WHITE PAPER

Delivering Real-World Total Cost of Ownership and Operational Benefits

TREASURE**DATA**
CLOUD DATA WAREHOUSING

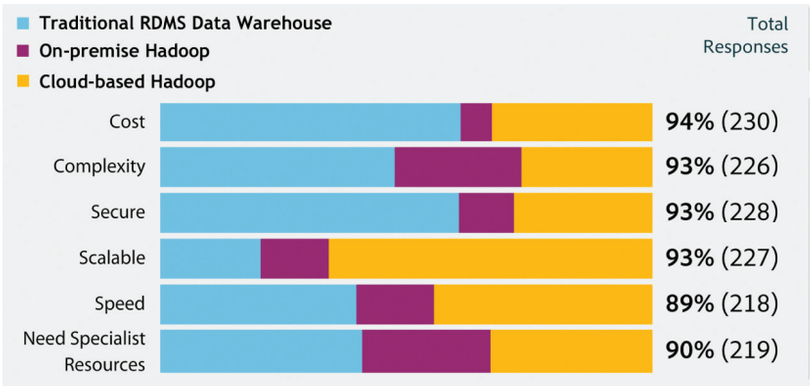
Background

Big data is traditionally thought of as a Big Ticket problem and this remains true for many big data solutions currently in the market from traditional vendors such as IBM, HP, Oracle and Teradata. To address this perception, many of these vendors have acquired appliance-based solutions and this has enabled them to reduce big data solution costs to between $25,000 and $100,000 per terabyte (Tb) of data. With this level pricing, even a modest appliance supporting 25Tb of data could easily cost $1 million+, while large (Petabyte+) configurations are affordable only by the largest organizations.

It is therefore no surprise that data-driven start-ups or new business units in brick and mortar companies have so readily embraced Hadoop with its open source "zero cost" model and efficient use of commodity hardware for storage and processing. A recent InformationWeek report[1] compared appliance vs. Hadoop costs and concluded "Running 30 Hadoop machines and an Arista switch costs less than $500,000 vs. at least $7 million for an appliance. No wonder everyone is looking out for Hadoop talent… It's a no-brainer when you plug in real numbers".

Despite Hadoop's cost advantage, it remains a complex solution that needs specialist staff for all phases of the development and operations life cycle. This has led several vendors to offer Cloud-based pre-configured Hadoop solutions in an effort to reduce the complexity inherent in Hadoop's flexible, distributed architecture.

Treasure Data recently conducted a survey of almost 250 big data specialists[2] to find out how they viewed traditional, Hadoop-based and Cloud-based big data architectures. One of the most striking things about the survey results was that all three approaches were rated as being high in terms of complexity and the need for specialist resources.



To understand these results more completely, Treasure Data conducted live research to investigate the resource needs of these approaches from initial experimentation to large-scale deployment. This whitepaper outlines their findings and contrasts an on-premise Hadoop solution with Amazon's Elastic MapReduce platform and the Treasure Data Cloud Data Warehouse service.

Executive Summary

Hadoop is a breakthrough technology that has made big data a reality for most organizations. However, with its power and flexibility come complexity and cost and a demand for expert resources that have exceeded the supply of such expertise in the market. This has made hosted Hadoop options such as Amazon EMR an attractive alternative. While Amazon addresses many of the complexity, scaling and support issues inherent in the on-premise option, Amazon's support is limited to their own components, leaving customers responsible for much of their own developer and end-user support.

Treasure Data's Cloud Data Warehouse service extends the hosted model to its logical conclusion and provides a fully managed big data solution, comprehensive SLAs and a single point of contact for all support issues. Treasure Data can do this because they actively curate the best open source components, integrate and test them in a complete big data systems environment ensuring a DevOps platform built on the latest software innovation and best practices. The economies of scale inherent in the Cloud model, combined with open source software and Treasure Data's technology, result in a highly economical offering for the fast-moving, data-driven customer.

"The time to market on providing usage metrics to our customers was shortened to a fraction of the time compared to our past experiences of building things from scratch. Also, costs are much lower than hiring a full time systems engineer to maintain the associated infrastructure." — *Michi Kono, Founder, Splurgy*

One critical impact of the pre-configured Treasure Data approach is that the deployment time, from experimentation to production, is massively reduced. Taking the DevOps estimates identified in the five phases of the deployment model in this document, the total effort associated with this approach is likely to be:

	Range of Effort (DevOps Days)	Median Number of Days	Median Cost (at $1,500/day)
On-Premise Hadoop	60 to 160	110	$165,000
Amazon EMR	20 to 40	30	$45,000
Treasure Data	5 to 10	7.5	$11,250

Whether a company assesses cost based on the staffing costs of this effort (plus the associated hardware and software costs) or the opportunity cost of getting a mission-critical big data solution into production faster, it is clear that the Cloud delivers a more-effective solution than on-premise Hadoop implementations.

Treasure Data's pre-built Cloud Data Warehousing solution extends the cost benefits of the Cloud model and further accelerates time to value. The result is a more cost-effective, scalable and faster option than either on-premise Hadoop implementation or Amazon EMR solution.

Phase One: Experimentation

When most companies decide that they need a big data solution, their starting point is usually to evaluate two or three options in an experimental or sandbox environment. The objective of this phase is usually to test different approaches and understand which is most likely to be best suited to the big data use-case. Typically, these evaluations are carried out by one person or a very small team of people, using one or two data sources on a single-server with a very simple use-case.

ON-PREMISE HADOOP

Some hardware may be required for this, but most likely spare hardware capacity will be available. Implementing the necessary software can be a complex process, as you need to set up the operating system, file system, Java and Hadoop. Scripts may need to be developed to load the data into the Hadoop back-end. This process takes 5-10 days of effort and requires one or two highly skilled engineers.

AMAZON EMR

The beauty of any Cloud solution is the on-demand provisioning of the hardware and operating system software. In the case of Amazon EMR, the platform also includes a generic Hadoop/Hive environment but this must be configured to the needs of the use-case requiring expertise with Amazon Web Services, Hive and Hadoop as well as scripting for data load. This process takes 1-2 days and requires one or two skilled engineers.

TREASURE DATA

As a Cloud solution, Treasure Data provides the benefits of immediate provisioning but also provides installers and td-agent, a lightweight data load daemon, to accelerate the configuration and data load processes. Although Treasure Data uses many Hadoop components, Treasure Data provides a preconfigured, pre-tested environment that makes it possible for engineers with no knowledge of Hadoop to create, load and test a big data environment in hours not days.

Phase Two: Proof of Concept

Once the use-case and basic architecture components have been identified, most companies will proceed to a Proof of Concept (PoC) phase before the complete big data project is approved. Due to resource constraints, most companies will select the experimentation approach that best meets their needs rather than run the PoC on multiple platforms. Typically, they will create a small (2 or 3 person) project team to develop a PoC that accesses multiple data sources, is deployed on a small cluster of servers and will support a small number of real-world use-cases.

ON-PREMISE HADOOP

A small cluster of dedicated servers needs to be provisioned or allocated to the project to prove its scaling capability and predict production needs. As well as requiring the necessary hardware and software, this process also requires significant architecture and engineering expertise to configure and test the cluster. Additionally, as the number of data sources grows, so does the effort required to develop and test data load scripts. Developing and testing the use-cases may also require a significant amount of time and effort depending on their complexity. As a result, this phase often takes a team of 3-4 specialist staff between 30 and 60 days to complete.

AMAZON EMR

Although Amazon EMR can be provisioned to support multiple servers very quickly at low cost, Hadoop still needs to be configured to take full advantage of the clustering capability. This requires knowledge of Hadoop performance engineering best practices, a skill that may not exist in many IT teams. Where this skill does not exist, it may be expensive and difficult to recruit - especially for a short-term project. In addition, the Amazon EMR option shares the data load scripting and use-case development overhead of the on-premise Hadoop option. This phase can usually be accomplished within a 10 to 20 day window.

TREASURE DATA

Treasure Data is delivered as a service and can be scaled on-demand, but unlike Amazon EMR, this scaling is achieved without the developer needing deep Hadoop knowledge as it is managed by the pre-configured Treasure Data infrastructure. As the number of data sources increases, so does the benefit of td-agent, which, once installed, provides easy data load on a batch or continuous basis. Treasure Data also accelerates and simplifies the data extraction process by providing a SQL-like interface enabling anyone with SQL knowledge to query the database. As a result, most Treasure Data customers complete their PoC within 1 to 2 days.

Phase Three: Build-Out

Once the Proof of Concept is complete and project approval has been granted, the build-out of the production infrastructure can be started. Typically, this will be a large-scale project that will need highly skilled engineering resources to be recruited, contracted or redeployed from other projects. Moreover, these engineering resources must include diverse expertise covering hardware, operating systems, databases, development and systems management tools. Even if this diversity of resources exists within an organization, coordinating their availability around the timing of the big data project may be an issue.

ON-PREMISE HADOOP

Building out an on-premise Hadoop environment requires significant hardware expense, will challenge the IT capability of most companies and will almost certainly require the hiring of additional — and costly — resources to manage the resulting infrastructure. The build-out solution will probably use components from several vendors, putting the onus for compatibility testing on the user not the vendor. Such infrastructures can be extremely difficult to troubleshoot resulting in a company's most skilled engineers being diverted from productive work to handle compatibility and configuration issues. Build-out time will vary based on project size, but could easily take 30 to 60 days.

AMAZON EMR

The Cloud model makes it much faster to build-out on Amazon EMR than on-premise, and the pre-configured nature of the EMR solution ensures compatibility of the Amazon components. However, it is still likely that there will be third party components involved in the complete solution. Additionally, because many of the data sources will not be Cloud-based, interoperability with on-premise or third party data sources must be established, tested and supported. With this comes cost and customers should be aware that transferring 10Tb of data per month from Amazon to an on-premise solution could cost $14,400 per year. The estimated build-out of an Amazon EMR solution is 10 to 20 days.

TREASURE DATA

In the case of Treasure Data's service, processing, storage and network resources are completely elastic and can be scaled as business needs dictate. Furthermore, Treasure Data can be scaled without the need to redeploy a company's best resources from other projects. Because Treasure Data owns, integrates and optimizes the entire component stack, Treasure Data delivers 2 to 3 times the performance of an EMR solution. Additionally, td-agent can pipe data between applications of any kind, reducing the cost and complexity associated with Cloud to on-premise interoperability. Build-out with Treasure Data is typically a 4 to 5 day process.

Phase Four: DevOps

By their nature, big data applications are more dynamic than their transactional counterparts. As a result of this dynamism, the traditional IT distinction between developers and operations people has become blurred. To build a solution that can respond instantly to new trends detected in its data and to provide performance that meets user expectations, developers must understand database operations and vice versa, hence the DevOps concept.

ON-PREMISE HADOOP

While the on-premise option provides DevOps teams with complete control over their environment, the complexity of the architecture often means that DevOps teams spend much more of their time on operations work than development work. Along with this architectural complexity comes cost. Although on-premise support costs are "softer" than the monthly support cost of a Cloud solution, they can be at least as much as these costs and, unlike a hosted Cloud option, on-premise support costs rarely include the SLAs and accountability that are inherent in the Cloud model.

AMAZON EMR

The DevOps model fits the Amazon EMR model very well. Amazon's world-class support capability and comprehensive management tools means that DevOps is more development focused than in the on-premise Hadoop option. However, because Amazon's SLAs only cover their supported software, customers still need an operations capability to support their own code, and tasks like schema changes can still be major issues.

TREASURE DATA

Treasure Data extends the Amazon model to its logical conclusion and provides a fully managed big data solution, comprehensive SLAs and a single point of contact for all support issues. Treasure Data is also schema independent; making schema level changes a non-issue.

Phase Five: Scaling and Hardening

Just as big data applications tend to be more dynamic than transactional applications, successful big data applications are more likely to experience explosive growth than transactional applications. For example, take the case of a Treasure Data customer in the online games business that went from zero to over 60 billion rows of data in less than a year. Big data applications therefore need to have open-ended scaling and be capable of being "hardened" through high-availability (HA) to 365x7x24 data access. This level of scaling and hardening provide unique challenges particularly to on-premise infrastructures that are often not suited to rapid scaling or enhancement to support HA.

ON-PREMISE HADOOP

Scaling an on-premise Hadoop implementation comes down to adding hardware to the cluster and tuning the systems and applications software to take best advantage of this. Similarly, adding an HA capability to an on-premise Hadoop implementation means replicating some, or all, of the hardware configuration or seriously compromising the capability that will be supported in the event of failure. Eliminating single points of failure in such an architecture is an involved and difficult process that has to be addressed by specialists who are in high-demand.

AMAZON EMR

Amazon EMR's scaling and HA capability is first-rate and, relative to the on-premise Hadoop option, can be rapidly deployed at low-cost.

TREASURE DATA

Treasure Data is deployed at Amazon and shares the scalability and HA capability of the Amazon EMR option. However, unlike Amazon EMR, HA is an inbuilt capability in the Treasure Data Cloud Data Warehousing solution — all Treasure Data users, regardless of size or status, benefit from this inbuilt HA capability.

The Benefits of the Treasure Data Service-Based Approach

The Treasure Data Cloud Data Warehouse is delivered in a service delivery model. This approach massively simplifies and accelerates the implementation process and extends the value of the Treasure Data core technology and open source innovation to offer substantial technical and economic benefits.

IMMEDIATE, FULLY ELASTIC DEPLOYMENT

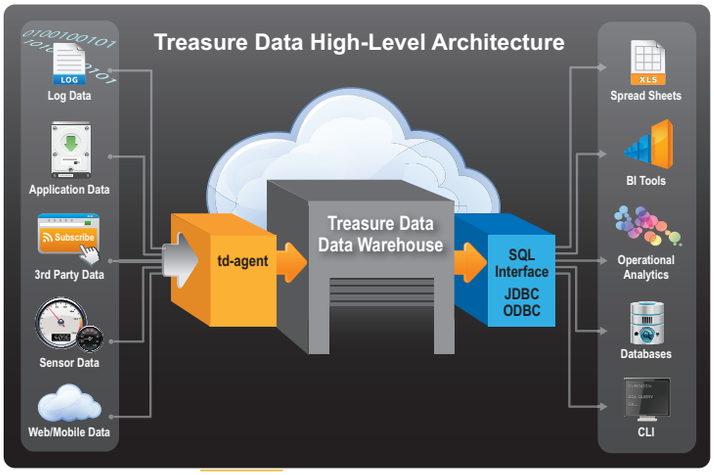
A production environment can be immediately deployed without the need to pull IT staff from other projects. Processing, storage and network resources are completely elastic, and can be scaled up or down as requirements dictate.

BUILT-IN BULK DATA-LOADER

To make the initial data load process as easy as possible, Treasure Data provides a bulk data-loader that can import any amount of data you need to load into the Treasure Data Cloud Data Warehouse.

STANDARDIZED DATA COLLECTOR

Treasure Data provides td-agent — a fully supported Fluentd data collector. This lightweight daemon is installed on each data source and provides both batch and continuous data feeds to the data store. The collector supports standard JSON format transformation for structured, semi-structured and unstructured data types enabling sophisticated ETL routines to be easily developed.



CURATION AND SUPPORT

Treasure Data staff actively curates open source components (e.g. Apache Hadoop, etc.) for each system layer, and integrates and tests them in a complete systems environment prior to production deployment, and ensures you have the latest stable software innovation and best practices.

ROBUST ENVIRONMENT

Treasure Data resides in an Amazon-hosted operational environment, a reliable and secure cloud infrastructure platform that delivers 99.95% availability for on-demand computing, storage, and networking services to power your Cloud Data Warehouse.

FULLY MANAGED

The Treasure Data Cloud Data Warehouse service is managed 24x7 by an expert operations staff familiar with all aspects of the system and service delivery best practices. This eliminates the need to dedicate in-house operations staff to this task and the overhead costs associated with managing an on-premise environment.

TIE ANALYTICS TO RETURNS

The Treasure Data Cloud Data Warehouse service is a subscription-based, pay-per-use solution that ties your investment in analytics directly to your return on investment.

ABOUT TREASURE DATA

Treasure Data combines the power of open source with the scalability and economics of the Cloud and their own patented software innovation, to deliver a big data analytic service that can be running in days not months without specialist IT resources and for a tenth the cost of other alternatives. With customers ranging from social selling start-ups, through online gaming leaders to global corporations, the Treasure Data Cloud Data Warehouse is already hosting over 100 billion rows of live data and is processing 10,000 new messages per second. Treasure Data was founded by some of the smartest engineering minds in the open source world and is backed by leading angel investors from the venture capital, open source and enterprise software markets. For more information, please visit www.treasure-data.com.

NOTES

[1] The Big Data Management Challenge, InformationWeek Reports, April 2012, www.reports.informationweek.com.

[2] Summary survey results of 244 responses are available on request from Treasure Data at info@treasure-data.com.