White Paper

Getting Real About Big Data: Build Versus Buy

By Nik Rouda, Senior Analyst

July 2014

This ESG White Paper was commissioned by Oracle and is distributed under license from ESG.

Contents

Executive Summary

In Search of Big Data Infrastructure

Enterprises of all types are betting on big data analytics to help them better understand customers, compete in the market, improve products and services, tune operations, and increase profits. In a recent ESG survey, 26% of IT and 53% of marketing professionals responsible for their organizations' strategies, technologies, and processes considered enhancing analytics a top spending priority.[1] Given the high degree of interest in analytics and the vast quantity of unutilized data, the first step most organizations need to take on the path toward realizing the benefits of big data analytics is to implement an appropriate infrastructure to process big data.

Though most enterprises are not starting entirely from scratch, having already developed data warehouses and related business intelligence (BI) solutions, most realize that big data analytics require a different infrastructure than what they have used historically for data warehousing and BI. Many organizations, therefore, plan to invest in a new solution infrastructure to realize the promise of big data. ESG research has found that 56% of respondents responsible for data initiatives expect their investments to increase by more than 10% in 2014 compared with 2013[2]. ESG also discovered that buying purpose-built appliances for use on-premises, such as Oracle's Big Data Appliance, will prove to be the preferred option for 22% of organizations seeking to implement a big data infrastructure.[3]

Do-it-yourself Hadoop Can Be Difficult

Web 2.0 companies, such as Google and Yahoo, have successfully built big data infrastructures from scratch. Those same firms also have been primary participants in the birth and nurturing of Hadoop, the Apache open source project deservedly given credit for catalyzing the big data movement. Hadoop has matured over the past year, due in part to feature enhancements and in part to growing support from a widening range of both startups and more established IT vendors.

For many organizations, Hadoop-based solutions will represent the first new explicitly big data investment. However, successful Hadoop implementations put into full production using do-it-yourself infrastructure components may require more time and effort than initially expected. Hadoop lures many big data hopefuls due to its apparent low infrastructure cost and easy access; as an open source technology, anyone can download Hadoop for free, and can spin up a simple Hadoop infrastructure in the cloud or on-premises. Unfortunately, many organizations also quickly find that they lack the time, resources, and expertise to make do-it-yourself Hadoop infrastructure work for big data, with reported shortages of skilled staff in IT architecture and planning, business intelligence and analytics, and/or database administration at approximately 20% of companies.[4]

The still somewhat rare and expensive expertise comes in two forms: (1) The Hadoop engineer, who can architect an initial Hadoop infrastructure, feed applicable data in, help the data analyst squeeze useful analytics out from the data repository, and evolve and manage the whole infrastructure over time; and (2) The data scientist or analyst, who knows how to render the tools of statistics in the context of big data analytics, and also can lead the human and business process of discovery and collaboration in order to yield actionable results.

Thus, despite the hope and hype, Hadoop on a commodity stack of hardware does not always offer a lower cost ride to big data analytics, or at least not as quickly as hoped. ESG asserts that many organizations implementing Hadoop infrastructures based on human expertise plus a purchased commodity hardware and software infrastructure may experience unexpected costs, slower speed to market, and unplanned complexity throughout the lifecycle.

[1] Source: ESG Research Report, *2014 IT Spending Intentions Survey*, February 2014.
[2] Source: ESG Research Report, *Enterprise Data Analytics Trends*, May 2014.
[3] Ibid.
[4] Source: ESG Research Report, *2014 IT Spending Intentions Survey*, February 2014.

Save 21% in Costs and Time Using Preconfigured Big Data Appliances

Based on ESG validation of an Oracle model for a medium-sized Hadoop-oriented big data project, a "buy" infrastructure option like Oracle Big Data Appliance will yield approximately 21% lower costs than a "build" equivalent do-it-yourself infrastructure. And using a preconfigured appliance such as Oracle's will greatly reduce the complexity of engaging staff from many IT disciplines to do extensive platform evaluation, testing, development, and integration. For most enterprises planning to take big data beyond experimentation and proof-of-concept, ESG suggests exploring the idea of skipping in-house development, on-going management, and expansion of your own big data infrastructure, to instead look to purpose-built infrastructure solutions such as Oracle's Big Data Appliance.

Delivering Big Data's Value Proposition

Introduction: In Search of Big Data Reality

The media bandies about the term "big data" as if it were a fait accompli at most enterprises. To the contrary, many enterprises have only recently begun what will turn into a multi-year commitment and effort toward enhancing the state of data analytics. Similarly, much of the source of the interest in big data springs from the vendors offering products and services based on the continually expanding Apache Hadoop project, which legitimately catalyzed interest in big data, yet their marketing sometimes contributes to the perhaps misleading notion that realizing the promise of big data will come easily and inexpensively. For example, in terms of infrastructure for Hadoop, the predominant delivery model suggests that on-premises or cloud-based commodity servers and storage ("good enough" solutions) will suffice for all environments, while the reality is that there are several viable options.

What will be the reality of big data? Assuming many enterprises will invest in new solutions to attain their respective big data goals, will Hadoop become the universal standard and revolutionize how IT departments and their partners in business deliver big data solutions, or will other approaches also have their place? ESG believes that Hadoop will play at least a part in many big data solutions. Assuming Hadoop plays some role in most organizations, what related infrastructure decisions will deliver the best ROI for big data solutions —do-it-yourself commodity infrastructures, or more purpose-designed, integrated alternatives?

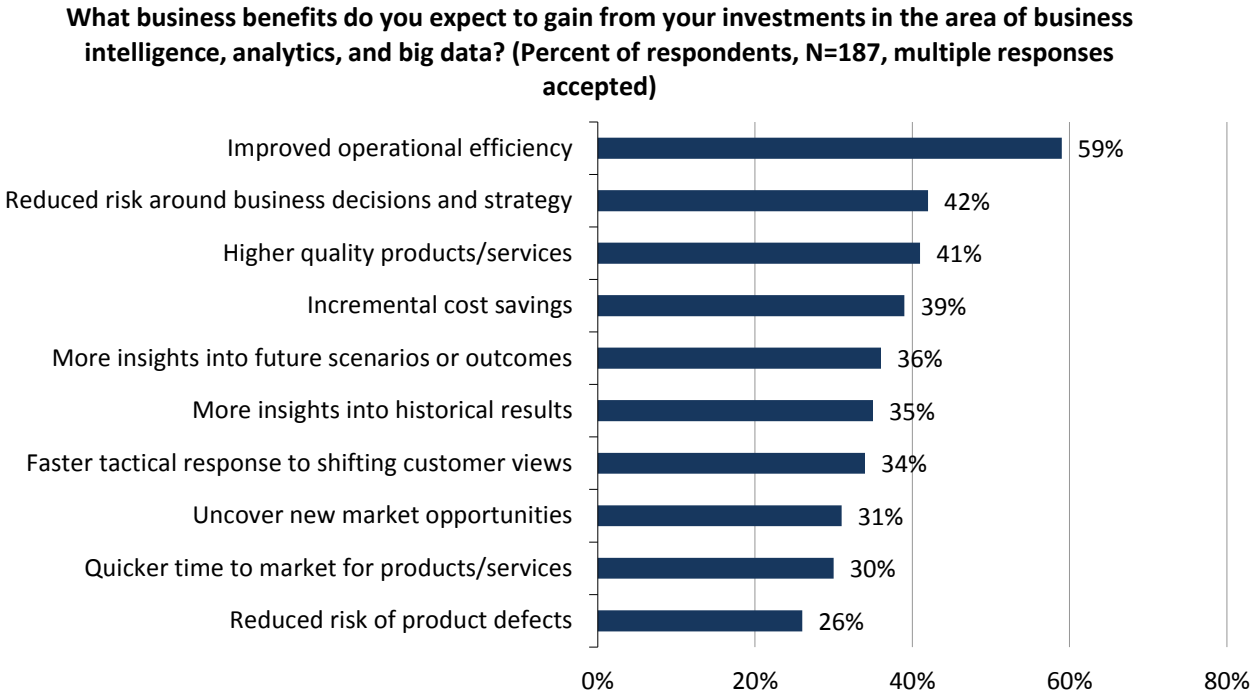
The Enterprise Big Data Value Imperative

Viewpoints abound about the importance of big data analytics. ESG holds the strongly optimistic point of view that in the long term, big data will propel significant business value and innovation across almost all industries. And we don't believe we are alone, based on recent ESG research where a wide variety of expected benefits were cited (see Figure 1) by respondents looking at their IT spending priorities in 2014.[5] ESG found it particularly telltale that roughly a quarter of respondents ranked enhancing analytics as a top business priority.

ESG holds the strongly optimistic point of view that in the long term, big data will propel significant business value and innovation across almost all industries.

[5] Ibid.

Figure 1. *Expected Business Benefits from Data Analytics Investments*



Source: Enterprise Strategy Group, 2014.

Big data being of primary importance to organizations might seem like taking a bold position, but it may prove to be an understatement; while many IT initiatives can help organizations do things better, big data helps organizations know what to do better. It is one thing to know that bookings decreased when compared with the same quarter last year (an example of business intelligence), but it is better to know why they decreased (analytics), and it is best to model and predict how they might increase going forward (big data analytics).

Despite ESG's belief that big data will eventually deliver on all the promise and hype, ESG does not believe organizations will successful reach their big data goals by cutting corners. In fact, enterprises should avoid falling for the false promise of "something earth-shattering for nearly nothing" that seems associated with big data—why not get real about big data? In order for enterprises to get real, they will need to select the right big data infrastructure.

Hadoop's Role in an Enterprise Big Data Platform

Based on the aforementioned ESG research survey, many respondents are already using—22% in total—MapReduce framework technologies to process large and diverse sets of data for big data analytics.[6] The fact that IT vendors of all types, from start-ups to well-established vendors, have jumped on the Hadoop bandwagon has changed Hadoop's status at most enterprises from a mere curiosity a few years ago to a serious choice for a key element of the big data solution set going forward.

The Hadoop MapReduce technology, while a key feature of Hadoop, only tells part of the Hadoop story because if one looks at all of the elements that make up Hadoop, it looks more like a big data platform. What Hadoop has most obviously lacked is the equivalent of integrated systems management software, but vendors like Cloudera, Hortonworks, and MapR have developed such software and support via their respective Hadoop distributions.

Enterprises stand at different points in terms of enhancing their data analytics. For example, companies that have successfully implemented data warehouses, which already do a good job at data governance, and that actively use business intelligence and analytics solutions from before the "big data" era definitely have a leg up. Those

[6] Ibid.

companies will determine how best to leverage and integrate Hadoop into their wider analytics solutions set. Other companies may choose to lean more directly on Hadoop as their primary data platform, moving to build what is now being called a data lake or data hub.

Debunking Three Common Assumptions of Hadoop Big Data

Hadoop will play a role in the big data efforts of many, perhaps a majority of, larger organizations. However, some myths are associated with big data and Hadoop. ESG believes that these myths present some risks for enterprises. Making a serious investment in big data is not to be entered into lightly, given the potential strategic impact. What are the most important big data and Hadoop assumptions that enterprises should take care to understand beyond the superficial level?

Assumption 1: Big data is not mission-critical.

If one believes that big data will indeed rank as one of the most critical capabilities to serve the business, then how should IT view big data from an architectural and operational perspective? Simple: Consider big data mission-critical, just like ERP. That means not considering big data experimental or a series of projects, but rather an enterprise-class IT asset that will deliver essential decision-making insight and direction every business day.

Organizations will initially play with big data, primarily as a learning experience. But the CIO and the business leaders should plan to architect, build, and operate a big data facility that delivers dependably, yet with flexibility for the entire organizational value chain—including customers and partners—on an ongoing basis. CIOs already know the mission-critical drill: 24x7x365 reliability, five 9s availability, appropriate performance, scalability, security, serviceability, and quick adaptability to business requirements changes.

ESG understands that thinking about big data in a mission-critical context runs counter to much of the way big data has been positioned in the media: The idea that an enterprise-class big data solution can only run effectively on inexpensive, commodity hardware, using purely open source software, simply strikes ESG as a somewhat limited viewpoint. The amount and diversity of data; the number of points of internal and external integration; the wide variety and large quantity of users; and the expertise to ensure the analytic models and results can be trusted suggests that enterprises may want to consider a similar approach to what IT has used to deliver mission-critical solutions for the business in the past.

Assumption 2: Hadoop is free.

You can visit the Apache website for Hadoop and download all the elements you need—for free. Alternatively, you could tap into emerging Hadoop-as-a-service (HaaS) offerings that you find in a variety of public clouds for something only a little more expensive than free, at least initially. Why then isn't Hadoop truly free, or nearly so?

First, the expertise required to implement and use Hadoop is expensive, whether you bring in data scientists and Hadoop engineers at several hundreds of dollars per hour, or shift some of your top business analysts and engineers and retrain them on Hadoop, or a combination thereof. Enterprises should ask themselves, "Do I have engineers on staff deeply familiar with configuring, allocating, and managing Hadoop infrastructure?" While the distribution software from vendors like Cloudera certainly helps, enterprises need to realize that big data will not remain an experiment, but in fact become a platform that will support many projects of discovery and analytics applications—it requires an enterprise lifecycle viewpoint and requisite infrastructure.

Second, ESG does not expect any enterprise to only depend on Hadoop for big data solutions—Hadoop is just part of the larger puzzle, and connecting other data sources and tools to Hadoop carries hidden costs on the human, software, networking, and hardware fronts. Existing data warehouses, and related integration and BI solutions, will certainly count as part of the overall big data solution at most companies. In fact, over the course of time, if Hadoop sits somewhere near the middle of an enterprise's big data solution set, connecting to Hadoop for data ingest and analytics visualization purposes may require significant administrative effort.

Assumption 3: I already own the infrastructure I need for big data Hadoop or can come by it inexpensively, and am staffed to configure and manage that infrastructure.

Consider a scenario where an enterprise recently shifted to a cloud-based Microsoft Exchange implementation (away from a self-managed server farm) and the enterprise thus owns a large number of generic servers and storage that are all paid for and fully depreciated. What a bonanza for big data! Simply repurpose all those nodes for your Hadoop project. Alternatively, you can buy a rack of commodity servers with DAS storage for the data nodes at white box prices.

Unfortunately, the inexpensive infrastructure myth may not work for all companies. If you are looking at big data from an enterprise-wide, mission-critical perspective, it can make sense to use "purpose built" hardware. That is, while generic servers may be fine for smaller projects and proofs-of-concept, for large-scale-production big data solutions you may want to use enterprise-grade servers, storage, and networking specifically designed for big data.

ESG believes that one of the primary big data ROI factors involves choosing the right infrastructure. Simply looking at the cost per server of commodity hardware misses much of the expense in a full production deployment. An enterprise-class big data infrastructure requires a robust and reliable architecture, and if your company does not currently possess sufficient personnel with the skills needed to architect and implement all the hardware, network, and systems software necessary for your big data solution, then you should consider an appliance instead. Identification, evaluation, and integration of a complete technology stack carries many hidden costs, not least in many hours of human effort on the project.

The Real Costs and Benefits of Big Data Infrastructure

A Model for Hadoop Big Data Infrastructure Cost Analysis: Build Versus Buy

What would an enterprise experience in terms of costs if (a) it rolled its own Hadoop infrastructure versus (b) it used an appliance that was purpose-designed and integrated for enterprise-class big data? One of the possible myths about Hadoop has been that companies can save money by rolling out and managing their own Hadoop commodity infrastructure. Of course, big data using Hadoop isn't just about clusters, it is about infrastructure: The infrastructure includes, for example, systems management software, networking, and extra capacity for a variety of analytics processing purposes.

A very important note is that "soft costs" of labor time to evaluate, procure, test, deploy, and integrate a full stack of hardware and software is not examined here. These costs in time and money can be extreme, ranging to hundreds of hours or more, and this topic alone is one of the most compelling reasons to consider a purpose-built big data appliance. That said, many organizations look to their existing staff to manage this effort, and would heavily discount or exclude these human costs, real as they may be.

ESG conducted a review of the Oracle Big Data Appliance for a three-year total cost comparison with a closely matched commodity "build" infrastructure. While the exact pricing will of course vary by vendor and over time, this model uses the closest directly comparable equipment, including HP Proliant DL-series servers and Infiniband networking, and costs that are publically available at the time of writing. Evaluators should always use this kind of ROI model as a reference while conducting their own calculations based on their specific environmental requirements. This exercise is primarily a validation of the comparable costs and financial calculations provided by Oracle, and a similar comparison may be found at Oracle blog, Price Comparison for Big Data Appliance and Hadoop.

Though some readers will argue that lower cost servers and networking components are frequently chosen in real-world deployments, these options can be lacking in performance and reliability features and aren't as directly comparable to the equipment included in the Oracle offering. The reason that the Oracle Big Data Appliance itself works well for this comparison is (a) it would serve well as an infrastructure for a medium-sized big data project, as depicted in Table 1, and (b) the cost and infrastructural details of the "buy" option—Oracle Big Data Appliance in this case—are publicly disclosed.

In order to make such a comparison, we will need a model project, and here are the assumptions for our theoretical medium-size, enterprise-class big data project:

- ***Users***: 200 end-users: 150 business end-users, 50 data scientist/analysts.
- ***Analytics Consumption***: 1/3rd ad-hoc, 1/3rd daily, 1/3rd monthly.
- ***Servers:*** Enterprise grade with newer chipsets, backup power supply, high storage density (to ensure, for example, that Hadoop doesn't become storage-bound), plenty of cores to support parallelization with more cores in the name node, plenty of memory to handle complex queries, and columnar-based analytics.
- ***Storage***: Total raw data capacity of 864 TB, in a balanced mix of refreshes (i.e., monthly, weekly, daily, real-time); note that in the example in Table 1, the math suggests more terabytes (18 nodes * 48 TB/node = 864 TB), but one has to account for replication (3x replication in pure Hadoop for example), peaks, compression, and sharding. We will use a usability rate of 30%, which in this case, rounded up, yields about 260 terabytes.
- ***Network***: Dedicated and particularly fast bandwidth, with multiple switches to deal with contention; given the amount of replication and data movement associated with, for example MapReduce, a big data infrastructure needs to ensure it doesn't become network-bound.
- ***Queries***: A set that spans from simple select statements to complex joins.
- ***Integration and Information Management***: Five new points of integration/transformation; in the process of design, additional data sources will be added in addition to, for example, a data warehouse. Those sources will require integration/transformation and information management work, and related licenses and hardware.
- ***Project Time***: Six months total running, which includes one month for architecture, design, procurement; one month for hardware/network configuration/implementation; two months for various development elements, MapReduce queries (assuming the use of Hive), statistics, user experience, integration, training, etc.; one month for final integrated/system testing and go live; one month for slack and over-runs.

For a full listing of all the resulting elements of cost associated with the theoretical project, see Appendix, Table 5.

Specific Costs for Build versus Buy Comparison

Table 1 lists those project items where ESG believes there is a pricing choice between build and buy. The table reflects estimated pricing for the "build" consumption option only.

Table 1. Medium Big Data Project Three-year TCO – Summary of Buy Cost Items

Item	Cost	Notes
Build Versus Buy Elements (Using Build Pricing)		
Servers	$410,500	18 @ $22.8k each; enterprise class with dual power supplies, 48TB of serial-attached SCSI (SAS) storage, 48-64 gigabytes memory, 1 rack
Networking	$40,000	3 @ estimated $6k for InfiniBand, 1 @ $11k for admin switch, 18 @ $0.6k for cables, looms, patch panels, etc.
Hardware support (three years)	$67,600	@15% of list cost
Hadoop licensing	$388,800	Cloudera: 18 nodes @ estimated $7.2k each per year
Installation	$14,000	Licenses and dedicated hardware
Build Project Costs	***$920,900***	Those project items where a "buy" option exists

Oracle Big Data Appliance as an Alternative to "Build"

ESG believes that for a true enterprise-class big data project of medium complexity, the appliance option will deliver nearly 20% cost savings compared with IT architecting, designing, procuring, configuring, and implementing its own big data infrastructure.

While it was convenient to use the Oracle Big Data Appliance as a "buy" candidate in this analysis, it is a well-conceived big data infrastructure solution that will serve most enterprises well for both initial big data projects and as organizations grow their big data facility. Reasons for this include:

Software: The appliance includes software often required in big data projects, such as Cloudera's Distribution including Apache Hadoop (CDH) and Cloudera Manager for infrastructure administration; Oracle Linux, the Oracle JDK, the Oracle NoSQL Database (Community Edition) to support advanced analytic queries; and the popular open source R statistical development tool. You will need to supply your own visualization tool(s), however. At time of writing, the latest version of CDH 5.0 is included, with additional features such as those from the rebase on Hadoop 2.2GA, MapReduce 2.0 on YARN, with backward compatibility to MapReduce version 1, and new functionality in Impala, Search, Spark, Accumulo, and more. Although beyond the scope of this particular paper, Cloudera's full capabilities are relatively broad and deep as a Hadoop distribution, and should be explored and considered as a preferred option for the data platform.

Additionally, security software functionality, including both network and disk encryption, Kerberos-based authentication, and more, is pre-installed and available as "check box" features, which will make protecting all that sensitive data more easily accomplished.

Hardware: Oracle's full rack Big Data Appliance includes 864 TB of raw storage capacity, 288 cores, 64 GB memory per server, and InfiniBand networking between nodes and racks—and adding racks is noninvasive. ESG believes that optimized networking throughout a big data infrastructure can be the secret sauce to big data performance: The InfiniBand included in Oracle Big Data Appliance plus the storage and core design of the appliance will enable enterprises to eliminate network bottlenecks for a Hadoop cluster. Again, much more information on the full specifications of the Big Data Appliance is readily available on Oracle's website, but the quality and capacity of the total rack is admirable and will add value over a "cheaper" commodity hardware-sourced environment.

Services: With an Oracle Big Data Appliance, you receive Oracle Premier support; the Automated Service Request (ASR) feature that auto-detects major problems and maintains a "heartbeat" with Oracle support; configuration and installation support; a warranty; and straightforward options and techniques for expansion—or you can do it all yourself with a home-built "commodity" infrastructure. It is worth noting that the appliance is supported as a whole, reducing the risk and finger-pointing of having multiple vendors involved in troubleshooting inevitable problems of a complex multi-component system.

And finally, and this is particularly important for primarily Oracle shops, using the same InfiniBand, you can connect the Oracle Big Data appliance to other Oracle engineered systems, such as the Oracle Exadata Database Machine and the Oracle Exalytics In-Memory Machine. But whether your organization is purely an Oracle shop or not, the Oracle Big Data Appliance presents a compelling alternative to do-it-yourself, Hadoop-oriented, big data infrastructures for those companies that are serious about big data.

Looking for Big Data Savings? The Infrastructure Buy Option

Where can you save in a "buy versus build" scenario for big data? One big bucket of cost comes from the big data infrastructure. Of that three-year TCO of $920,900 in our theoretical build versus buy project item comparison, using build pricing, well over half of the costs come from hardware and networking plus related support. The other large portion of costs come from software licensing. It is very important to note that the labor times to order, assemble, and manage all the hardware and software were excluded, but will surely be very significant as well. The primary "buy" option for big data infrastructure considered, being Oracle's Big Data Appliance, would deliver everything in "build" for approximately $728,150 list for three-year TCO, fully loaded (includes the Hadoop distribution, storage, network/bandwidth, hardware support, etc.).

Table 2. Buy Three-year TCO for the Oracle Big Data Appliance (full rack)

Item	Cost	Notes
Appliance (18 node)	*$525,000*	Oracle's list price
Hardware and Software Support	*$189,000*	$63k/year for 3 years of full support
Installation	*$14,150*	By Oracle professional services
Total	***$728,150***	

Source: Oracle, 2014.

ESG believes that for a true enterprise-class big data project of medium complexity, the appliance option could deliver roughly 21% cost savings versus IT architecting, designing, procuring, configuring, and implementing its own big data infrastructure.

Table 3. Medium Big Data Implementation Three-year TCOs – Buy Wins

Item	Cost	Notes
Buy Total	*$728,150*	Cost of Oracle Big Data Appliance including installation and three years of support
Build Total	*$920,900*	See Appendix for inventory
Buy Savings	*$192,750*	Over three-year period
ESG Estimated Savings	***~21%***	Oracle BDA appliance lowers costs versus do-it-yourself

Source: Enterprise Strategy Group, 2014.

But the "build versus buy" big data story doesn't end with a single project: there are additional and longer-term infrastructure costs and concepts to consider.

Big Data "Build" Proof-of-concept and Cluster Costs Misleading

Some would argue that for proof-of-concept projects, a major investment in a big data appliance, like Oracle's Big Data Appliance, at a list price of $525,000 for the full rack of 18 servers is simply too expensive, but a smaller "starter" option is available for a $186,406 list price. This entry point may be easier for many organizations, and with six nodes supporting the same software environment, could be a very desirable "scale on demand" approach. The ultimate costs of expanding from six to 18 nodes would exceed purchasing the fully equipped system at the start, but may help avoid initial overprovisioning for smaller environments. While ESG understands the reticence to make such any such capital investment, the problem with the "keep it cheap, it is only a proof-of-concept" argument rests with the nature of proof-of-concept: The proof-of-concept for a big data infrastructure must include all the elements of complexity, or it may not represent the true scenario upon enterprise-wide rollout. If IT personnel spin up a commodity cluster, implement Hadoop, load some data into Hadoop, and run a few queries, the primary thing that has been proved is that Hadoop works. Such a proof-of-concept does not necessarily indicate the infrastructure's and the related personnel's ability to reliably support and process enterprise-grade big data analytics over the long-term. In fact, making long-term big data architectural decisions based on a simplistic proof-of-concept could engender unforeseen long-term costs.

Oracle's Big Data Appliance offers a short cut to many of the integration efforts, with a full software stack including Cloudera's full Enterprise Data Hub edition, with all the latest innovations and updates. The Cloudera distribution is presented as an included perpetual license, not a per-server, ongoing subscription, and this alone has a significant impact on both cost and effort. The functionality of the data platform will be an asset because Cloudera's EDH is recognized as a technology leader in both range and quality of features, offering flexibility, scalability, performance, security, and management tools. Security features including strong authentication, network encryption, and on-disk encryption are an added bonus, if not a necessary consideration for any enterprise. Also included and fully integrated "out-of-the-box" are Oracle's Linux, JDK, and a NoSQL database license, all of which again increase the value of the offering versus procuring and assembling these software components individually.

ESG suggests that customers use references from vendors, rather than relatively simply proof-of-concept projects, for making initial decisions on big data infrastructure.

Big Data Longer-term Infrastructure Costs Favor Consistency

Though ESG believes that for medium-sized big data projects the "buy" option can deliver significant savings over "build," companies also should consider what will happen with their next project—will you reuse the same infrastructure from the initial project, or will you create a series of big data project islands? Let's go back to lessons learned with ERP, a comparably heavy investment for many companies in years past, to consider an approach.

One of the key issues many organizations faced in the early days of ERP, and that some still face, was piecemeal ERP implementations, resulting in a variety of infrastructures, databases, schemas, and user experiences. The result was a huge dependency on integration technologies, and a never-ending "ERP upgrade" lifecycle.

The time for IT and the business is now to realize that big data projects ultimately lead to something larger and more complex, and best practices like architectural guidelines should apply to big data just as they do today for ERP and CRM.

ESG believes that the customers choosing a consistent infrastructure for multiple big data projects across different departments will enjoy compounded savings due to the elimination of the learning curve associated with managing, maintaining, and tuning the infrastructure, plus the potential for infrastructure reuse.

Big Data "Build" Risk Never Disappears

Look forward three years from now: Your organization has become a successful big data organization, where IT easily adapts to new big data demands and where your executives, business users, and extended value chain all benefit and compete better due to big data.

Will your organization easily reach that goal if every big data project is a "build" project? The IT department will constantly swim upstream to deal with new demands, constantly tuning and updating custom infrastructures. The "buy" decision, however, may enable your organization to focus more on value and visualization, versus procurement and deployment.

ESG suggests that organizations take infrastructure development and deployment variability out of the pool of risk associated with reaching big data success. The risk of constantly reinventing and managing do-it-yourself infrastructures for big data may grow with the volume and complexity of big data deliverables.

Better Time to Market Is the Repeatable Benefit: The Benefit of the Buy Option

> ESG believes that a "buy" versus "buy" approach will yield roughly one-third faster time-to-market benefit associated with big data analytics projects for discovery, improved decision making, and business process improvement.

Beyond costs, determining the "benefits" is often the hardest part of calculating ROI for big data projects. The difficulty of predicting and detecting big data benefits comes from the fact that big data projects do not end per se, but rather turn into analytics applications, which grow, evolve, and lead to more discovery projects—in total, providing benefits in continually learning ways. Unfortunately, making this claim will not satisfy the CFO who is wondering when and from where the big data benefit will arrive. ESG research found that 67% of respondents wanted to see positive business impact within one year.[7]

It needn't be that difficult, however. Usually some kind of basic business metric will offer an acceptable framework for benefit, which may receive tuning over time. For example, "Better understanding of what products our customers like best" may link to "increased product sales" and "streamlined supply chain." Or, "More accurately

[7] Source: ESG Research Report, *Enterprise Data Analytics Trends*, May 2014.

predicting energy demand" may link to "a new premium pricing scenario for peak times" or "a well-tuned electrical grid bidding scheme." An agreed-upon metric for the benefits side of the equation forms the basis for ROI. But one variable seems to fit neatly into every single big data project—and that variable is time.

Quite simply, the faster the business and IT is able to deliver the big data solution, the faster the business will realize any associated benefit. The tried and true time-to-market variable applies in every big data project. ESG performed a time-to-market benefit analysis associated with a medium-sized big data project (see Appendix, Table 5 for details), and the analysis unveiled that a "buy" infrastructure will deliver the project about 30% faster, cutting the medium-sized project time from 24 weeks to 17 weeks.

ESG believes that a "buy" versus "buy" approach will yield roughly one-third faster time-to-market benefit associated with big data analytics projects for discovery, improved decision making, and business process improvement.

The Bigger Truth

Oracle's Big Data Appliance has many advantages over a home-grown solution, and for those considering their deployment options for Hadoop, it is well worth taking the time to evaluate the total picture. Certainly many organizations will be more comfortable working within their existing Oracle relationship and enjoy the confidence of quality services and support that will accompany the fully vetted and pre-integrated appliance. Cost is outlined as one factor, but ultimately the satisfaction of the many users in the enterprise will determine the success of the big data initiative.

Serious Big Data Requires Serious Commitment

In order for big data to deliver on the promise of helping organizations look forward, just as ERP for example has helped automate business processes and business intelligence has helped organizations look backward, big data will require an enterprise-class facility. Big data carries unique requirements for hardware, networking, software, and human skills. Enterprises should plan to build mission-critical infrastructures to support a series of big data projects and applications, yielding a big data facility that continually benefits the organization over time.

Avoid Common Assumptions about Big Data

Hadoop has been associated with sometimes over-hyped promises, like "you already have the data you need," "you already have the people you need," and "you can use inexpensive commodity infrastructures," which may not be true for many organizations. Hadoop is a tool that can be used in many ways to help organizations achieve big data results, but meeting end-user expectations of performance and availability often requires costly expertise and an enterprise-class infrastructure that spans the total needs for storage, processing, and lifecycle management.

"Buy" Will Often Trump "Build" for Both Big Data Infrastructure Costs and Benefits

ESG asserts that a shorter path to lowering big data costs for the vast majority of enterprises can involve buying a preconfigured and purpose-built infrastructure—such as an appliance. "Rolling your own" infrastructure often involves a long series of hidden risks and costs, and may undermine your organization's ability to deliver big data in the long term. In addition, the "buy" option for big data infrastructures compresses project durations. Thus, well-designed appliances often helps deliver the one common benefit metric for big data: time to market.

Appendix

Table 4. Elements of a Medium Enterprise Big Data Implementation Cost/Benefit – Build Cost Elements

Item	Value	Metrics/Comments
Hardware/Network		
Nodes	18	Servers - each 2 x 8-core processors
Cores	16	16 per node, as above
Memory	64	GB/server (with option for expansion)
Racks	1	Assumes up to 18 nodes/rack
Storage for Nodes	48	TB/node for primary clusters, server-based storage
Storage for Information Management	50	Terabytes; assumes one-fourth of total data in motion maximum at any given time
Switches	3	Assumes Infiniband, assumes 3x throughput improvement over generic 10GBe; plus an administration switch and assorted cabling
Hardware Support	15%	Of total hardware costs; third-party support
Software		
Hadoop License Costs	18 @ $7,200/node	Typically new licenses during early big data adoption, allow for two more licenses for warm backup

Table 5. Medium Big Data Project – Build Versus Buy Project Time Analysis

Phase	Build Time	Buy Time/Savings (assumes 4 week months)
		Reason for Reduction
Architecture, design, procurement	1 month	2 weeks/2 week savings
		Hardware/network pre-designed, single source procurement
Hardware, network configuration, and implementation	1 month	2 weeks/2 week savings)
		Appliance pre-configured; less “pieces” to implement separately and certify
Development, integration, training, etc.	2 months	7 weeks/1 week savings
		Appliance stability provides more solid foundation for development, testing, etc.
Integrated system test, go live	1 month	3 weeks/1 week savings
		Expect smooth system test and go live due to appliance’s fewer “moving parts”
Slack and over-run	1 month	3 weeks/1 week savings
		More dependable infrastructure will lead to more dependable scheduling
Project Totals	***6 months (24 weeks)***	***4.25 months (17 weeks)***
“Buy” Time to Market Benefit = 7 weeks or ~30%		

What Table 5 suggests is that a medium-sized project will complete nearly one-third sooner due to a "buy" infrastructure decision over "build.” The reduced procurement time, reduced configuration and design time,

reduced implementation time, and a more dependable infrastructure out of the box, which moves along development and testing time, are all factors that contribute to the idea of "buy" speeding along big data projects. The time-to-market benefit of "buy" keeps giving, too, because the same time savings keeps repeating in every big data project.

Enterprise Strategy Group | **Getting to the bigger truth.**

20 Asylum Street | Milford, MA 01757 | Tel: 508.482.0188 Fax: 508.482.0218 | www.esg-global.com