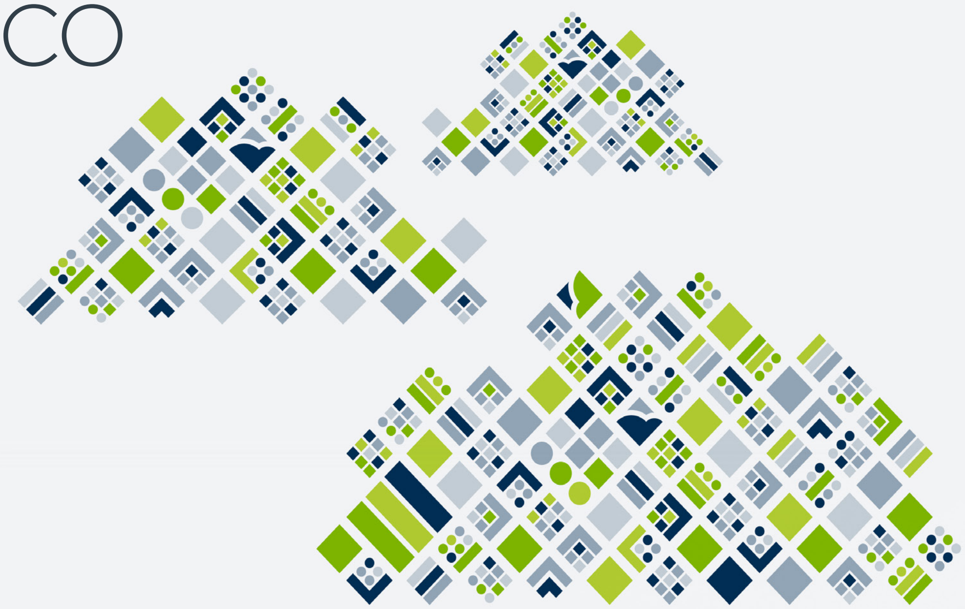
wanDISCO



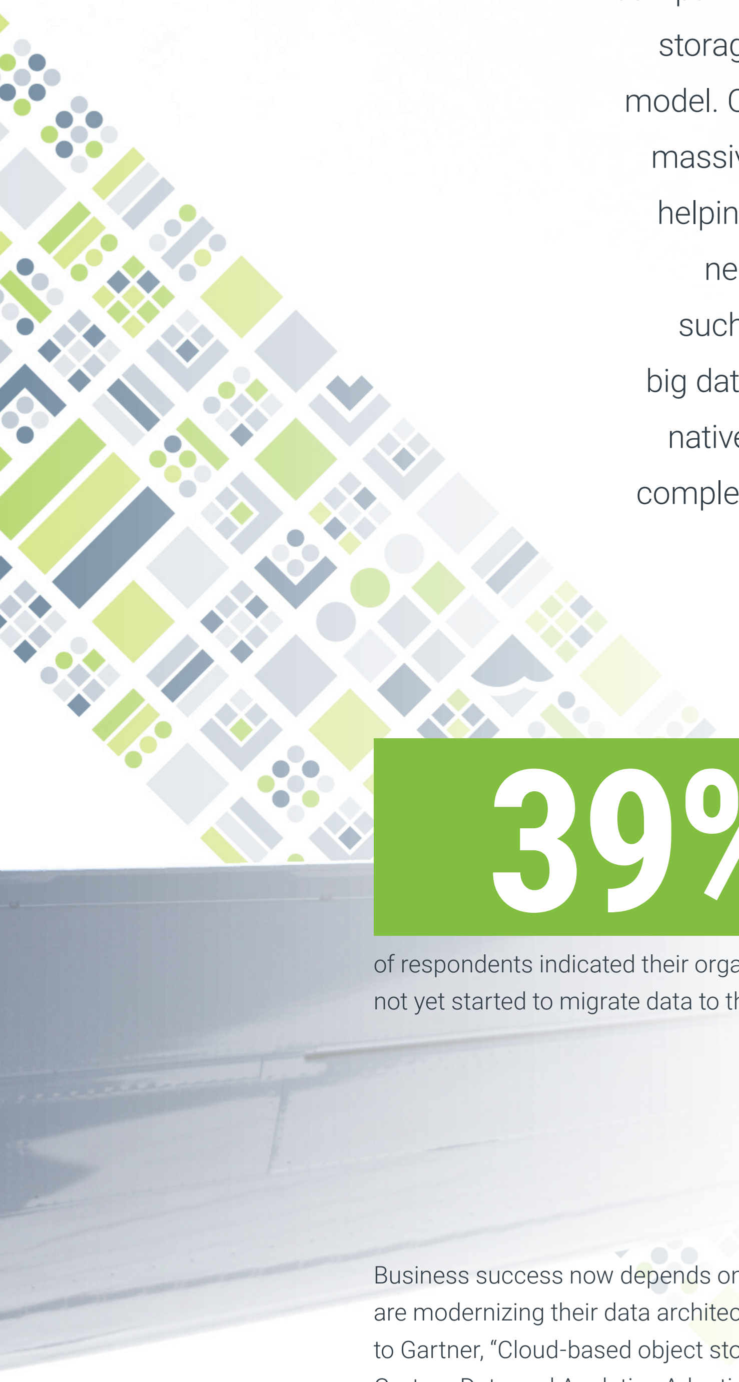
eBook

HADOOP DATA MIGRATION SURVEY AND REPORT

BENCHMARK REPORT

ABOUT THIS DOCUMENT

This document is for cloud and data architects that are responsible for or planning to migrate their on-premises Hadoop implementations to a modern cloud architecture. The document explores some of the business risks associated with big data to cloud migrations. It highlights third-party research as well as results of a survey conducted with 220 cloud and data architects, and other technical specialists to provide insights on their data migration plans and strategies, and compares the costs and risks associated with manual vs automated big data migrations.



The cloud is a natural home for big data. In the cloud, companies can take advantage of inexpensive, scalable storage and compute platforms leveraging an OPEX model. Cloud Service Providers (CSPs) are able to spin massive compute clusters up or down automatically, helping organizations optimize costs and reduce the need for people with deep expertise in managing such deployments. CSPs have also enhanced their big data solutions providing greater functionality over native Hadoop offerings, and greatly simplifying the complexities for organizations having to manage their own Hadoop implementations.

39%

of respondents indicated their organizations have not yet started to migrate data to the cloud.[1]

71%

of survey respondents indicated that the key reason they are moving onpremises data to the cloud is to adopt scalable cloud storage to free systems from capital constraints.[1]

Other key use cases cited include Disaster Recovery, Cloud Analytics and Data Lake consolidation.

Business success now depends on how effectively organizations utilize their data. To do so companies are modernizing their data architecture, and moving their Hadoop deployments to the cloud. According to Gartner, “Cloud-based object storage was the most widely deployed technology identified in the 2019 Gartner Data and Analytics Adoption Survey, with 57% already using and 39% planning to within the next two years.”2 While we believe the move to the cloud and cloud analytics is making companies more competitive, lean and nimble, many companies are worried about the complexities of migrating such large volumes of data, the cost and effort of manual migration requirements, the perceived need for extended periods of downtime, and fear of lost productivity.

MANUAL DATA MIGRATION BUSINESS RISKS

Manual data migration is a custom, tactical approach to copying big data. When administrators manually migrate data, they create, manage, schedule and maintain custom scripts to migrate large and continually increasing data sets, which leads to the following business risks.

57%

of survey respondents indicated that zero or only hours of total downtime was acceptable for data migrations[1]

65%

of survey respondents indicated that they were only somewhat, slightly or not at all confident that their on-premises to cloud replication is/will be accurate.[1]

Respondents indicated the top-3 items driving their data migration costs are:[1]

1. Create, manage, schedule and maintain custom scripts
2. Manual intervention for anticipated failures
3. Manually bringing data back into synch

BUSINESS DISRUPTION

Manual migration of continually growing and actively changing data sets often requires significant disruption to on-premises applications. Administrators who choose incremental migration strategies that bring data sets to the cloud over many months, face handling disruptive updates and incur the risk of not meeting their enterprise SLAs. How much downtime is acceptable to your business?

DATA INCONSISTENCY

With manual migration relying on custom scripts that focus on copying data, how does the team validate that the replication is accurate? Manual reconciliation at scale does not guarantee completely consistent data outcomes. Also, how will administrators handle new updates that occur during migration? Typically, data that are being modified or created during migration are not catered for with manual migration approaches.

HIGH IT RESOURCES

The overhead of activities to attempt non-disruptive, no-downtime big data migration are significant. It requires resources to create, test, manage, schedule and maintain custom migration scripts, as well as validate the ongoing results. Due to the custom nature of manual migrations, the program is prone to delays. According to Gartner, "Data migration projects often exceed their budget by 25% to 100% or more, due to a lack of proactive attention to data quality issues (a problem that persists postmigration)."3to the following business risks.

AUTOMATED DATA MIGRATION BENEFITS

To reduce business risks as well as the burden of custom coding data migration scripts, many organizations have started to use packaged tools. According to Gartner, 57% have adopted data migration/consolidation as a data integration tool and 34% have adopted data migration/consolidation as a data quality tool.[3]

BUSINESS CONTINUITY

In order to avoid business disruption, companies need to leverage solutions that allow existing on-premises environments to operate in parallel with the new environments during the migration process. This can only be achieved with solutions that support true active transactional replication, capable of moving data as it changes in both the old and new environments.

Data and analytics have become part of business-critical workloads, which require the systems to be highly available.

According to the Ponemon Institute, downtime can cost businesses nearly $9,000 per minute.[4]

DATA CONSISTENCY

Enterprises are recognizing the benefits of greater flexibility and functionality, stronger resilience and disaster recovery, and improved business performance when using a multi-cloud strategy. This makes data consistency across different cloud platforms and regions essential. Manual intervention to ensure data consistency across actively changing data sets is not a scalable or feasible approach. Automated capabilities to ensure data consistency are an imperative.

“A recent Gartner survey on cloud adoption revealed that

80%

of respondents using the public cloud were using more than one Cloud service provider (CSP).”[5]

IT EFFICIENCY

According to Gartner, “Given the volume and complexity (the number of structures, such as database tables) of data involved in most application modernization efforts, automated assistance is required.”[3] While many organizations have started to leverage data migration tools to reduce the burden, many still rely on custom scripts. Moving to an automated approach would provide these latter companies with significant opportunities for improving their IT efficiency.

34%

of respondents use migration tools to reconcile updates. However, over 26% still do this manually, while another 13% don’t have any plans to reconcile the new updates.[1]

AUTOMATED DATA MIGRATION TOOLS REQUIREMENTS

It is clear that in order to reduce the business risk associated by big data cloud migrations, companies need to leverage modern solutions that allow existing and new environments to operate in parallel thereby ensuring business continuity, data consistency and IT efficiency. Our survey asked participants to rate the capabilities required in such solutions. Capabilities and their response rates are listed below:[1]

SELECTIVE MIGRATION - Allows selection of which data sets should be migrated, or excluded from migration to specific clusters in the new environment (71%)

HADOOP & OBJECT STORAGE - Works across a variety of big data source and target environments, including all major Hadoop and object storage technologies (67%)

AUTOMATIC OUTAGE RECOVERY - System eliminates the need for manual response to system failures, including network outages and other disruptions (64%)

RAPID AVAILABILITY - Minimizes the time required to bring workloads to the cloud by making each data location available for use as soon as bandwidth allows (61%)

PETABYTE SCALE - Migrates big data sets at any scale to cloud storage without needing to halt changes made to the data sets during migration (58%)

ONE CLICK - Following data set selection, allows users to and push start to begin replication. System notifies users when migration is complete. No scripts, no code maintenance, no transfer devices, no scheduling, and no need for reviewing (55%)

BANDWIDTH MANAGEMENT - Optimizes bandwidth use by eliminating the need for repeated transfer of data, and enforcing limits on bandwidth use (44%)

ACTIVE REPLICATION - As changes occur anywhere in the source system, the solution creates and ensures target system data consistency (40%)

ONE PASS - Migrates existing data sets with a single pass through the source storage system, eliminating the overhead of repeated scans (40%)

IN SUMMARY

The advantages of moving on-premises Hadoop implementations to the cloud are clear. Reduced operational overhead, cost optimization, as well as enhanced big data services and object storage provided by CSPs are just some of the benefits. Another key driver for these migrations is cloud analytics. Organizations need to modernize their data infrastructure in order to make more effective use of their valuable data assets, which is so critical in today's digital revolution. The companies that can most effectively leverage their data will have a competitive advantage, while those that don't will cease to exist.

There are few big data migrations that can be switched overnight from an on-premises to a public cloud platform without significant risk for the business. Migrating a data lake to the cloud, with no business disruption, while keeping massive volumes of data consistent with the on-premises platform, presents complex technical challenges, and can require many IT resources resulting in high costs or potential project delays. These risks are pervasive in manual data migration projects and are preventing some organizations from their big data to cloud migration due to their fear of lost productivity.

Automated data migration tools are being used by many companies to reduce the above business risks. As more companies modernize their data infrastructure even more will need to adopt this strategy, as it will be essential for successful big data to cloud transformations.

REFERENCES

[1] WANdisco Hadoop data migration survey and report, March 2020

[2] Gartner "Choosing the Right Path When Exploring Hadoop's Future," Merv Adrian, Rick Greenwald, 27 February 2020

[3] Gartner "Make Data Migration Boring: 10 Steps to Ensure On-Time, High-Quality Delivery," Ted Friedman, 13 December 2019

[4] "Cost of Data Center Outages", Published January 2016, Ponemon Institute© Research Report

[5] Gartner "Are You Ready for Multicloud and Intercloud Data Management?," Adam Ronthal, et al, 24 May 2019

5000 Executive Parkway, Suite 270
San Ramon, CA 94583

www.wandisco.com

Talk to one of our specialists today

US	+1 877 WANDISCO (926-3472)
EMEA	+44 (0) 114 3039985
APAC	+61 2 8211 0620
All other	+1 925 380 1728

Join us online to access our extensive resource library and view our webinars.

Follow us to stay in touch

EB-HDMBR-200908