Azure Data Lake Storage Gen1 to Gen2 Migration

Azure Data Lake Storage (ADLS) Gen2 is a highly scalable and cost-effective data lake solution for big data analytics. It combines the power of a high-performance file system with massive scale and economy to help organizations speed their time to insight. ADLS Gen2 extends Azure Blob Storage capabilities, is optimized for analytic workloads, and is the most comprehensive data lake available.

As more customers migrate from ADLS Gen1 to Gen2 they typically follow one of four migration approaches. These approaches are described in this document, and the final section provides information on WANdisco LiveMigrator, which minimizes the risks and costs associated with large scale data migration initiatives, and is an ideal and Microsoft recommended solution for migrating data from ADLS Gen1 to Gen2 with zero downtime during migration, zero data loss and 100% data consistency.

MIGRATION APPROACHES

Migration from ADLS Gen1 to Gen2 typically follows one of four migration patterns, which are described in more detail below. The patterns are also discussed in the Microsoft documentation at: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-migrate-gen1-to-gen2#migration-patterns

Lift and Shift

A lift and shift approach migrates an application and data from one environment to another without redesigning the application for the target environment. A lift and shift approach is typically the simplest approach requiring the following high level steps:

- Stop all writes to Gen1
- Move the data from Gen1 to Gen2
- Point ingest operations and workloads to Gen2
- Decommission Gen1

Typically, this approach is best suited for small scale migrations, where all applications can be upgraded to the new environment at one time, and for which downtime is acceptable. Once organizations need to migrate 100s of TBs or PBs of data, the amount of time required just to physically move the data is usually longer than the acceptable downtime that is required. Additionally, while upgrading all applications at one time can be a pro, many organizations like to phase the migration in order to minimize risk. This phasing is not possible with a big bang lift and shift approach.

<table>
<thead>
<tr>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplest approach</td>
<td>Requires downtime during migration and cutover periods</td>
</tr>
<tr>
<td>All applications upgraded at one time</td>
<td>All applications upgraded at one time</td>
</tr>
</tbody>
</table>
INCREMENTAL COPY

An incremental copy approach is where the new and modified data is periodically copied from the source to target destination. To execute the incremental copy approach requires that the destination must have all data from the source system before the incremental copy process can be initiated. Steps for this approach are as follows:

- Start moving data from Gen1 to Gen2
- Incremental copy of new and modified data from Gen1 to Gen2
- Once incremental copy is complete, stop all writes to Gen1 and point workloads to Gen2
- Decommission Gen1

An incremental copy approach is typically used when needing to migrate larger data sets and the copy requires more time. Since it allows writes to continue in the Gen1 environment it does not require as much application downtime. However, just as was the case for lift and shift, once organizations need to migrate 100s of TBs or PBs of data, the incremental copy approach is likely also not acceptable. The new and modified data in Gen1 must continuously be reconciled and incrementally copied to the Gen2 environment. Manual reconciliation becomes unacceptable for large scale data sets, and the incremental copy process may take too long to complete. In addition, just as for lift and shift, all applications must be upgraded at one time which may not be acceptable for many organizations.

DUAL PIPELINE / INGEST

A dual pipeline or dual ingest approach is where new data is ingested simultaneously into both the Gen1 and Gen2 environments. Steps for this approach are as follows:

- Start moving data from Gen1 to Gen2
- Ingest new data into both Gen1 and Gen2
- Point workloads to Gen2
- Stop all writes to Gen1 and then decommission Gen1

While a dual ingest approach can support a zero downtime migration, and allow for a phased cutover of applications, it introduces much higher complexity and requires many more resources to manage this complexity during the setup, maintenance, testing and validation activities. Once the dual ingest is started in both environments, reconciliation needs to be continuously performed to identify data changes that occur in Gen1 and make sure those same changes get applied to Gen2. As discussed previously, manual reconciliation may not be feasible or acceptable for large scale data sets. The longer that changes continue in Gen1 the greater the chance of introducing data inconsistency, and given this approach is typically used for migration of large data sets where downtimes introduced by the previous patterns would not be acceptable, the amount of time this approach requires before it is completed can be very lengthy. The migration projects often exceed expected timelines and budgets.

PROS

- Requires less downtime than lift and shift approach
- All applications upgraded at one time

CONS

- Requires downtime during cutover period
- All applications upgraded at one time
- Requires reconciliation to identify new & changed data
- Lengthy process for large scale migrations
BIDIRECTIONAL SYNCHRONIZATION

A bidirectional synchronization approach is needed when downtime is not acceptable, and for large scale data sets that are undergoing active change. Steps for this approach are as follows:

- Set up bidirectional replication between Gen1 to Gen2. Microsoft recommends the use of WANdisco LiveMigrator
- When all moves are complete, stop all writes to Gen1 and turn off bidirectional replication
- Decommission Gen1

A bidirectional synchronization approach using WANdisco LiveMigrator is ideal for complex scenarios that may involve a large number of pipelines, where downtime is not acceptable, and where organizations want to minimize the risks and costs of their ADLS Gen1 to Gen2 migrations. The following section provides more details on the WANdisco LiveMigrator solution.

PROS

Assuming use of WANdisco LiveMigrator

- Supports zero downtime migration
- Ensures 100% data consistency and zero data loss
- IT efficiency, requiring fewer resources to conduct the migration
- Shorter migration; faster time to value
- Allows phased migration of applications

CONS

- High complexity if custom developed and not using WANdisco
WANdisco Livedata Migrator

WANdisco LiveMigrator allows organizations to migrate data in a single pass while keeping up with all changes occurring in the source (ADLS Gen1) system. This results in guaranteed consistency between source (ADLS Gen1) and target (ADLS Gen2) at the completion of the migration.

LiveMigrator uses a single scan iterator which visits every source file once and only once, replicating data from ADLS Gen1 to ADLS Gen2. This ensures ADLS Gen2 has a complete copy of the data. If data is changed, which the iterator hasn’t seen yet then the change will simply be picked up when the iterator reaches that content and replicated to the target environment. If changes occur in an area that has already been scanned by the iterator then LiveMigrator transfers the change as soon as it occurs. The result is a completely consistent data copy at the end of migration. Users can manually trigger a migration whenever you are ready to begin.

LiveMigrator provides an easy to use user interface from where users can perform configuration, define replication rules, monitor usage, and much more.

![WANdisco Fusion User Interface](image)

Create replication rules

Start migration and perform consistency checks
WHY LIVEMIGRATOR IS DIFFERENT?

- **Hadoop & Object Storage**: Ideal for ADLS Gen1 to ADLS Gen2 migration. Also works across a variety of other big data source and target environments, including all major Hadoop and object storage technologies.

- **Petabyte Scale**: Migrates big data sets at any scale to cloud storage without needing to halt changes made to the data sets during migration.

- **Selective Migration**: Allows selection of which data sets should be migrated; administrators define replication policies that control what data is replicated between environments.

- **One Pass**: Migrates existing data sets with a single pass through the source (ADLS Gen1) storage system, eliminating the overhead of repeated scans.

- **Consensus-Based Consistency**: Incorporates changes made to the data sets under migration with a strongly consistent distributed consensus engine, so that applications can continue to modify the source system’s data without causing divergence between source and target.

- **Bandwidth Management**: Optimizes bandwidth use by eliminating the need for repeated transfer of data, and enforces limits on bandwidth use.

- **Automatic Outage Recovery**: The consensus engine used by LiveMigrator eliminates the need for manual response to system failures, including network outages and other disruptions.

- **Rapid Availability**: Minimizes the time required to develop new workloads by making each data location available for use as soon as bandwidth allows.

LIVEMIGRATOR BUSINESS BENEFITS

**Business Continuity**

- Zero downtime during migration
- High Scalability and better performance with Big Data Sizes (100’s TB - Multi PB)
- Immediate availability of migrated data
- Unaffected by outages

**Ensures Data Consistency**

- As changes occur anywhere in the source system, LiveMigrator creates and ensures target system data consistency.

**Cost Avoidance/IT efficiency**

- No code maintenance
- One Click automated replication across all major cloud storage, commercial Hadoop distributions, and analytic services
- One Pass of the source storage lowers project cost, faster project completion, faster adoption of ML and AI
- Multi-Cloud (Bi-Directional) Data Replication compatible with All Cloud Vendors
- No "Big Bang" cutover for applications

About WANdisco

With zero downtime and zero data loss, WANdisco LiveMigrator solves the problem of moving petabyte scale data to the cloud and LiveAnalytics Solution provides immediate analytic data access through continuous automated replication from on-premises Hadoop analytics to Spark based cloud analytics. WANdisco Fusion keeps geographically dispersed data at any scale consistent between on-premises and cloud environments allowing businesses to operate seamlessly in a hybrid or multi-cloud environment. WANdisco has over a hundred customers and significant go-to-market partnerships with Microsoft Azure, Amazon Web Services, Google Cloud, Oracle, and others as well as OEM relationships with IBM and Alibaba.

Talk to one of our specialists today

<table>
<thead>
<tr>
<th>Region</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>+1 877 WANdisco (926-3472)</td>
</tr>
<tr>
<td>EMEA</td>
<td>+44 (0) 114 3039985</td>
</tr>
<tr>
<td>APAC</td>
<td>+61 2 8211 0620</td>
</tr>
<tr>
<td>All other</td>
<td>+1 925 380 1728</td>
</tr>
</tbody>
</table>

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

Follow us to stay in touch