The State of Hadoop and Data Lifecycle Management

September 2015

INTRODUCTION

Thought leaders and Big Data practitioners completed a Talena survey in which they detailed their adoption and use of Hadoop technologies. The respondents also highlighted the challenges associated with data lifecycle management processes for their Hadoop environment. The most important findings are outlined below.

KEY TAKEAWAYS

- Over 50% of the respondents are actively looking for or implementing a data management solution for their Hadoop infrastructure
- 85% of Hadoop users highlight analytics as their primary use case
- Over 50% of the respondents have deployed three or fewer Hadoop clusters
- Budgets grow substantially as projects move from research to production
- 30% of the respondents store over 100 TB in their largest cluster

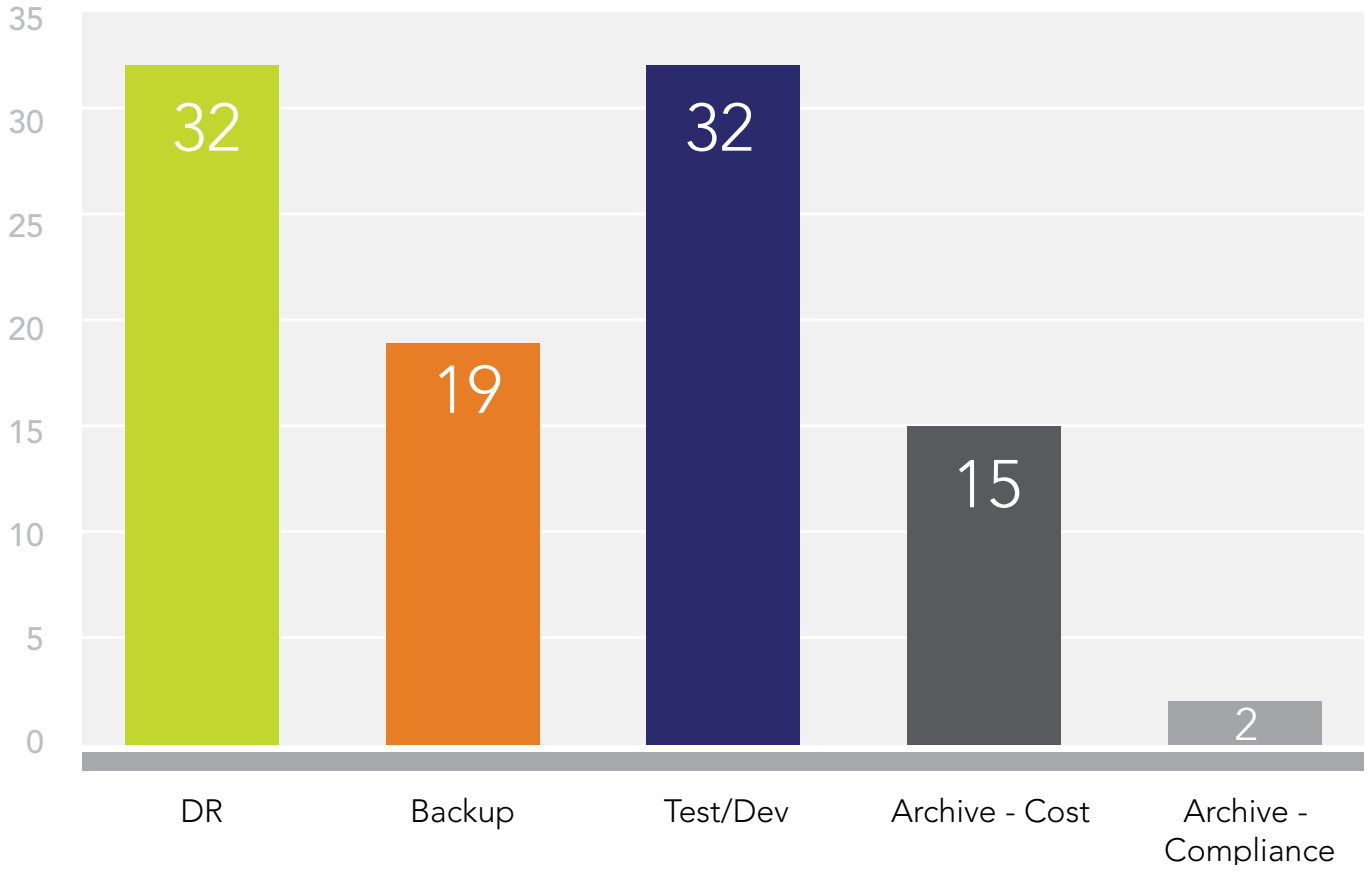
Methodology

We invited people whom we met at Hadoop conferences, thought leaders, or contacts made via our own research into the Hadoop community to fill out the survey. Over 100 responses were recorded. All the raw responses were converted into percentages for the purposes of the charts below, and for a few questions there were multiple responses allowed which made the percentage total above 100.

1

Disaster Recovery and Copying Data for Test/Dev Are Pain Points

Respondents were asked to prioritize various data management processes associated with their Hadoop environment. Disaster recovery and using production data in test & dev environments were highlighted as the two most critical challenges with Hadoop for companies. While the former seems self-evident, the latter could be because the increased presence of DevOps "shops" makes it imperative to support rapid application iterations while still ensuring PII and other sensitive data are masked and protected. These issues play a role in delaying production rollouts.

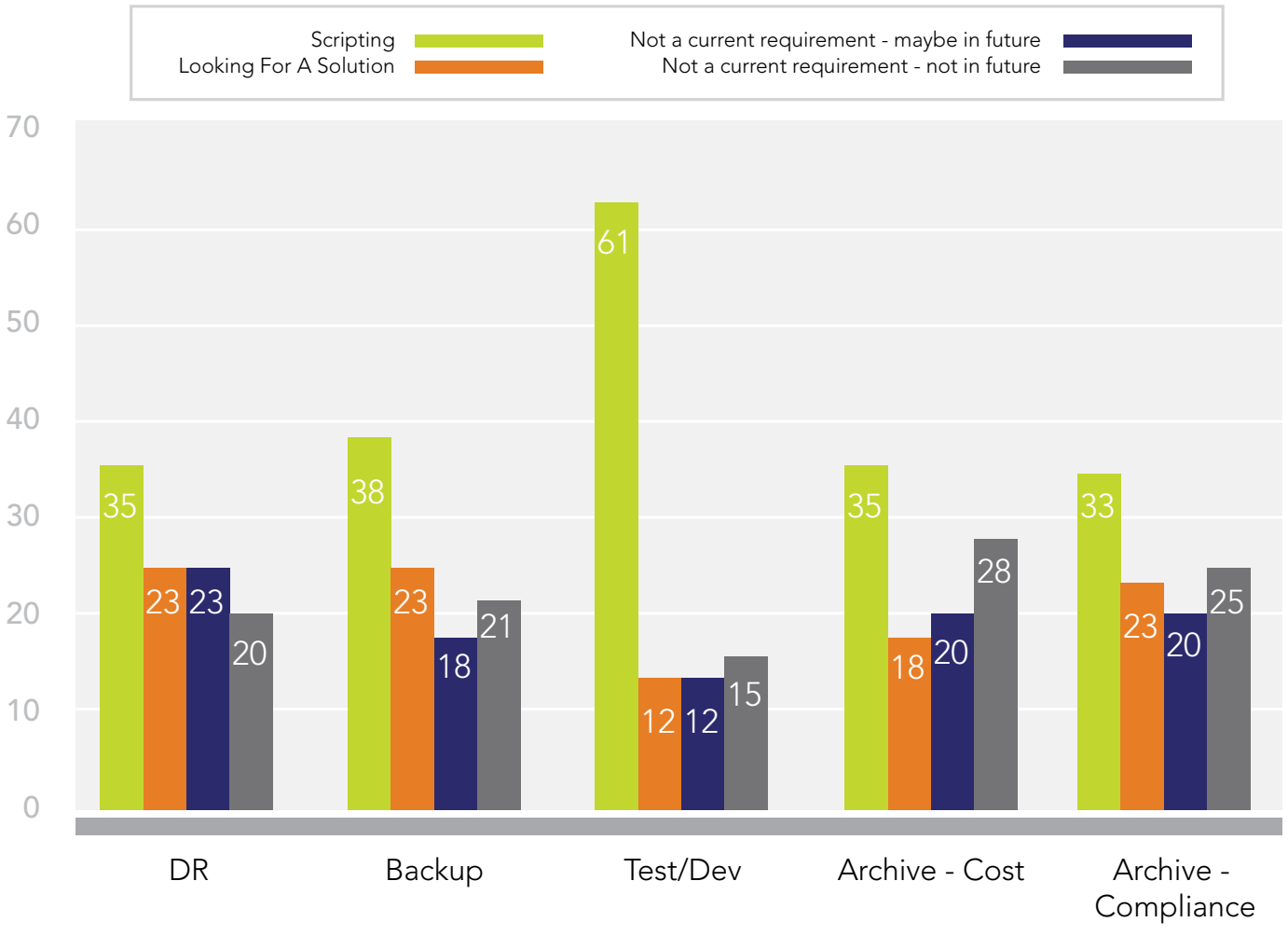


What is your #1 Hadoop challenge as it relates to data management?

2

Scripting Is An Answer, But Is It The Solution?

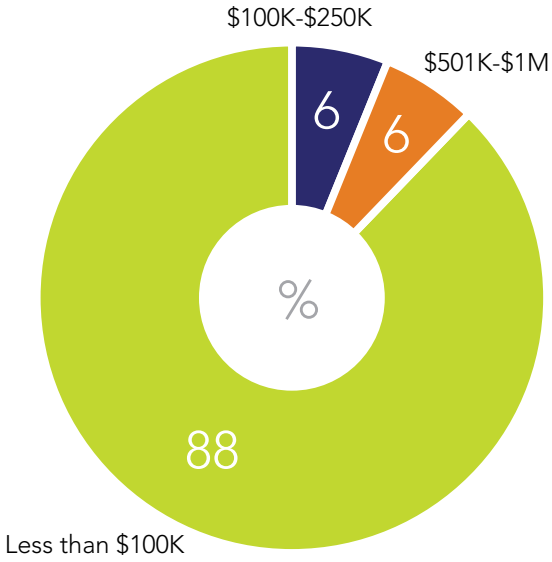
When asked how they were currently solving these challenges, over 60% of those looking for a solution for test/dev/analysis turned to scripting. Over 30% of those who saw disaster recovery, archiving or backups as their biggest issue used scripting while another 20% were actively looking for a solution. In short, implementing some form of data management (even scripting) is relevant to over 50% of the survey group, independent of use case. However, it brings up an interesting question as to whether these increased deployment budgets are wisely spent on a solution that neither scales not supports efficient deployments.



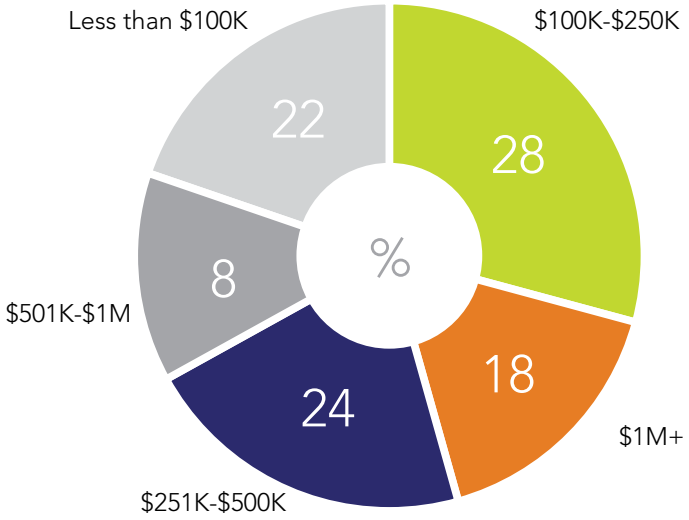
3

There is a Strong Correlation Between Budget and Deployment Phase

Nearly 90% of the enterprises who are still researching their Hadoop options have allocated less than $100K for their future project needs. As these projects move into pilot and production phases, the budgets grow substantially. Just over 20% of those companies that are in pilot or production have budgets less than $100K, while the majority spend significant amounts for infrastructure and engineers to build these next-generation applications (and the scripts for managing their data).



2015 Budget for Research

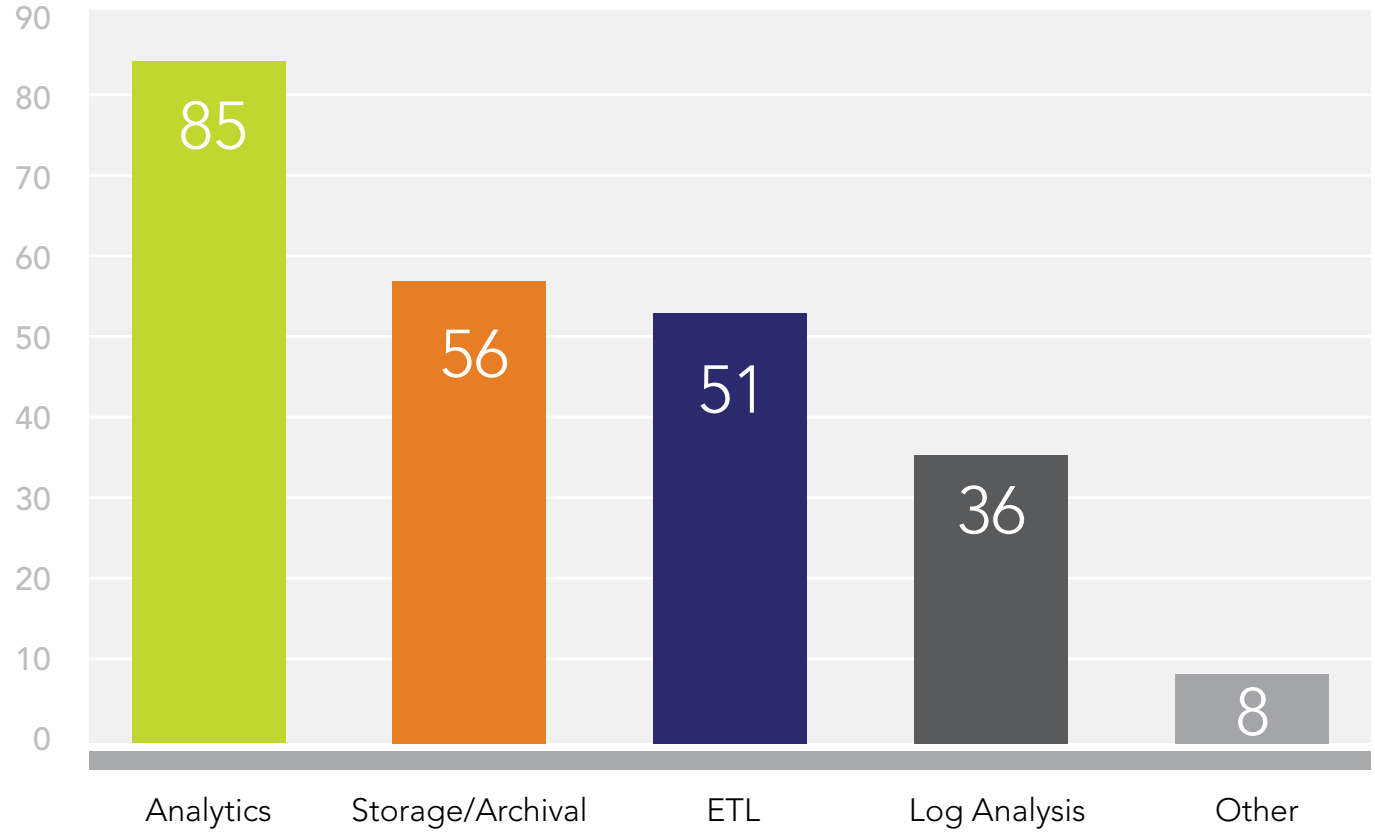


2015 Budget for Pilot/Production

4

Analytics Remains Predominant Use Case With Log Analysis Surprisingly Prevalent

Respondents labeled "analytics" as their #1 use case with over 85% of respondents pinpointing this need. 35% of the respondents use Hadoop for log analysis. Given the emergence of products purposed-built for log analysis like Splunk, Sumo Logic, ELK, and Loggly, we thought this number to be high. Evidently the processing power of Hadoop for large data sets like logs still provides users with compelling value despite all the real-time, machine learning capabilities of these other tools.



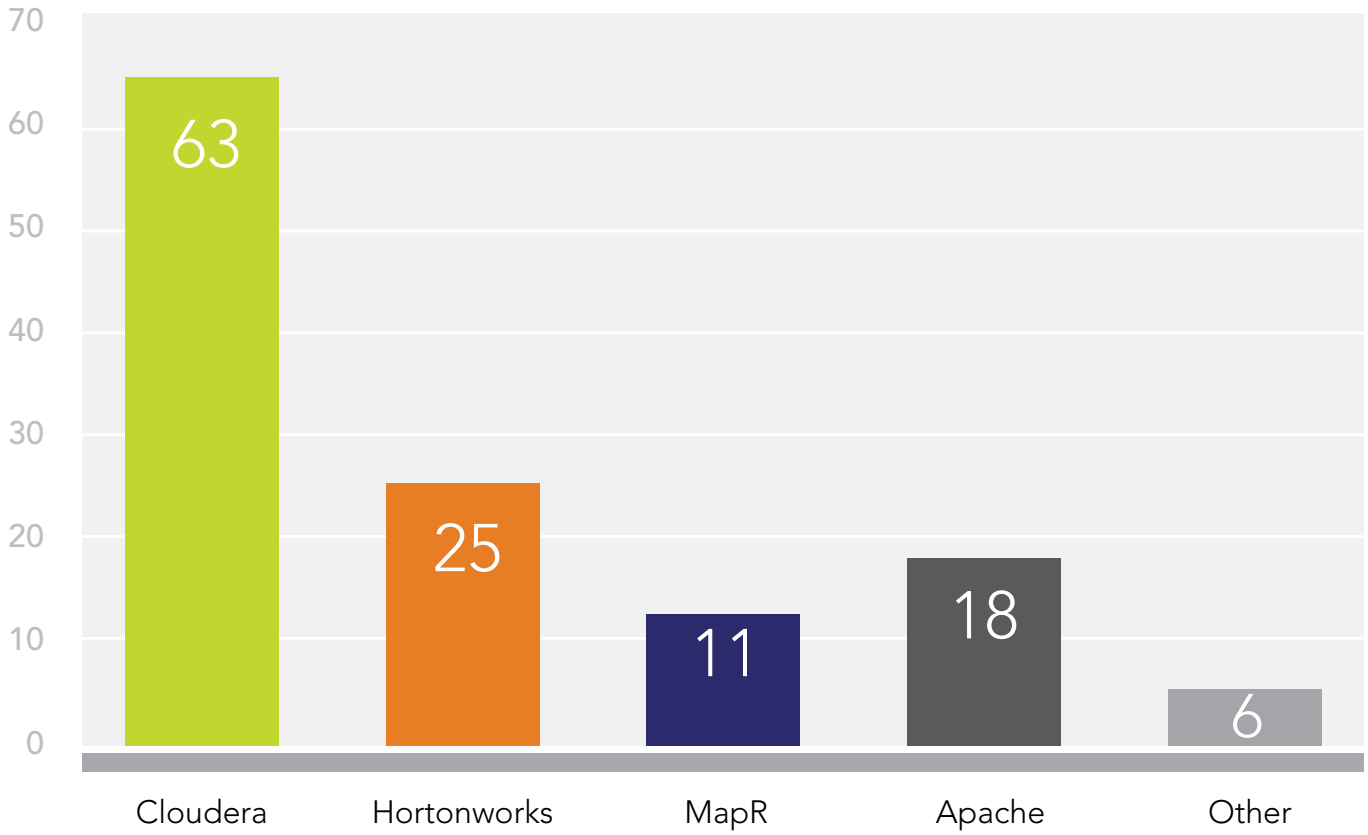
How is your organization using Hadoop?*

*Multiple responses allowed

5

Cloudera Remains The Distribution of Choice

Cloudera was by far the most favored distribution with a greater than 2x advantage over Hortonworks, and nearly 6x over MapR, reinforcing the popular opinion around the relative penetration of the different Hadoop distributions.



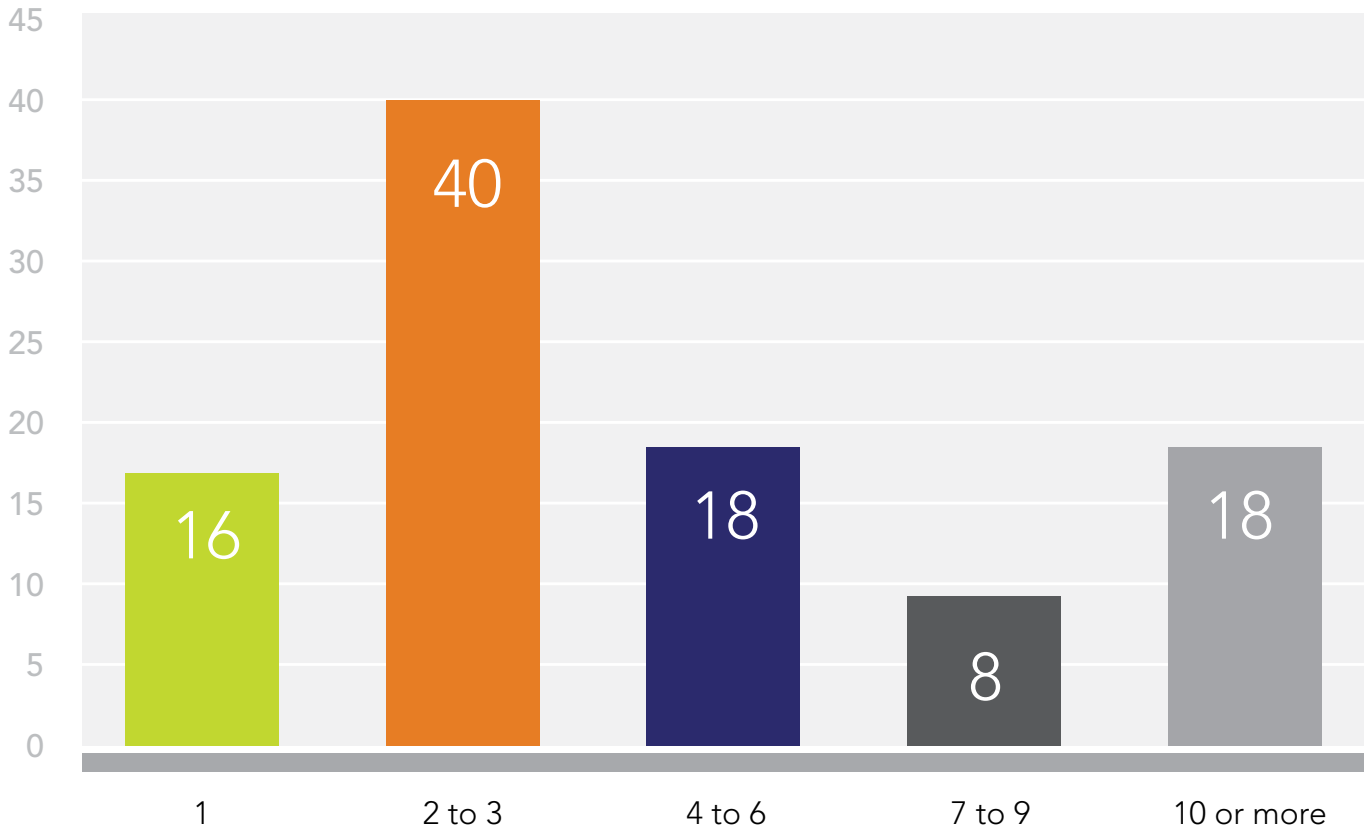
What Hadoop Distribution Do You Use?*

*Multiple responses allowed

6

Most Companies Currently Implementing Three or Fewer Clusters

Over 50% of the companies are using three or less Hadoop clusters (whether production or non-production), while a small percentage of companies have over ten clusters in their environment. This supports the popular theory that there is a smaller group of very large Hadoop deployments, while the broader Hadoop ecosystem is built around more modest cluster sizes that, over time, will grow as companies obtain value from their Big Data applications.

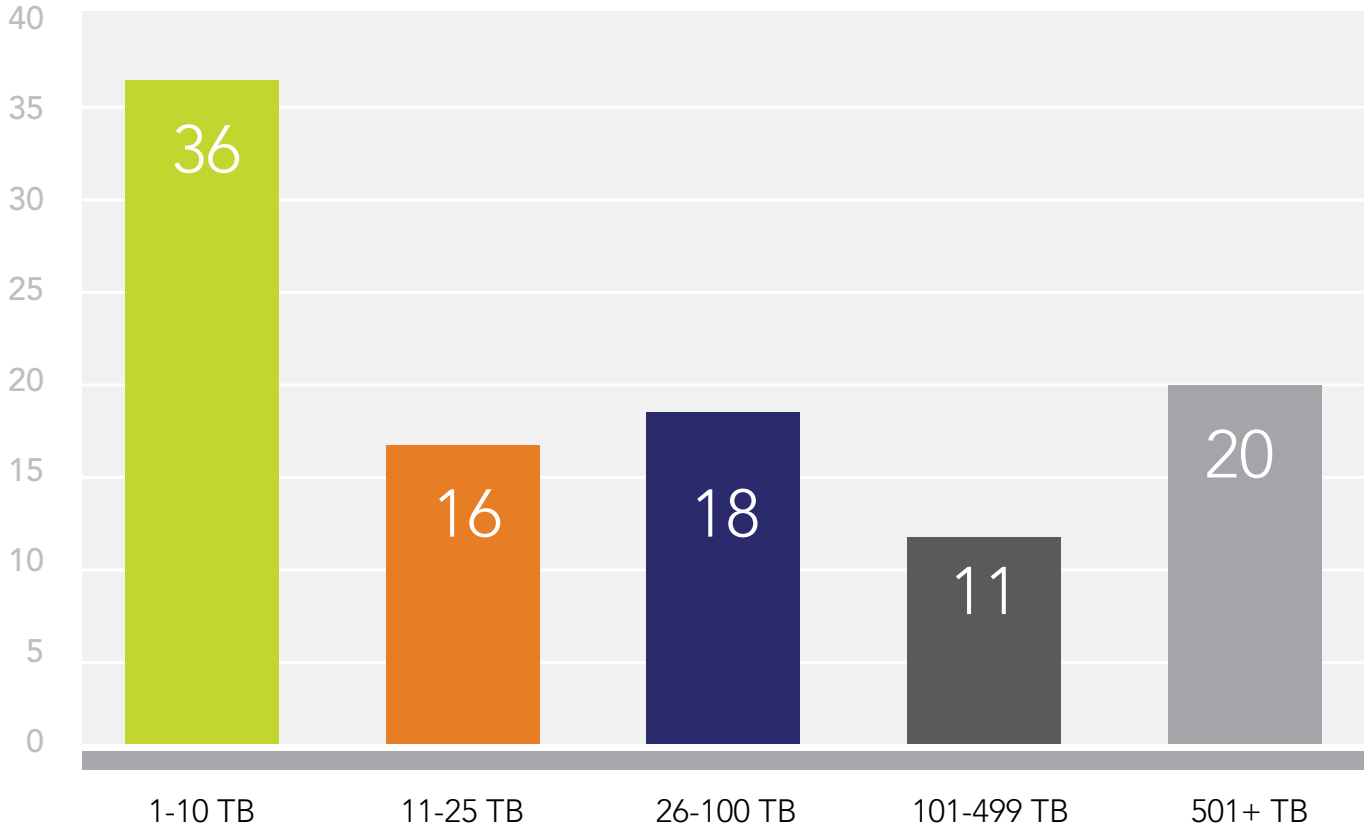


How Many Hadoop Clusters Do You Operate?

7

Nearly a third of the largest clusters contain over 100 TB of data

While companies aren't necessarily deploying large numbers of clusters, they are starting to put large amounts of data in their existing clusters. Over 30% of our respondents store over 100 TB in their largest Hadoop cluster, presumably running analytics, ETL or log analysis on these data sets.



How Much Data Resides In Your Largest Hadoop Cluster?

SUMMARY

With regard to the general adoption of use of Hadoop, our findings support what has been written about in the popular press, with a greater percentage of users actually in production than the overall population. The latter is undoubtedly a result of our sample bias which explicitly includes Hadoop users or those who expressed a preference for Hadoop.

About Talena

Talena, the next-generation data availability management company, solves the problems associated with unavailable or lost data, and potential compliance risk related to Big Data applications.

Our exabyte-scale solution automates backup, test/dev, archive and disaster-recovery functions. With Talena, companies enable rapid application iteration, save engineering and infrastructure resources, and prevent data loss from user error or application corruption.

Please contact us for more information at info@talena-inc.com or visit us at www.talena-inc.com.

Talena, Inc. | 830 Hillview Court, Suite 138, Milpitas, CA 95035 | +1.408.649.6338 | talena-inc.com