoop* RandomTextWriter. ▪ Representative of real-world MapReduce jobs that extract a small amount of interesting data from a large data set.
	TeraSort	▪ A standard benchmark for large-size data sorting that is generated by the TeraGen program.
	Enhanced DFSIO	▪ Tests Apache* HDFS* system throughput of a Hadoop cluster. ▪ Computes the aggregated bandwidth by sampling the number of bytes read or written at fixed time intervals in each map task.
Web search	Apache Nutch* Indexing	▪ This workload tests the indexing subsystem in Nutch*, a popular Apache open-source search engine. The crawler subsystem in the Nutch engine is used to crawl an in-house Wikipedia* mirror and generates 8.4 GB of compressed data (for about 2.4 million web pages) total as workload input. ▪ Large-scale indexing system is one of the most significant uses of MapReduce (for example, in Google* and Facebook* platforms).
	Page Rank	▪ This workload is an open-source implementation of the page-rank algorithm, a link-analysis algorithm used widely in Web search engines.
Machine learning	K-Means Clustering	▪ Typical application area of MapReduce for large-scale data mining and machine learning (for example, in Google and Facebook platforms). ▪ K-Means is a well-known clustering algorithm.
	Bayesian Classification	▪ Typical application area of MapReduce for large-scale data mining and machine learning (for example, in Google and Facebook platforms). ▪ This workload tests the naive Bayesian (a well-known classification algorithm for knowledge discovery and data mining) trainer in the Apache Mahout* open-source machine learning library.
Analytical query	Apache Hive* Join	▪ This workload models complex analytic queries of structured (relational) tables by computing the sum of each group over a single read-only table.
	Hive* Aggregation	▪ This workload models complex analytic queries of structured (relational) tables by computing both the average and sum for each group by joining two different tables.

Five Steps and a Checklist: Get Started with Your Big Data Analytics Project

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 emerging technologies that can help you get insights out of these 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 Hadoop initiatives.

You can get started with your big data analytics project by following these five steps.

Step 1: 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.
- ☐ Determine the skills you need to successfully accomplish your initiative.

Step 2: Do your research to get up to speed on the technology.

- ☐ Talk with your peers in IT.
- ☐ Take advantage of Intel IT Center resources for big data.
- ☐ Understand vendor offerings.
- ☐ Take tutorials and examine user documentation offered by Apache.

Step 3: Develop 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.

Step 4: Identify 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?
- ☐ 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.

Step 5: 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

About Big Data

Big Data Analytics (Collection Page)

This web page aggregates key Intel resources that can help you implement your own big data initiatives. Visit this page in the Intel IT Center for planning guides, peer research, vendor solution information, and real-world case studies.
intel.com/bigdata

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

Inside IT: Big Data

In this podcast, Moty Fania, who leads Intel's strategy team for big data for business intelligence, talks about developing the necessary skills and the right platform to deal with big data.
http://connectedsocialmedia.com/intel/5773/inside-it-big-data/

Peer Research: Big Data Analytics

Read the results of a survey of 200 IT managers that provide insights into how organizations are using big data analytics today, including what organizations need to move forward and what the research means for the IT industry. Highlights are reported in the video "IT Managers Speak Out about Big Data Analytics."
intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html

Big Thinkers on Big Data

A series of interviews with thought leaders about big data, including LiveRamp CEO Auren Hoffman on the big data revolution driving business competition; Forrester Principal Analyst Mike Gualtieri on what's next; and Cognito CEO Joshua Feast on big data, human behavior, and business outcomes.
intel.com/content/www/us/en/big-data/big-thinkers-on-big-data.html

About Hadoop Software

Apache Hadoop Spotlights

Visit this page to hear from Apache Hadoop open-source community experts explaining how software components of the Hadoop stack work and where future development will lead. Podcasts of interviews are with Alan Gates (Hortonworks) on HCatalog and Pig, Konstantin Shvachko (AltoScale) on HDFS, Deveraj Das (Hortonworks) on MapReduce, and Carl Steinbach (Cloudera) on Hive.
intel.com/content/www/us/en/big-data/big-data-apache-hadoop-framework-spotlights-landing.html

Intel® Cloud Builders Guide to Cloud Design and Deployment on Intel® Platforms: Apache Hadoop**

This reference architecture is for organizations that want to build their own cloud computing infrastructure, including Apache Hadoop clusters to manage big data. It includes steps for setting up the deployment at your data center lab environment and contains details on Hadoop topology, hardware, software, installation and configuration, and testing. Implementing this reference architecture will help you get started building and operating your own Hadoop infrastructure.
intelcloudbuilders.com/docs/Intel_Cloud_Builders_Hadoop.pdf

Optimizing Hadoop Deployments*

This white paper provides guidance to organizations as they plan Hadoop deployments. Based on extensive lab testing with Hadoop software at Intel, it describes best practices for establishing server hardware specifications, discusses the server software environment, and provides advice on configuration and tuning that can improve performance.
intel.com/content/www/us/en/cloud-computing/cloud-computing-optimizing-hadoop-deployments-paper.html

Additional Resources

Big Data: Harnessing a Game-Changing Asset

This report from the Economist Intelligence Unit and sponsored by SAS looks at big data and its impact on companies. The survey examined the organizational characteristics of companies already adept at extracting value from the data and found a strong link between effective data management and financial performance. These companies can provide models for how organizations need to evolve to effectively manage and gain value from big data.
sas.com/resources/asset/SAS_BigData_final.pdf

The Forrester™ Wave: Enterprise Hadoop Solutions, Q1 2012

This report by James Kobielus at Forrester reviews 13 enterprise Hadoop solutions providers, applying a 15-criteria evaluation to each. Leaders include Amazon Web Services, IBM, EMC Greenplum, MapR, Cloudera, and Hortonworks.
forrester.com/The+Forrester+Wave+Enterprise+Hadoop+Solutions+Q1+2012/quickscan/-/E-RES60755

Endnotes

1. Gens, Frank. *IDC Predictions 2012: Competing for 2020.* IDC (December 2011). http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf
2. "Big Data Infographic and Gartner 2012 Top 10 Strategic Tech Trends." *Business Analytics 3.0* (blog) (November 11, 2011). http://practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and-gartner-2012-top-10-strategic-tech-trends/
3. "Global Internet Traffic Projected to Quadruple by 2015." *The Network* (press release) (June 1, 2011). http://newsroom.cisco.com/press-release-content?type=webcontent&articleId=324003
4. *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* McKinsey Global Institute (May 2011). mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.pdf
5. *Peer Research on Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data.* Intel (August 2012). intel.com/content/www/us/en/big-data/data-insights-peer-research-report.html
6. Nutch* software was initially an independent open-source project originated by Doug Cutting and Mike Cafarella. In 2005, Nutch began being managed by the Apache Software Foundation, first as a subproject of Apache Lucene* search software, and then in 2010 as a top-level project of the Apache Software Foundation. Source: "Nutch Joins Apache Incubator" (press release). Apache Software Foundation (January 2005). nutch.apache.org/#January+2005%3A+Nutch+Joins+Apache+Incubator
7. "Hadoop Hits Primetime with Production Release." Datanami (January 6, 2012). datanami.com/datanami/2012-01-06/hadoop_hits_primetime_with_production_release.html
8. 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, Intel's program for IT professionals. The Intel IT Center is designed to provide straightforward, fluff-free information to help IT pros implement strategic projects on their agenda, including virtualization, data center design, cloud, and client and infrastructure security. Visit the Intel IT Center for:

- Planning guides, peer research, and solution spotlights to help you implement key projects
- Real-world case studies that show how your peers have tackled the same challenges you face
- 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

0213/RF/ME/PDF-USA 328687-001

Sponsors of Tomorrow.™