



The World Leader in
Active Data Replication™

A large, complex network diagram composed of numerous orange dots (nodes) connected by thin orange lines (edges). The nodes are arranged in a roughly horizontal line, with many connections between them, creating a dense web of relationships. The overall shape is somewhat irregular, with some nodes having more connections than others.

Hadoop migration on-premises and to cloud

Why an active data
migration tool is required



Contents

Executive summary	1
Selecting the right tools and processes	2
The three phases of migration	3
Post-migration	4
Conclusion	5

Executive summary

There are numerous business and technical benefits to be gained by migrating from one Hadoop distribution to another, whether on-premises or into the cloud. These include:

- Improved functionality and performance offered by a different Hadoop distribution, or an updated version of the same distribution, which in effect becomes a migration if the underlying Hadoop file system format has changed
- Lower support costs offered by competing Hadoop distribution vendors
- Enterprise-wide consolidation on a single Hadoop distribution
- Economies of scale offered by cloud-based options offering Hadoop as a service (HaaS) with a vast array of analytics applications and processing capacity that would be impossible to deploy and maintain in-house.

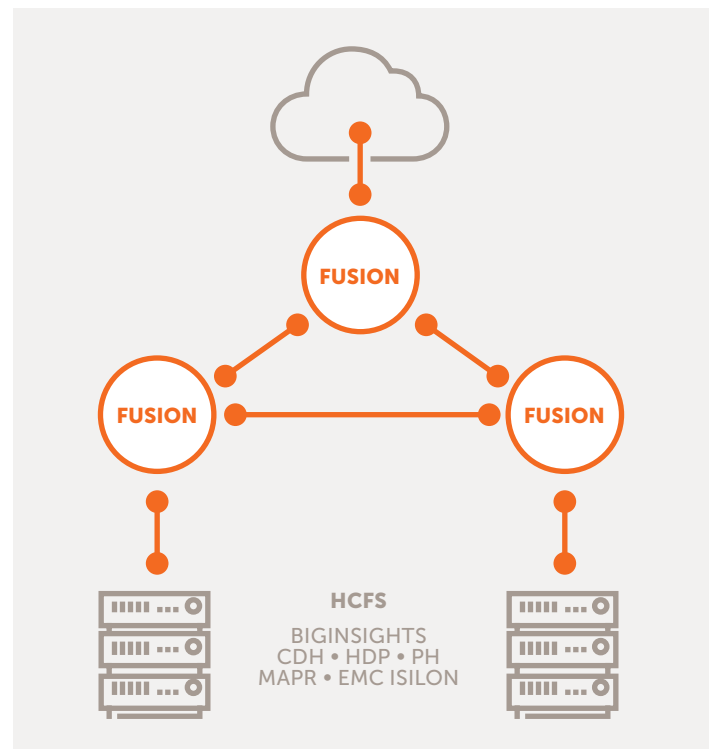
Unfortunately, there are major obstacles to obtaining these benefits, including the extended downtime and business disruption caused by the inherent limitations of the one-way batch-oriented tools typically used. These solutions require data to remain static in both the source and target clusters during migration - as a result, up to a week of downtime isn't unusual. Because of this, the risk of unrecoverable data loss is significant and there is no way to know with any certainty if all the data migrated successfully, or whether applications will perform and function as expected in the new environment until migration is complete.

This is unacceptable in virtually any production environment, and in the context of cloud environments means that typically only cold, static data can be migrated.

However, given the compelling business and technical benefits migration promises, how can these obstacles be overcome and the risks mitigated?

This white paper answers the question by exploring the challenges surrounding Hadoop migration both on-premises and to the cloud and takes the position that:

- The tool used for migration is the most critical factor in avoiding downtime and business disruption
- In order to avoid migration downtime, the tool must be active and multi-directional to enable old and new clusters to operate in parallel while data moves between them as it changes
- A comprehensive migration plan is critical regardless of the tool used, to ensure that organizational goals are met.



Selecting the right tools and processes

Migration challenges – traditional tools and processes

Hadoop migration on-premises

Hadoop migration projects most often rely on DistCp, the unidirectional batch replication utility built into Hadoop. DistCp is at the heart of the backup and recovery solutions offered by the Hadoop distribution vendors. It's the tool they and their systems integrator partners most frequently rely on to deliver migration services, its limitations are at the root of the downtime and disruption migration projects face. With DistCp, significant administrator involvement is required for setup, maintenance and monitoring. Replication takes place at pre-scheduled intervals in what is essentially a script-driven batch mode of operation that doesn't guarantee data consistency. Any changes made to source cluster data while the DistCp migration process is running will be missed and must be manually identified and moved to the new target cluster.

In addition, DistCp is ultimately built on MapReduce and competes for the same MapReduce resources production clusters use for other applications, severely impacting their performance. These drawbacks require production clusters to be offline during migration, and they're the same reasons cluster backups using DistCp during normal operation must be done outside of regular business hours. This risks data loss from any network or server outages after the last after hours backup.

Another migration technique is to physically transport hard-drives between old and new clusters. In addition to downtime and limited resource utilization during migration, there are other challenges with this approach:

- If the underlying Hadoop file system format is different between the source and target clusters, custom software development may be required to support complex data transformation requirements. Data loss often results from incorrectly translating, overwriting, or deleting data
- Even a small Hadoop data node server will have at least 10 physical disks. In a cluster of any size, it's almost inevitable that one or more may be lost or damaged in transit.

Hadoop to cloud migration

Hadoop distribution vendors have also added support to their DistCp solutions for moving data to the cloud, but the same challenges faced with on-premises Hadoop migration remain. For large-scale data migration, some cloud vendors offer an appliance-based approach. Typically a storage appliance is delivered to the customer's data center and data is copied from the customer's servers to the appliance. The appliance is then shipped back to the cloud vendor for transfer to their servers to complete the process, which can take over a week. While this may be suitable for archiving cold, less critical data to the cloud, it doesn't address migration of on-premises data that continues to change. In addition, such an approach doesn't address elastic data center, or hybrid cloud use cases for on-demand burst-out processing in which data has to move in and out of the cloud continuously. This also doesn't meet requirements for offsite disaster recovery with the lowest possible RTO (recovery time objective) to get back up and running after a network or server outage, nor does it enable the lowest possible RPO (recovery point objective) to minimize potential data loss from unplanned downtime. In many industries, both are mandated by regulatory as well as business requirements to be a matter of minutes.



Overcoming migration challenges

The only way to avoid migration downtime and disruption is to use a tool that allows existing and new clusters to operate in parallel. This kind of migration experience can only be achieved with a true active replication solution capable of moving data as it changes in both the old and new clusters, whether on-premises or in the cloud, with guaranteed consistency and minimal performance overhead.

With an active migration tool, applications can be tested to validate performance and functionality in both the old and new environments while they operate side-by-side. Data, applications and users move in phases and the old and new environments share data until the migration process is complete. Problems can be detected when they occur, rather than after a period of downtime when they may be impossible to resolve without restarting the entire migration process, extending downtime even further.

In addition, the tool must be agnostic to the underlying Hadoop distribution and version, the storage it runs on, and in the case of cloud migration, the cloud vendor's object storage. The migration tool should also be capable of handling data movement between any number of clusters if the goal is consolidation onto a single big data platform, whether on premises or in the cloud. WANdisco Fusion is such a solution.

WANdisco Fusion overcomes migration challenges by:

- Eliminating migration downtime and disruption with patented one-way to N-way active data replication that captures every change, guaranteeing data consistency and enabling old and new clusters to operate in parallel. Fusion delivers this active data replication across clusters deployed on any storage that supports the Hadoop-Compatible File system (HCFS) API, local and NFS mounted file systems running on NetApp, EMC Isilon, or any Linux-based servers, as well as cloud object storage systems such as Amazon S3.

This eliminates many restrictions that would otherwise apply during migration.

- Simplifying consolidation of multiple clusters running on any mix of distributions, versions and storage onto a single platform. Clusters and data in the new post-migration environment can automatically be distributed in any configuration required both on-premises and in the cloud. This makes it easy to bring new data centers online, or retire existing ones as part of a migration project
- Allowing administrators to define replication policies that control what data is replicated between clusters and selectively exclude data from migration to specific clusters in the new environment, or move it off to be archived
- Providing forward recovery capabilities that allow migration to continue from where it left off in the event of any network or server outages.

The three phases of migration: strategy, planning, and execution

Even with the best technologies, a clear strategy supported by a comprehensive migration plan is required to ensure that organizational goals are met.

Strategy

The first stage is to define a strategy that outlines:

- Organizational goals and objectives based on the priorities and expectations including those of your development, operations and end-user organizations, both pre- and post-migration
- The scope of the migration effort. WANdisco Fusion can support projects that require moving data across any number of clusters running on a variety of distributions, file systems and cloud storage environments simultaneously with no disruption. This allows projects with a much broader scope to be completed in a far shorter timeframe than migrating a single active cluster from one Hadoop distribution to another
- A clear description of the expected benefits and acceptance criteria for your migration project

- A complete list of risks and their impact on the organization (e.g., an estimate of the cost of any downtime)
- Well-defined roles and responsibilities for migration tasks and deliverables.

Planning

- Clearly define the order and timing of each task during the execution phase.
- Produce a detailed test-plan that reflects the acceptance criteria defined with stakeholders during the strategy phase.

Execution

1 Establish the new environment

Step one is to establish the new environment, and validate its correct implementation before moving data to it. New clusters can be used for DR (disaster recovery) during migration and old clusters can be used for DR post-migration. However, with WANdisco Fusion's patented active replication technology, your old clusters can support much more than DR. With WANdisco Fusion, all clusters are fully active, read-write at local network speed everywhere, continually synchronized as changes are made on any cluster, with automatic recovery from each other after an outage.

2 Migrate and test

WANdisco Fusion allows data transfer to take place while operations in both the old and new clusters continue as normal. You can test applications and compare results in both the old and new environments in parallel and validate that data has moved correctly and applications perform and function as expected. WANdisco Fusion can also replicate data selectively to control which directories go where. Data not needed post-migration can be moved off for archiving.

If network or server outages occur during migration, WANdisco Fusion has built-in forward recovery features that enable migration to automatically continue from where it left off without administrators having to do anything.

3 Adopt the new environment

This must incorporate feedback on the actual results set against the acceptance criteria defined in your migration plan. If you have hardware and other infrastructure from the old environment available post-migration you can implement WANdisco Fusion in your new environment and take advantage of features that were huge benefits during migration.

Benefits of WANdisco Fusion post-migration

Post-migration WANdisco Fusion enables:

Continuous availability with guaranteed data consistency

WANdisco Fusion guarantees continuous availability and consistency with patented active replication for the lowest possible RTO and RPO across any number of clusters any distance apart, whether on-premises or in the cloud. Your data is available when and where you need it. You can lose a node, a cluster, or an entire data center, and know that all of your data is still available for immediate recovery and use. When your servers come back online, WANdisco Fusion automatically resynchronizes your clusters after a planned or unplanned outage as quickly as your bandwidth allows.

Full use of cluster resources

WANdisco Fusion eliminates read-only backup servers by making every cluster fully writable as well as readable and capable of sharing data and running applications regardless of location, turning the costly overhead of dual environments during migration into productive assets. As a result, otherwise idle hardware and other infrastructure becomes fully productive, making it possible to scale up your Hadoop deployment without any additional infrastructure.



Selective replication on a per folder basis

Fusion allows administrators to define replication policies that control what data is replicated between Hadoop clusters, on-premises file systems and cloud storage. This enables global organizations to only replicate what's required, and keep sensitive data where it belongs to meet business and regulatory requirements.

Minimal data security risks

In addition to working with all of the available on-disk and network encryption technologies available for Hadoop, WANdisco Fusion only requires the Fusion servers to be exposed through the firewall for replication between data centers. This dramatically reduces the attack surface available to hackers. In contrast, DistCp solutions require every data node in every cluster to be able to talk to every other through the firewall. This creates an untenable level of exposure as well as an unreasonable burden on network security administrators as cluster size grows.

Active hybrid cloud

With WANdisco Fusion, data arrives in the cloud environment as it changes on-premises and vice versa, so the cloud vendor storage and applications can be leveraged seamlessly, regardless of the Hadoop distribution running on-premises. In addition, it's far more efficient to use the cloud provider's network and servers to scan and process large volumes of data that can be collected and distilled down to provide a smaller data set that's more meaningful for analysis, than to burden in-house resources with these tasks. WANdisco Fusion can then replicate this analytics-ready data to on-premises Hadoop clusters for further analysis. In addition, because WANdisco Fusion can replicate across any HCFS compatible storage, Big Insights on Cloud can be used with on-premises clusters deployed on BigInsights, Cloudera, Hortonworks, MapR, Oracle BDA, Pivotal, or any other Hadoop distribution or storage that supports the HCFS API.

Conclusion

Hadoop migration projects and the tools that support them need to account for a wide variety of requirements.

In summary, with WANdisco Fusion's unique patented active replication technology you have the ability to:

- Operate both old and new clusters in parallel, without stopping operation in the old cluster either during, or after migration
- Make data produced in your new production cluster available in the old cluster infrastructure as part of a DR strategy
- Test applications in parallel in the old and new environments to validate functionality and performance
- Phase your migration of data, applications and users
- Consolidate multiple clusters in a distributed environment running on a mix of Hadoop distributions and storage onto a single on-premises or cloud platform, with the data distributed and synchronized in any manner your organization requires
- Eliminate the need to restrict your production environment to a single cluster, or only those resources and applications available in-house. Both old and new, or a combination of multiple clusters and cloud environments, can be operational and work on the same underlying content on an opt-in basis using WANdisco Fusion's patented active replication capability after migration.

For more information, please visit

wandisco.com

Notes



Talk to one of our specialists today
US TOLL FREE +1-877-WANDISCO (926-3472)
EMEA +44 114 3039985
APAC +61 2 8211 0620
ALL OTHER +1925 380 1728


5000 Executive Parkway, Suite 270
San Ramon, California 94583

Copyright © 2016 WANDISCO, Inc. All rights reserved.

wandisco.com

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

Follow us to stay in touch

 [@WANDISCO](https://twitter.com/WANDISCO)