White Paper

Getting Real About Big Data: Build Versus Buy

By Evan Quinn, Senior Principal Analyst

February 2013

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

Contents

Executive Summary 3
In Search of Big Data Infrastructure 3
Do-it-yourself Hadoop Usually the Wrong Choice 3
Save One-third in Costs and Time Using the Preconfigured Oracle Big Data Appliance 3
Delivering Big Data's Value Proposition 4
Introduction: In Search of Big Data Reality 4
The Enterprise Big Data Value Imperative 4
Hadoop's Role in an Enterprise Big Data Platform 5
Debunking Three Myths of Hadoop Big Data 5
The Real Costs and Benefits of Big Data Infrastructure 7
A Model for Hadoop Big Data Infrastructure Cost Analysis – Build Versus Buy 7
Specific Costs for Build versus Buy Comparison 7
Looking for Big Data Savings? The Infrastructure Buy Option 8
Better Time to Market Is the Repeatable Benefit: The Benefit of the Buy Option 9
The Bigger Truth 10
Serious Big Data Requires Serious Commitment 10
Avoid the Traps of Hadoop Hype 10
"Buy" Trumps "Build" for Both Big Data Infrastructure Costs and Benefits 10
Oracle Big Data Appliance as an Alternative to "Build" 11
Appendix 12

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, more rapidly discover insights to help improve products and services, and increase profits. In a recent ESG survey, two-thirds of IT and business professionals responsible for their organization's data analytics strategies, technologies, and processes considered enhancing analytics a top-five business priority, and more than one-quarter cited enhancing analytics as their most important business priority.[1] Given the high degree of interest in analytics and the vast quantity of unmined data, the first step most organizations need to take on the path towards realizing the benefits from 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 estimates that by the end of 2013, more than half of larger organizations will have made such investments. ESG believes that buying the preconfigured Oracle Big Data Appliance will prove to be the better option for most larger organizations trying to successfully implement a big data infrastructure over a "build your own" solution.

Do-it-yourself Hadoop Usually the Wrong Choice

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 established IT vendors.

For many organizations, Hadoop-based solutions will represent the first new explicit big data investment. However, the number of successful Hadoop implementations put into full production using do-it-yourself infrastructure is rare. 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, essentially all organizations also quickly find that they lack the hard-to-find and expensive expertise to make do-it-yourself Hadoop infrastructure work for big data.

The 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 Hadoop, and evolve and manage the Hadoop 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 does not offer a low-cost ride to big data analytics. ESG asserts that most organizations implementing Hadoop infrastructures based on human expertise plus commodity hardware will experience unnecessarily high costs, slower speed to market, and unplanned complexity throughout the lifecycle.

Save One-third in Costs and Time Using the Preconfigured Oracle Big Data Appliance

Based on ESG's modeling of a medium-sized Hadoop-oriented big data project, the preconfigured Oracle Big Data Appliance is 39% less costly than a "build" equivalent do-it-yourself infrastructure. And using Oracle Big Data Appliance will cut the project length by about one-third. For most enterprises planning to take big data beyond experimentation and proof-of-concept, ESG suggests skipping the idea of 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 Big Data Appliance.

[1] Source: ESG Research Report, *The Convergence of Big Data Processing and Integrated Infrastructure*, July 2012.

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, most enterprises have just begun what will turn into a multi-year commitment and effort towards enhancing the state of data analytics. Similarly, much of the source of the hype associated with big data springs from the Apache Hadoop project, which has legitimately catalyzed interest in big data, but also 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 Hadoop hype suggests that commodity servers and storage ("good enough" solutions) will suffice.

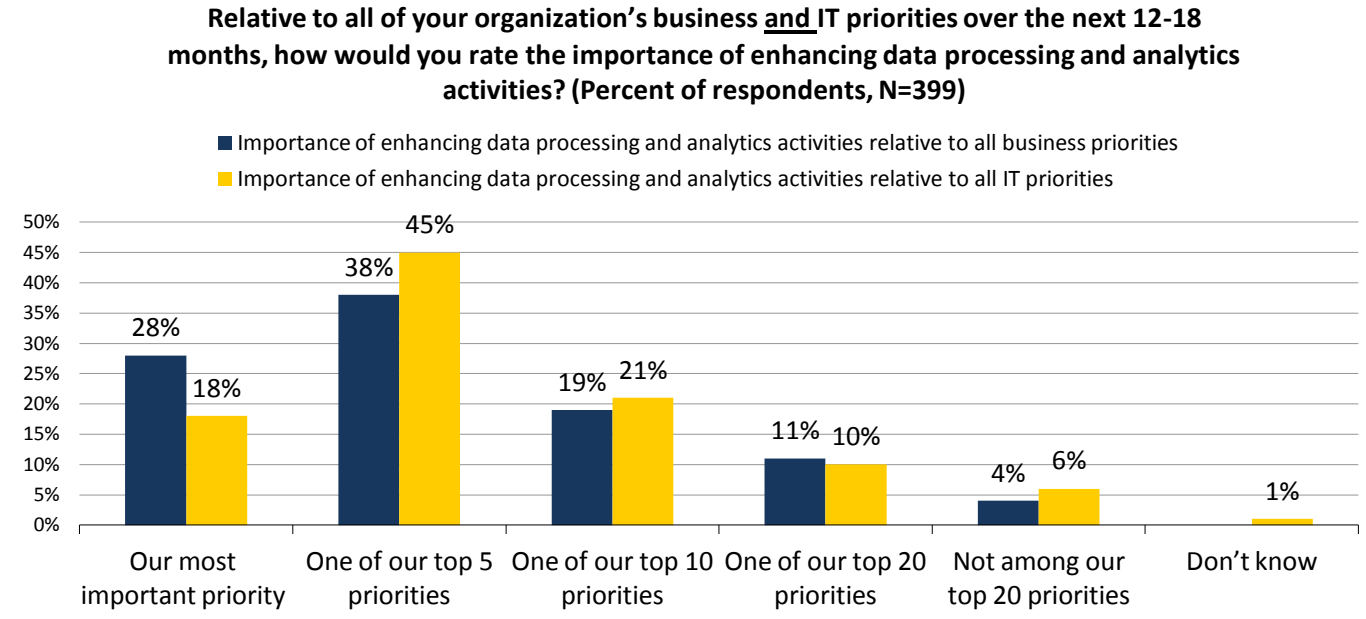
What is, or for more organizations will be, the reality of big data? Assuming many enterprises will invest in new solutions to attain their respective big data goals, will Hadoop revolutionize how IT departments and their partners in business deliver big data solutions, or will Hadoop play a less starring role? 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 Hadoop-infused big data solutions–do-it-yourself commodity infrastructures, or more purpose-designed, integrated alternatives?

The Enterprise Big Data Value Imperative

> ESG believes that in the long term, big data will roughly equate to ERP in terms of value to enterprises.

Viewpoints abound about the importance of big data analytics. ESG believes that in the long term, big data will roughly equate to ERP in terms of value to enterprises. We don't believe we are alone, based on recent ESG research where nearly two-thirds (see Figure 1) of respondents ranked enhancing analytics as either their most important, or a top-five IT and business priority over the coming year to year and a half[2]. ESG found it particularly telltale that over a quarter of respondents ranked enhancing analytics as the number one business priority.

Figure 1. *Importance of Enhancing Data Processing and Analytics Activities*



Source: Enterprise Strategy Group, 2013.

[2] Source: Ibid.

Stating that big data will eventually equal ERP in terms of importance to organizations might seem like taking a bold position, but it may prove to be an understatement; while ERP helps organizations do things better, big data helps organizations know what to do better. It is one thing to know that bookings decreased when compared to 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 reach big data 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 a big data infrastructure.

Hadoop's Role in an Enterprise Big Data Platform

Based on the aforementioned ESG research survey, the majority of respondents have either used or have interest in using—62% in total—MapReduce framework technologies to process large and diverse sets of data for big data analytics.[3] The market-leading source of MapReduce technology is Hadoop, the Apache open source project responsible to a large degree for the rising consciousness about big data.

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. ESG estimates that by the end of 2013, roughly half of the Global 2000 will have at least experimented with Hadoop.

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."

Different enterprises stand at different points in terms of enhancing their data analytics. For example, companies that have successfully implemented data warehouses, that 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 companies will determine how best to leverage Hadoop and how best to integrate Hadoop into their wider analytics solutions set. Other companies may choose to lean more directly on Hadoop.

Debunking Three Myths of Hadoop Big Data

Hadoop will play a role in the big data efforts of many, perhaps a majority, of larger organizations. However, there are some myths associated with big data and Hadoop. ESG believes that these myths present significant 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 myths that enterprises should take care to understand beyond the superficial hype?

Myth 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 certainly 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

[3] Source: Ibid.

already know the mission-critical drill: 24x7x365 reliability, five 9s availability, appropriate performance, scalability, security, and serviceability, and the ability to quickly adapt 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 will run effectively on inexpensive, commodity hardware, using purely open source software, simply strikes ESG as naïve. 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 need a similar approach to what IT has used to deliver mission-critical solutions for the business in the past.

Myth 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 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 help, enterprises need to realize that big data is not an experiment, but in fact 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 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 as much attention as managing Hadoop.

Myth 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 bunch of skinny boxes with server-based storage for the data nodes, and one beefier box for the name node, at white box prices. Pretty close to free again?

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

ESG believes that one of the two primary big data ROI factors involves choosing the right infrastructure. If you are a multi-business-unit organization with data analysts and users all over the world, speaking different languages, and constantly churning through new analytics scenarios, a farm of inexpensive nodes all servicing all of the elements required for big data—from load through visualization—is a recipe for ROI disaster. An enterprise-class big data infrastructure requires a unique architecture, and if your company possesses the personnel with the skills needed to architect and implement all the hardware, network, and systems software needed for your big data solution, then your company is in a small minority.

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 it (a) rolled its own Hadoop infrastructure versus (b) using 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.

ESG decided to focus on using Oracle Big Data Appliance for cost comparison purposes versus commodity "build" infrastructure. The reason that Oracle Big Data Appliance works well for this comparison is (a) ESG believes it would serve well as an infrastructure for a medium-sized big data project, as depicted below, 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***: Two hundred TB of data, in a balanced mix of refreshes (i.e., monthly, weekly, daily, real-time); note that in the example below, the math suggests more terabytes (18 nodes * 36 TB/node = 648 TB). But one has to account for replication (3x replication in pure Hadoop for example), peaks, compression, and sharding. ESG will use a usability rate of 30%, which in this case rounded up yields 200 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, map reduce, queries (assuming the use of Hive), statistical, 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 I. Some project items do not avail themselves of a build or buy choice, and those items are referred to as "always built" herein. An inventory of "always built" items and costs are in Appendix II.

Specific Costs for Build versus Buy Comparison

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

*Table 1. Medium Big Data Implementation – Summary of "Build Versus Buy" Cost Items (Priced for Build)**

Item	Cost	Notes
Build Versus Buy Elements (Using Build Pricing)		
Servers	$400,000	@ $22k each; enterprise class with dual power supplies, 36TB of serial attached SCSI (SAS) storage, 48-64 gigabytes memory, 1 rack
Server support	$60,000	@ 15% of server cost
Switches	$15,000	3 @ $5k for InfiniBand; in older network switches (e.g., 10 GbE) will run at least 3x the costs of InfiniBand**
Distribution/systems management software	$90,000	Cloudera: 18 nodes @ $5k each
Integration	$100,000	Licenses and dedicated hardware
Information Management Tools	$20,000	4 @ $5k each
Architect, design, procure	$32,000	320 hours @ $100/hour human cost
Node Configuration and Implementation	$16,000	8 hours/node, 20 nodes = 160 hours, $100/hour
Build Project Costs	***$733,000***	Those project items where a "buy" option exists
*Does not include "always built" items, see Appendix II for an "always built" project inventory and costs		
**Assumes in-house expertise for InfiniBand exists, though most IT departments do not have InfiniBand experts		

Source: Enterprise Strategy Group, 2013.

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 total $733,000 in our theoretical build versus buy project item comparison, using build pricing, nearly two-thirds of the costs come from hardware and networking plus related support. The other roughly one-third of costs come from software and human projects costs. ESG estimates that Oracle's "buy" offering in Big Data , Oracle Big Data Appliance, would deliver everything in "build" for approximately $450,000 list, fully loaded (includes the Hadoop distribution, storage, network/bandwidth, hardware support, etc.). ESG believes that for a true enterprise-class big data project of medium complexity, Oracle Big Data Appliance will deliver nearly 40% cost savings versus IT architecting, designing, procuring, configuring, and implementing its own big data infrastructure.

ESG believes that for a true enterprise-class big data project of medium complexity, Oracle Big Data Appliance will deliver nearly 40% cost savings versus IT architecting, designing, procuring, configuring, and implementing its own big data infrastructure.

Table 2. Medium Big Data Project - Oracle Big Data Appliance (Buy) Wins

Item	Cost	Notes
Build Total	*$733,000*	See Table 1 for inventory
Buy (Oracle Big Data Appliance)	*$450,000*	Cost of Oracle Big Data Appliance for same infrastructure and tasks costs (list)
Buy (Oracle Big Data Appliance) Savings	*$283,000*	Not lifecycle costs, just for initial project
ESG Estimated Savings	***~39%***	Oracle Big Data Appliance lowers costs versus do-it-yourself

Source: Enterprise Strategy Group, 2013.

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 fully formed big data appliance, like Oracle's Big Data Appliance, at a list price of $450,000 is simply too expensive. While ESG understands the reticence to make such a 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 proves nothing. If IT personnel spin up a commodity cluster, implement Hadoop, load some data into Hadoop, and run a few queries, the only thing that has been proved is that Hadoop works. Such a proof-of-concept does not prove anything about the infrastructure's and the related personnel's ability to 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 massive long-term costs.

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 "Buy"

Though ESG believes that for medium-sized big data projects the "buy" option will deliver significant savings over "build," companies need to 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 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 and standard vendor lists should apply to big data just as they do today for ERP and CRM.

ESG believes that the customers choosing the same "buy" over "build" infrastructure for multiple big data projects 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 ever 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, will enable your organization to focus on value and visualization, versus procurement, cores, and terabytes.

ESG suggests that most organizations take infrastructure 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 grows with the volume and complexity of big data deliverables.

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

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, that grow, evolve, and lead to more discovery projects—in total,

providing benefits in entirely unforeseen ways. Unfortunately, making this claim will not satisfy the CFO who is wondering when and from where the big data benefit will arrive.

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 what products our customers like best" may link to "increased product sales" and "streamlined supply chain." Or, "More accurately 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 the medium-sized big data project (see Appendix III for details), and the analysis unveiled that a "buy" infrastructure will deliver the project about 30% faster, cutting the medium-size project time from 24 weeks to 17 weeks.

ESG believes that a "buy" versus "do-it-yourself" 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.

ESG believes that a "buy" versus "do-it-yourself" 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

Serious Big Data Requires Serious Commitment

In order for big data to deliver on the promise of helping organizations look forward, just as ERP has helped automate business processes and business intelligence has helped organizations look backwards, big data will require an ERP-class facility. Big data, like ERP, 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 the Traps of Hadoop Hype

Hadoop has been associated with some 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 do not hold water even under a cursory examination. Hadoop is a tool that can be used in many ways to help organizations achieve big data results, but it requires costly expertise and an enterprise-class infrastructure that spans the total needs for storage, processing, and lifecycle management.

"Buy" Trumps "Build" for Both Big Data Infrastructure Costs and Benefits

ESG asserts that the shortest path to lowering big data costs for the vast majority of enterprises will involve buying a preconfigured and purpose-built infrastructure—usually an appliance. "Rolling your own" infrastructure 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, "buy" helps deliver the one common benefit metric for big data: time to market.

Oracle Big Data Appliance as an Alternative to "Build"

While it was convenient to use 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; the Oracle NoSQL Database 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.

Hardware: Oracle Big Data Appliance includes 648 TB of storage, 288 cores, and InfiniBand networking between nodes and racks—and adding racks is noninvasive. ESG believes that optimized networking throughout a big data infrastructure is often 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 fly through MapReduce.

Services: With an Oracle Big Data Appliance you receive Oracle Premier support; the Automated Service Request (ASR) feature which 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.

And finally, and this is particularly important for primarily Oracle shops, using the same InfiniBand, you can connect 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, Oracle Big Data Appliance presents a compelling alternative to do-it-yourself, Hadoop-oriented, big data infrastructures for those companies serious about big data.

Appendix

Table 3. 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	6 or 12	12 for "name node," 6 for slaves
Memory	64	GB/server (with option for expansion)
Racks	1	Assumes up to 18 nodes/rack
Storage for Nodes	36	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; may also require Ethernet administration switch - assumes internal expertise for InfiniBand
Hardware Support	15%	Of total hardware costs; third-party support
Software		
Distribution/Systems Management	20 @ $5,000/node	Typically new licenses during early big data adoption, allow for 2 more licenses for warm backup
Integration	$100,000	Incremental enterprise software licenses plus dedicated hardware – to support point integration, transformation
Information Management Tools	4 @ $5,000/user	Software licenses; assume 4 incremental @ $5k each – assume existing servers will support new licenses
Statistical/analytics tools	$0	Assumes use of open source "R"
Visualization/collaboration software	$5k/analyst $1k/business user	Assumes 100 business users, 50 analysts
Human		
Architect, Design, Procure	4 weeks, 2 experts	Assumes experts have previous experience in big data
Node Configuration and Implementation	2 days/node	Includes networking – 20 nodes
Load and Information Management Design/Implementation	2 weeks, 2 experts	Assume 5 data sources, using existing information management tools, includes testing and go live
Map Reduce Coding	4 weeks, 2 experts	
Query and User Experience Development, Phase 1	8 weeks, 2 experts	Assumes 50 queries, 5 dashboards, testing, data validation, go live
R statistical coding	8 weeks, 2 experts	Assumes working in parallel with query/user experience developers throughout
End-User Training	2 trainers, 5 (30 users per class) 1-day end-user sessions; 2 (25 users per class) 2-day data scientist/analyst sessions	Assumes all sessions completed during last 2 weeks before go live

Source: Enterprise Strategy Group, 2013.

Table 4. Medium Big Data Implementation – "Build vs Oracle Big Data Appliance(Buy)" Versus "Always Built" Costs

Item	Cost	Assumptions
Always Build Costs		
Visualization/collaboration software	$350,000	$5k for 50 analysts, $1k for 100 users initial license
Storage for information management	$40,000	$8k/terabyte for tier-2 storage, need 50 tb for backup/restore/staging
Load and Information Management Design/ Implementation	$16,000	80 hours for 2 experts each = 160 hours @ $100/hour
Map Reduce Coding	$32,000	160 hours for 2 experts each = 320 hours @ $100/hour
Query and User Experience Development, Phase 1	$32,000	160 hours for 2 experts each = 320 hours @ $100/hour
R statistical coding	$32,000	160 hours for 2 experts each = 320 hours @ $100/hour
End-User Training	$4,000 $3,200 $120,000 $80,000	40 hours for 2 trainers for user classes @ $100/hour 32 hours for 2 trainers for analyst classes @ $100/hour 8*150 = 1200 hours for business users @ $100/hour 16*50 = 800 hours for scientist/analysts @ $100/hour
Always Built Sub-total	***$709,200***	

Source: Enterprise Strategy Group, 2013.

Notes on Assumptions:

1. ESG believes that its cost estimates are optimistic. There are a variety of possible scenarios, mainly using the build option, where many of the project costs and work will run well higher than what is estimated.
2. We have not included lifecycle costs of a big data facility in terms of additional development work for MapReduce, query, and other elements. Our belief is that Oracle Big Data Appliance would trump a "build" scenario deeper in the lifecycle, simply because the build option includes more moving parts to maintain.
3. In order to make the labor costs analysis simple, we assumed $100/hour, fully loaded, for all labor. We realize that MapReduce Java developers/consultants will cost significantly more than "sunk cost" internal personnel spinning up nodes, but the math proves out that, in the bigger picture, these individual variances do not significantly impact a model of this nature.

Table 5. Medium Big Data Project – Build Versus Oracle Big Data Appliance (Buy) Project Time Analysis

Phase	Build Time	Oracle Big Data Appliance (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,	2 months	7 weeks/1 week savings

training, etc.		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)***
Oracle Big Data Appliance ("Buy") Time to Market Benefit = 7 weeks or ~30%		

Source: Enterprise Strategy Group, 2013.

What Table 5 suggests is that a medium-sized project will complete nearly one-third sooner with Oracle Big Data Appliance, 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 Oracle Big Data Appliance, a "buy", speeding along big data projects. The time-to-market benefit of Oracle Big Data Appliance, a "buy" solution, keeps giving too, for 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