Hadoop Migration Made Simple:

A SINGLE APPROACH TO CLOUD, ON-PREMISE AND MULTI-VENDOR MIGRATIONS

By Steve Jones, Capgemini Global VP, Big Data and Analytics

Hadoop Migration Made Simple: A Single Approach To Cloud, On-Premise and Multi-Vendor Migrations

1. Executive Summary

Many firms are facing the challenge of transitioning departmental and niche Big Data programs into the information fabric of their enterprise. This shift often involves revisiting previous decisions regarding vendors and approach. It also requires migration and consolidation to be core competencies. Without them, any strategy will be based on what is currently available and not what's needed for the future.

To be successful a migration or consolidation needs to be able to overcome key hurdles including:

- Downtime during migration.
- Consistency of security models.
- New environment verification before a switch over.

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

- Business consistency which helps drive greater degrees of collaboration.
- Consolidated investments across multiple business areas.
- Improved functionality and performance offered by a different Hadoop distribution, or an updated version of the same distribution, which effectively becomes a migration if the underlying Hadoop file system format changes between releases.
- Lower support costs offered by competing Hadoop distribution vendors.
- Consolidation on a single Hadoop distribution or Hadoop-as-a-Service (HaaS) cloud platform.
- Economies of scale offered by cloud-based storage and processing, with access to a wide range of analytics applications and other services that would be virtually impossible to deploy and maintain in-house.

This white paper provides a systematic approach to Hadoop migration, both on-premise and to cloud and shows:

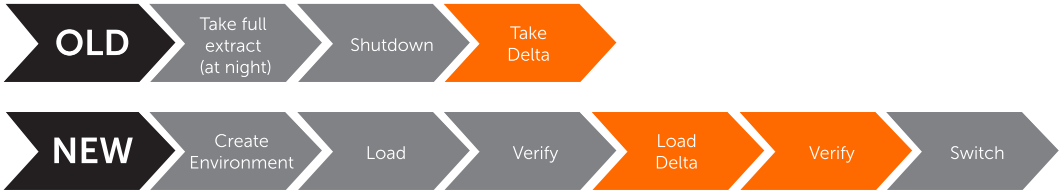
- The tool used for migration is the key to avoiding downtime and business disruption.
- In order to avoid migration downtime, the tool must be transactional and multi-directional, allowing a phased migration that enables old and new clusters to operate in parallel while data moves between them as it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

When looking at the shift to Big Data as a business platform for insight and its impact on current efforts it is essential to be able to answer the key question "how do I get from what I have now to what I need in the future?" Active migration is a central part of answering that question.

2. Selecting the Right Tools and Processes

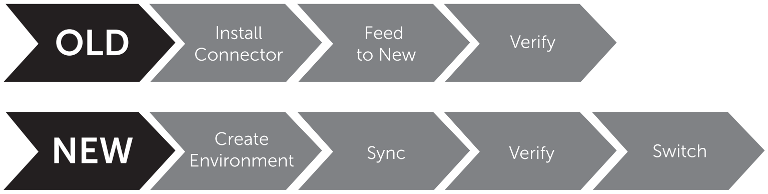
Passive or Active?

When it comes to migration there are two broad choices: passive or active. Passive migration is what Data Warehouse people will be comfortable with, and what Hadoop vendors provide out of the box. It essentially involves taking a large extract at a given point in time, loading it into the new environment, shutting off the old environment, loading any additional data, and then turning on the new environment.



This means downtime. It also normally means sticking with one vendor as their tools are designed to move data from one version of their platform to another, not to migrate away.

Active migration enables you to continually synchronize between the old and new environments. Both are live and operate in parallel during migration. Data, applications and users move in phases. Complete transition to the new environment doesn't take place until it's proven to be identical and defined acceptance criteria are met. This approach eliminates downtime and enables you to shift vendors rather than stick with only one.



Migration Challenges – Why Old School Doesn't Work in the New World

"In the box" Hadoop Migration On-Premise

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, and 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 necessarily introduces the risk of data loss from any network or server outages occurring since 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-premise 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 often takes more than a week. While this may be suitable for archiving cold, less critical data to the cloud, it doesn't address migration of on-premise 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 transactional replication solution capable of moving data as it changes in both the old and new clusters, whether on-premise or in the cloud, with guaranteed consistency and minimal performance overhead.

With an active transactional 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.

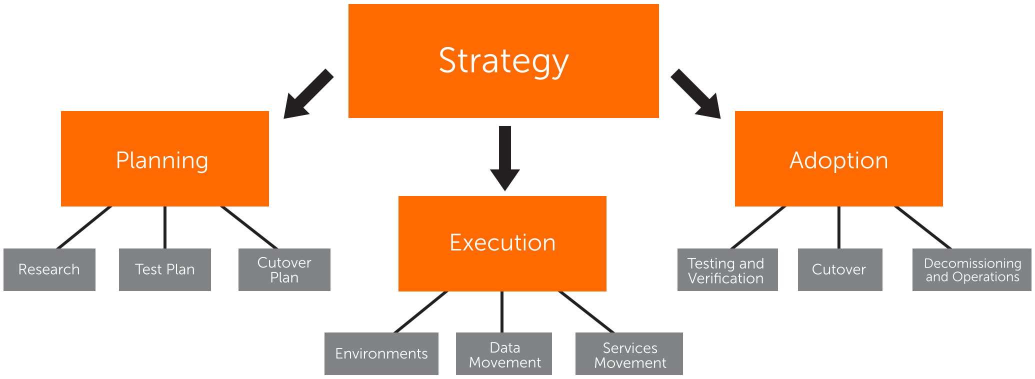
In addition, the tool must be agnostic to the underlying Hadoop distribution and version, the storage it runs on, and in the case of cloud migration, the cloud vendor's object storage. The migration tool should also be capable of handling data movement between any number of clusters if the goal is consolidation onto a single big data platform, whether on premise 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 transactional data replication that captures every change, guaranteeing data consistency and enabling old and new clusters to operate in parallel. Fusion delivers this active transactional 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-premise 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.

3. The Four Phases of Migration: Strategy, Planning, Execution and Cutover

Even with the best technologies, a clear strategy supported by a comprehensive migration plan is required to ensure that organizational goals are met. This is the case regardless of whether you're migrating from one on-premise cluster to another, or planning a more complex consolidation project across multiple data centers behind the firewall and in the cloud.



Strategy

The first stage is to define a strategy that outlines:

- Organizational goals and objectives based on the priorities and expectations 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 without disruption. This allows projects with much broader scope than migrating a single active cluster from one Hadoop distribution to another to be completed in a much shorter timeframe.
- 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 with a detailed project plan that includes estimates, dependencies, roles and responsibilities.
- Produce a detailed test-plan that reflects the acceptance criteria defined with stakeholders during the strategy phase. You must understand their priorities and expectations, both pre- and post-migration. This research will also help gather the information required to define an adequate test plan.

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- transactional 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 either cluster, and recover automatically from each other after an outage.

2: Migrate Data

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: Migrate Applications

A migration that doesn't include moving the analytical services and other applications that run on your existing Hadoop platform doesn't result in a successful migration. To facilitate this, it's crucial that services are moved to 'shadow running' often in a headless mode where they are disconnected from enterprise systems, but where functionality and performance can be tested between the old and new environments. This migration will often require applications to be modified in some way to take advantage of, or remove reliance on a particular vendor's tool set.

At the end of execution the company has two active Hadoop clusters which should be data, functional and analytically equivalent. At this point, the foundation for adoption of the new environment is in place.

Adoption

1: Testing and Verification

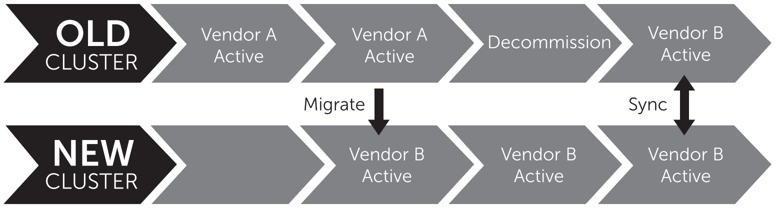
This stage measures the new environment against the set of acceptance criteria defined in your migration plan. Using approaches such as mRapid and LEAP from Capgemini this functional and analytical equivalence, also known as outcome equivalence, can be automated. These tools and solutions enable the automated testing of reports, HIVE and other Hadoop based data technologies as well as more complex analytical models such as R. By automating the process and not relying on visual confirmation a business is not only able to more rapidly and accurately verify outcome equivalence but also do so at a much lower cost base than human driven approaches.

2: Cutover

Cutover can be handled in a few ways depending on the business requirements and plan, but thanks to WANdisco's active replication technology a "bleed" of services between environments can be done with the active services tag being transitioned from the old environment to the new rather than needing a 'big bang' approach where everything moves in a single bound. By enabling this approach it becomes possible to performance test and verify the cutover, and in Cloud based environments to actively scale the cluster as new services switch from shadow running to taking enterprise load. At the end of cutover all services are transitioned and no enterprise or business functions are relying on the old cluster.

3: Decommissioning and Operations

The final stage is the transition towards standard operations for the new cluster and the decommissioning of the old cluster. This is crucial to plan as only when decommissioned can cost benefits be realized. At this stage the active migration approach comes with additional benefits. For instance the active migration approach can be now used in reverse to re-use the hardware for the old cluster as either a DR or 'dual active' environment running on a single vendor.



4. Post Migration

Post-migration WANdisco Fusion enables:

- **Continuous availability with guaranteed data consistency**
  WANdisco Fusion guarantees continuous availability and consistency with patented active-transactional replication for the lowest possible RTO and RPO across any number of clusters any distance apart, whether on-premise 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.
- **100% 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-premise 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 on-premise data centers, and to the cloud. 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, creating an untenable level of exposure as well as an unreasonable burden on network security administrators as cluster size grows.
- **Active-transactional hybrid cloud**
  The same unique capabilities that support parallel operation of on-premise and cloud environments during migration enable Fusion to support true public-private hybrid cloud deployments post-migration. Fusion transfers data as it changes between cloud environments such as Amazon S3 and on-premise Hadoop clusters, on-premise linux-based NFS and local file systems, or other cloud storage with guaranteed consistency.

  An active EMR cluster is not required to move data between S3 and on-premise Hadoop environments. However, Fusion supports EMR's ability to dynamically spin up and down on demand for hybrid burst-to-cloud support. This is ideal for hybrid cloud projects that require additional computing power without the hassle and expense of bringing additional server capacity and staff in-house.

5. Conclusion

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

In summary, with an active approach to migration 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 distributions and storage onto a single on-premise or cloud platform distributed in any manner your organization requires.
- Eliminate the need to restrict your production environment to a single cluster. Both old and new, or a combination of multiple clusters and cloud storage environments, can be operational and work on the same underlying content on an opt-in basis.

All of this is enabled by WANdisco Fusion's unique patented active-transactional replication technology.

For more information please visit: www.wandisco.com/hadoop/wd-fusion and www.capgemini.com/insights-data

wanDISCO®

5000 Executive Parkway
Suite 270
San Ramon, California
94583

US Toll Free: +1-877-WANDISCO (926-3472)
EMEA: +44 114 3039985
APAC: +61 2 8211 0620
All Other: +1925 380 1728