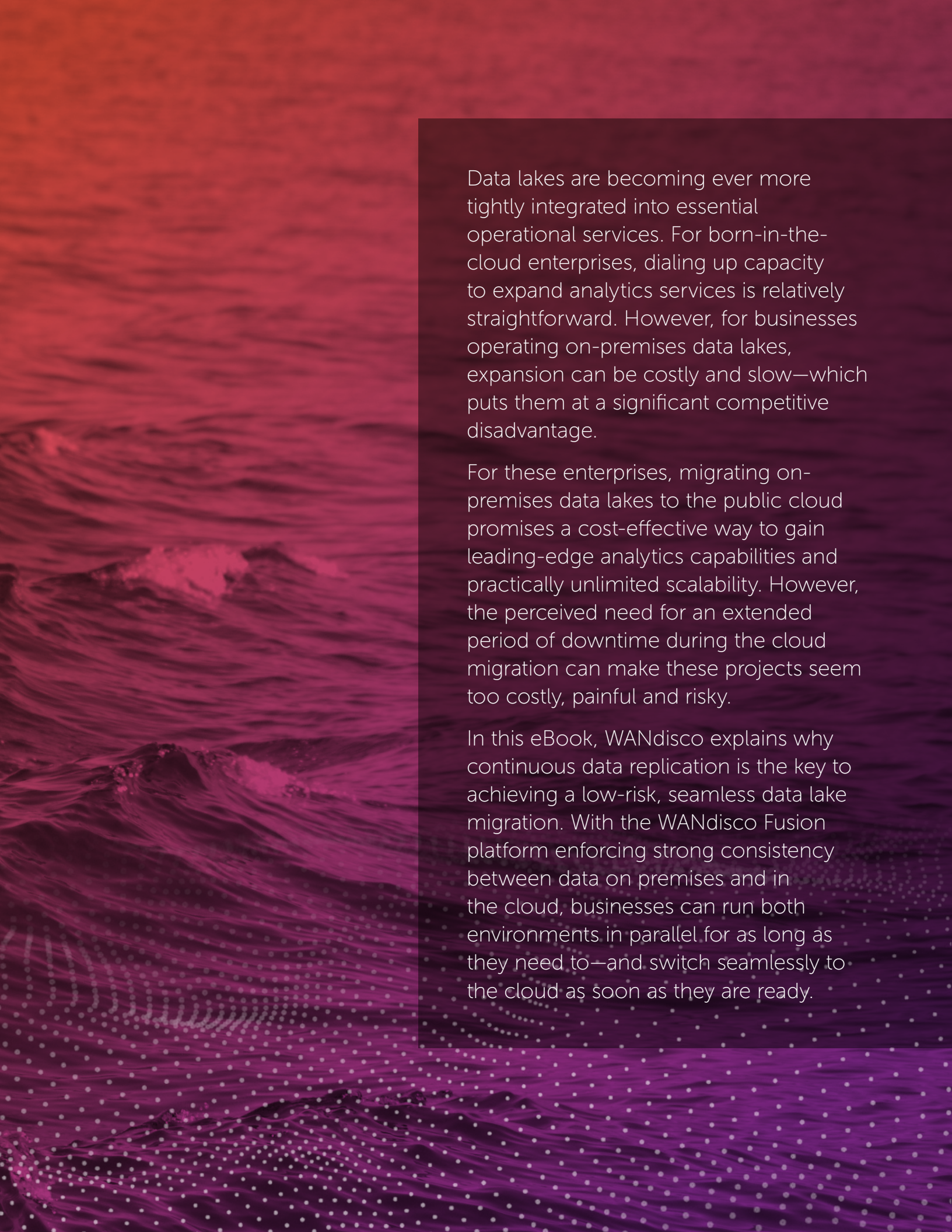
wanDISCO
The LiveData Company

MIGRATE YOUR DATA LAKE TO THE PUBLIC CLOUD WITH ZERO-DISRUPTION

Why continuous data replication is crucial to de-risk your data lake migration project



Data lakes are becoming ever more tightly integrated into essential operational services. For born-in-the-cloud enterprises, dialing up capacity to expand analytics services is relatively straightforward. However, for businesses operating on-premises data lakes, expansion can be costly and slow—which puts them at a significant competitive disadvantage.

For these enterprises, migrating on-premises data lakes to the public cloud promises a cost-effective way to gain leading-edge analytics capabilities and practically unlimited scalability. However, the perceived need for an extended period of downtime during the cloud migration can make these projects seem too costly, painful and risky.

In this eBook, WANdisco explains why continuous data replication is the key to achieving a low-risk, seamless data lake migration. With the WANdisco Fusion platform enforcing strong consistency between data on premises and in the cloud, businesses can run both environments in parallel for as long as they need to—and switch seamlessly to the cloud as soon as they are ready.

WHY USE WANDISCO FUSION FOR DATA LAKE MIGRATION?

- With WANdisco Fusion, businesses can move HDP data lakes to cloud platforms such as Microsoft Azure and Amazon S3 rapidly and without risk of downtime.
- Businesses can drive cloud migrations without making major changes to their data lake architecture, dramatically reducing the cost of cloud migration.
- Once in the cloud, organizations can unlock cost-effective scalability and access to cutting-edge analytics capabilities—delivering rapid ROI on WANdisco Fusion.

WHY MIGRATE YOUR DATA LAKE TO THE CLOUD?

From the total cost of ownership (TCO) perspective, on-premises platforms cannot hope to compete with leading cloud vendors. Because data lakes are so complex to build, manage and maintain, personnel costs often constitute as much as 70 percent of the TCO for a typical on-premises data lake—costs that all but evaporate after migrating a data lake to the cloud.

Although increasing numbers of CIOs are targeting a cloud-first or cloud-only future, many enterprises face the reality of an on-premises data lake that is growing steadily year after year. These businesses can find themselves in a Catch-22 situation: the costs of scaling out the data lake are spiraling out of control, but the platform is so critical to day-to-day operations that taking it offline for a cloud migration project is simply not an option. What's the way forward?

Because data lakes are so complex to build, manage and maintain, personnel costs often constitute as much as 70 percent of the TCO for a typical on-premises data lake—costs that all but evaporate after migrating a data lake to the cloud.

WHAT ARE THE OPTIONS?

When it comes to cloud migration, businesses have three options: lift-and-shift their entire data lake environment to the cloud, re-create their data lake in the cloud, or a hybrid of these two approaches. Here, we consider the advantages and disadvantages of each method.

LIFT-AND-SHIFT

Timeline: 6-18 months

Cost: Lowest

Lift-and-shift is the fastest approach to cloud migration, as it allows businesses to avoid making changes to their analytics environment. As a result, the business can use a service such as AWS Snowball or Azure Data Box to transport data to the cloud extremely quickly. However, the lift-and-shift method means that any technical debt—i.e., sub-optimal software systems and business processes—in the on-premises analytics environment will be carried over into the cloud.

To ensure a non-disruptive migration, the lift-and-shift approach requires some specific technical capabilities. The business must be able to copy their production environment without shutting down their data lake, and then synchronize that copy with all the changes that happened while the data was being shipped to the cloud. And to avoid the risks of a big bang cutover, the business will also need to run their on-premises and cloud environments side-by-side for validation. Throughout the testing period, it will be crucial to ensure that data on premises and in the cloud are consistent at all times.

RE-CREATE-FROM-SCRATCH

Timeline: 3-5 years

Cost: Highest

Businesses that want to avoid bringing technical debt into the cloud have the option of building a data lake from the ground up. The key advantage of this approach is that it enables businesses to harness cloud-native analytics and security technologies, which are often many generations ahead of on-premises capabilities. However, these benefits come at a considerable cost in terms of time and investment.

Companies that decide to re-create-from-scratch in the cloud will also need to copy production data into the new environment without causing disruption for business users—and update that data with any changes that happened on-premises during transit. Because the cloud environment is likely to be extremely different from the on-premises environment, running the two platforms side-by-side may not be a practical way to validate the new solution. As a result, testing and cutting over to the cloud platform will present a greater challenge than a lift-and-shift approach.

HYBRID APPROACH

Timeline: 12-24 months

Cost: Medium

Enterprises that prefer to strike a balance between a timely migration and reducing technical debt may opt for a hybrid of the lift-and-shift and rebuild-from-scratch methodologies. In the hybrid approach, businesses analyze each component of their on-premises lake for data dependencies. Components with large numbers of dependencies with other systems will be strong candidates to lift and shift into the cloud, as this avoids the cost and complexity of additional development work.

Conversely, components with few data dependences may be ideal candidates to re-architect for platform-as-a-service offerings, such as Salesforce for customer relationship management or Adobe for digital analytics.

As with the lift-and-shift and re-create-from-scratch options, a hybrid approach will require the business to migrate data from their production environment without shutting down critical services. Rather than attempting to move large amounts of data over the wire, most enterprises will use a data transport solution—such as AWS Snowball or Azure Data Box—provided by their cloud vendor. And once migrated, it will be equally important to keep the data in the cloud instance consistent with the on-premises environment to enable a seamless switchover.

WHAT IS A **LIVEDATA STRATEGY?**

A LiveData strategy is both an offensive and defensive approach to data management that ensures consistent, accessible, accurate, and available data across all sites, centers, and even multi-cloud architectures. A LiveData strategy ensures that no data is lost, and that recovery from an outage is instantaneous.

WHAT IS **WANDISCO FUSION?**

WANdisco Fusion is the only platform on the market today that can enable a LiveData strategy over a standard internet connection. Built on a unique and high-performance coordination engine called DConE, WANdisco Fusion can replicate unstructured data in any mixed IT environment without any downtime for production systems, even at very large scale.

CLOUD MIGRATION **WITHOUT** A LIVEDATA STRATEGY INCREASES COST AND RISK

Businesses that attempt to migrate their data lake to the public cloud without a way to run their on-premises and cloud platforms in parallel will be at an immediate disadvantage. This approach significantly extends the testing and validation process, driving up costs substantially. Most importantly, businesses without a LiveData strategy will be unable to achieve a seamless cutover to the cloud—increasing the risk of business disruption.

CLOUD MIGRATION **WITH** A LIVEDATA STRATEGY ENABLES A SEAMLESS TRANSITION

Enterprises that can ensure consistent data across their on-premises and cloud platforms will be able to drive both systems side-by-side with ease—dramatically accelerating the testing cycle. Because the same, current data is available in both platforms at all times, businesses can switch over operations to the cloud in a way that is completely transparent to their users.

HOW CAN WANDISCO HELP?

There are few analytics processes that be switched overnight from an on-premises to public cloud platform without significant risk for the business. Enterprises can mitigate these risks with a period of testing in which the on-premises data lake runs side-by-side with the target environment in the cloud.

Migrating a data lake to the cloud, and keeping massive volumes of data consistent with the on-premises platform, presents complex technical challenges. Without a way to continuously replicate their data, businesses will create data consistency problems that ratchet up the risk of disruption and downtime.

Delivering a cloud migration project with zero disruption might seem like an impossible task—but it's not.

WANdisco Fusion is the only platform on the market today that enables businesses to ensure data consistency across on-premises, hybrid-cloud and even multi-cloud environments over an internet connection. Powered by a revolutionary consensus engine called DConE, WANdisco Fusion replicates data continuously between all platforms: a key enabler for data lake migration projects. Crucially, WANdisco Fusion supports data lake offerings from the leading cloud vendors, including AWS EMR, Microsoft Azure Data Lake Storage and Microsoft HDInsight.

As well as maintaining consistency between multiple data platforms, WANdisco Fusion can restore data sources to consistency if they go offline—a capability that is particularly valuable for cloud migration projects. For example, WANdisco Fusion enables businesses to make a copy of their data lake to a data transfer device such as Azure Data Box without any downtime for the production cluster. When this data is transferred to the cloud, WANdisco Fusion automatically restores it to consistency it with the production environment, and then keeps the two environments consistent thereafter.

Using this capability, businesses can run their cloud and on-premises environments in parallel for as long as required, and switch seamlessly to the cloud as soon as they have validated their new data lake. With the guarantee of consistent data, available everywhere, enterprises can extend their testing periods at minimal incremental cost—enabling them to identify potential switchover challenges and break down problems into manageable parts.

If you're ready to migrate your data lake to the cloud, WANdisco is ready to help. To get started today, click here to arrange a callback.

WHAT ABOUT LEGACY PLATFORMS THAT CAN'T MOVE TO CLOUD?

Established enterprises—particularly those in heavily regulated industries—may be unable to decommission all their legacy systems and migrate them to the public cloud. However, this doesn't mean that many of the benefits of cloud are out of reach. With WANdisco Fusion, businesses can maintain a consistent copy of data on-premises and in their cloud data lake. As a result, these companies can continue to operate their legacy platforms for compliance purposes while applying cutting-edge analytics tools in the cloud to extract greater business value from the business data.

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