Enterprise Strategy Group | Getting to the bigger truth.™

White Paper

Time for Prime Time: Effective Data Management for NoSQL and Hadoop Environments

Understanding Key Data Protection and Data Orchestration Requirements

By Christophe Bertrand, ESG Senior Analyst, and Mike Leone, ESG Senior Analyst

July 2019

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

Contents

Introduction

As organizations continue to look for ways to increase business agility, a need for a modern database architecture that can rapidly respond to the needs of business is more apparent than ever. While an RDBMS still serves as a lifeline for many organizations, the adoption of technologies such as NoSQL and Hadoop are enabling organizations to best address database performance and scalability requirements while also satisfying the goals of embracing hybrid cloud and becoming more data-driven. And with organizations relying so heavily on these new technologies to yield rapid insights that positively impact the business, the need to evaluate how those new technologies are managed and protected is essential. Hadoop and NoSQL workloads are now pervasive in production environments and require "production-class" data protection, yet few data protection solutions offer such capabilities today.

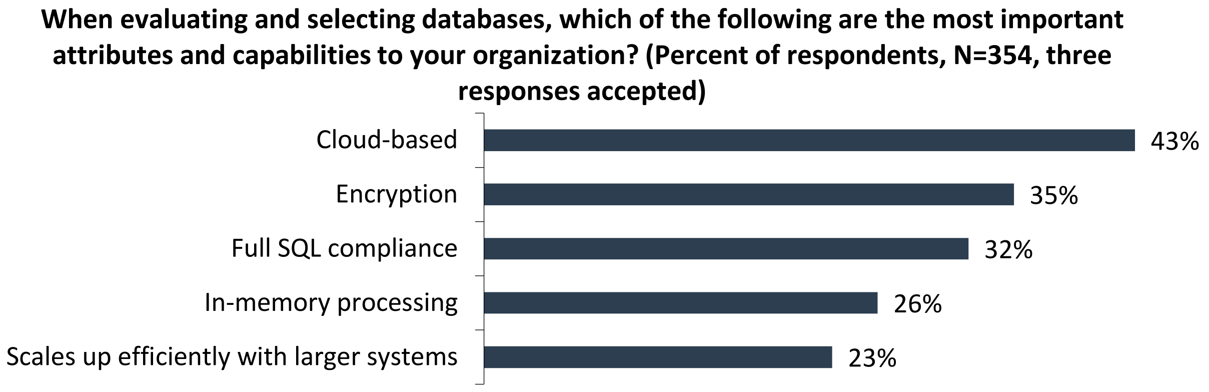
Database Proliferation

ESG research shows that 38% of organizations report that they have between 25 and 100 unique database instances, while another 20% have over 100.[1] The proliferation alone comes with significant cost, management, and data protection challenges, but the different types of database architectures (i.e., RDBMS versus NoSQL) and their locations (i.e., on-premises versus the public cloud) should not be overlooked.

Pervasiveness of NoSQL and the Cloud

NoSQL databases are now the norm, with 78% of organizations reporting in 2017 that they were using a NoSQL database, while another 18% reported having plans for or being interested in using one within the following year.[2] By benefiting from a distributed architecture that can easily scale to meet the demands of the business while enabling organizations to easily transition to the cloud, the pervasiveness of NoSQL is not that surprising. And with the cloud offering organizations cost savings, infrastructure flexibility, and higher availability, it makes sense that when asked about the most important attributes and capabilities that organizations look for when evaluating and selecting a database, being cloud-based was the top response (see Figure 1).[3] It is important to note that while cloud is being prioritized, this is not a scenario where all database instances are being migrated to the cloud. While 62% of organizations currently use a cloud-based production database, just 10% of organizations said their primary infrastructure deployment strategy for net-new database deployments is a public cloud service.[4]

Figure 1. Top Five Attributes and Capabilities When Evaluating and Selecting Databases



Source: Enterprise Strategy Group

[1] Source: ESG Survey, *Enterprise Database Trends*, January 2017. All ESG research references and charts in this white paper are taken from this report, unless otherwise noted.
[2] Source: ESG Brief, *Market Disruption: Next-generation Databases*, April 2017.
[3] Source: ESG Brief, *Database Purchase Criteria*, June 2017.
[4] Source: ESG Brief, *The Database Market: Radical Shift to the Cloud*, April 2017.

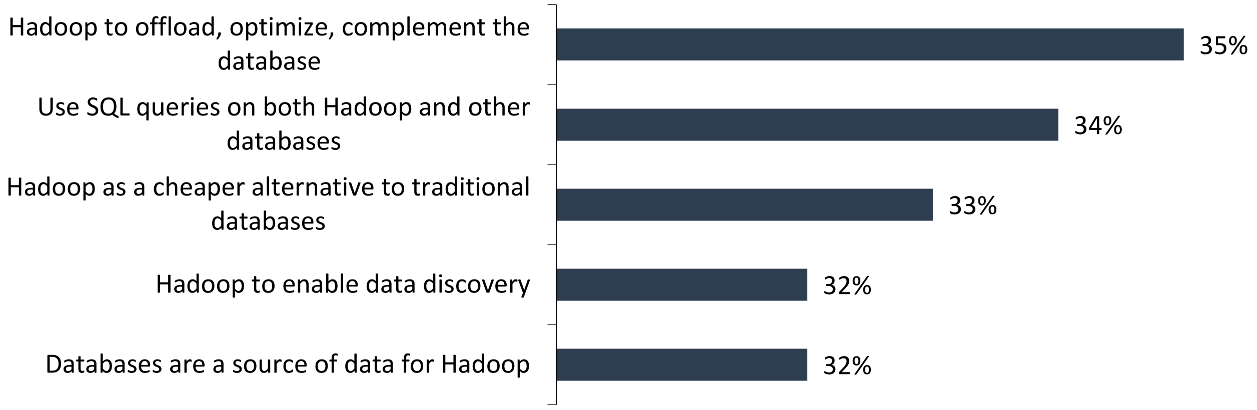
Enter Hadoop

Databases are the start, but to satisfy the data-driven initiatives, organizations must continue to enhance their database deployments with more robust infrastructure that can help with processing constantly growing datasets. This is where platforms like Hadoop and Spark come in.

For Hadoop specifically, despite post-hype pessimism, the technology is still widely used, with ESG research showing that 50% of organizations leverage the technology to some extent. So, how are organizations leveraging Hadoop alongside their databases? The most-often cited way, with 35% of responses, is by leveraging Hadoop to offload, optimize, and/or complement an existing database. Thirty four percent of organizations are looking to use SQL queries on both Hadoop and other databases, while 33% view Hadoop as being a cheaper alternative to traditional databases. Thirty two percent feel Hadoop better enables the discovery of data, and 32% also cite leveraging databases as a source of data for Hadoop.

Figure 2. Top Five Ways Organizations Use Databases Alongside Hadoop

Which of the following represents the manner(s) in which your organization uses or plans to use databases alongside Hadoop and/or data lakes implementations? (Percent of respondents, N=298, multiple responses accepted)



Source: Enterprise Strategy Group

Business and IT Imperatives

In recently completed research, respondents surveyed by ESG identified that cost and improving SLAs for data and applications are among the top data protection mandates from their organizations' IT leadership.[5] This should come as no surprise as impacts on production environments and their associated mission-critical applications and data can be very damaging, sometimes caused by the ransomware epidemic that has spread across industries around the world.

Impact of Downtime and Data Loss

ESG's real-world SLAs research points to organizations having very little tolerance for downtime for servers running "high priority" workloads. Fourteen percent of organizations report tolerating no downtime ever, while another 36% can accept less than 15 minutes.[6] This only highlights the criticality of having a solid backup and recovery infrastructure in place for all mission-critical production environments. NoSQL and Hadoop workloads, as outlined previously, fall in that category.

Impacts of downtime or lost data can be catastrophic for organizations. ESG's research identified direct loss of revenue, loss of customer confidence, missed business opportunity, loss of employee confidence, and damage to brand integrity as

[5] Source: ESG Master Survey Results, *2018 Data Protection Landscape*, to be published.
[6] Source: ESG Master Survey Results, *Real-world SLAs and Availability Requirements*, May 2018.

the five most commonly cited potential impacts.[7] Clearly, downtime and lost data are business issues that demand full executive attention and mitigation measures.

SLA Requirements for Production Applications and Databases

Zooming in on production environments and the amount of tolerable data loss, ESG identified that most organizations simply don't tolerate losing much data at all. IT leaders intuitively know that losing data that needs to be recreated is inherently an expensive exercise to be avoided. ESG's research shows that 51% of organizations indicate that the threshold is placed at 15 minutes of data loss, or how much data can be lost without significant impact to the business. The same research also highlights that 16% of organizations can only tolerate less than five minutes of data loss for "high priority" data.[8] From a data protection perspective, this means that these environments must be protected very frequently or continuously. Recent trends such as the rise of ransomware and its consequences on data loss only reinforce this requirement.

Mitigating the Ransomware Epidemic

There isn't a week that passes without news breaking of another cyber-attack or ransomware extortion scheme. Most recently, big data databases have been specifically targeted. Cyber-attackers have even organized "competitions" focused on MongoDB installations for example, with IT press reporting tens of thousands of attacks where the databases were publicly exposed.

A cyber event affecting the integrity or availability of data is akin to a logical corruption event in data protection terms, one that requires remediation via recovery. To recover and minimize data loss, a recent backup must be available and meet the service levels mandated by IT and organizational leadership. This means that data protection service levels such as recovery point objectives (RPOs) and recovery time objectives (RTOs) are aligned with best practices for production environments to minimize business impact and disruption. It also means that frequent backups that allow a "rollback" to a pre-attack time is a best practice to implement in combination with ransomware-specific mitigation capabilities, such as early detection, flagging/warnings, etc.

The Role of Cloud

The adoption of cloud as an extension to "traditional" on-premises environments is a reality across the board:

- ESG research shows that the adoption of cloud and multi-cloud environments has accelerated over the past few years. Overall adoption of public cloud has gone from 57% of organizations using in 2013 to 85% today.[9]
- As users consider deploying new applications, they often turn to cloud for those deployments either as a part of a "cloud-first" strategy (29%) or a hybrid on-premises/cloud approach (47%).[10]
- The number of organizations running production applications on public cloud infrastructure continues to climb with 46% of current IaaS users doing so in 2018, up from only 27% three years ago.[11]

The research is clear: Cloud has become a key component of production environments, replacing or extending the data center.

[7] ibid.
[8] ibid.
[9] Source: ESG Master Survey Results, *2018 IT Spending Intentions Survey*, December 2017.
[10] ibid
[11] Source: ESG Brief, *2018 Public Cloud Infrastructure Trends* , April 2018.

For this reason, the flexibility to move these workloads, to migrate them or repatriate them, is an important capability to consider as part of the deployment process to optimize operational efficiency. In the context of big data databases, this flexibility/mobility also supports the efficiency of the development and management teams.

Key Data Management Considerations for Hadoop and NoSQL Deployments

Hadoop and NoSQL environments are pervasive and support many organizations' business processes as outlined previously. Protecting these business-critical workloads is a mandate for IT, as is establishing processes for leveraging cloud destinations and movement in and out of the data center.

Protection Considerations

NoSQL environments and Hadoop environments do not come with easy-to-deploy, built-in data protection. While several utilities exist, they typically require expertise and custom coding that is not always easy to deploy or maintain.

There is a false sense of security because NoSQL and Hadoop databases are inherently self-replicating, which only covers the risk of hardware failure/node failure. Also, production NoSQL and Hadoop environments can quickly grow in scale and be very distributed, making the process of backup and recovery more time consuming unless scalable and modern data protection techniques are used, such as point-in-time backup and recovery, incremental-forever backup, granular recoveries, no agents, etc.

There are many "flavors" of NoSQL in the market today (Mongo DB, Cassandra, Couchbase, etc.). It is important that the backup and recovery solution not only scales, but also maintains awareness of the types of databases it is protecting, and their nuances, to deliver a unified, coherent, and consistent backup process as well as granular recoverability. Awareness of data can also support compliance or privacy requirements with data masking, for example.

Finally, hybrid/cloud support and movement off-premises and back on-premises are critical for operational efficiency and IT flexibility. This is where orchestration can play a role.

Orchestration Is Critical

The data protection process is a workflow of events or operations that need to take place in a predefined sequence, or runbook. Many steps need to take place to ensure the proper backup, recovery, or migration of a NoSQL environment. Leveraging a robust policy engine is an architectural requirement that powers many data operations such as data masking, data mirroring, and the very critical incremental backup process, among others.

With these data-aware orchestration capabilities, organizations can optimize their archiving process, support testing and developments functions by providing "real" data clones, and better manage cloud migration efforts.

Flexible Migration Is a New Imperative

As organizations have now extended their on-premises environments to include cloud destinations, "freedom" or flexibility of movement of production workloads is an operational must have, whether moving NoSQL and Hadoop workloads across data centers, from on-premises to cloud, or from any source to any destination.

The ability to leverage automation is much needed given the scale of complexity of these databases. It is necessary for obvious operational efficiency reasons, but also to improve the recoverability of these environments.

Taking human error out of the management and data protection loops is becoming a reality as more machine learning and artificial intelligence capabilities get integrated in modern data backup and management solutions. With more automation and intelligence accompanying IT, it is easier to reduce the impact of ransomware, optimize data recovery service levels, and operate more autonomous environments, freeing up precious IT or development resources to support the business.

The Bigger Truth

Hadoop and NoSQL backups using traditional data protection methods are no longer enough. The proliferation of Hadoop and NoSQL in what can be petabyte-scale creates business and IT imperatives of stringent data protection SLAs. The risk of data loss, the threats that exist with ransomware, and best recoverability practices to support business objectives mean that IT leaders must apply enterprise-class data protection solutions to be ready for production prime time. Protection only is not enough: Orchestration is key and requires intelligence capabilities to better enable IT to manage complex data protection operations and migrate these workloads seamlessly between/across cloud and on-premises environments.

That's where Cohesity can help. The company focuses on enterprise data management and has recently added a solution for Hadoop and NoSQL environments running on-premises or in the public cloud. Cohesity's solution is architected for petabyte scale, data-aware, and powered by machine learning. Its most recent version, version 6.3, added some strong capabilities in the areas of policy automation, ransomware mitigation, and advanced granular backup and recovery.

IT leaders evaluating their data protection strategies for production Hadoop and NoSQL environments should put Cohesity on their lists based on its breadth and depth of functionality for these environments, and for its innovative use of machine learning and automation.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.