

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 LiveData Plane, which minimizes the risks and costs associated with large scale data migration initiatives, and is an ideal and Microsoft recommended solution for bidirectional replication and 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.

PROS

- Simplest approach
- All applications upgraded at one time

CONS

- Requires downtime during migration and cutover periods
- All applications upgraded at one time



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.

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

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

- Supports zero downtime
- Allows phased migration of applications

CONS

- High complexity solution
- Requires more resources to manage the setup, maintenance and testing activities
- Requires reconciliation to identify data changes in Gen1 from initial copy and while dual ingest is active
- Higher potential for data inconsistency
- 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 LiveData technology
- When all moves are complete, stop all writes to Gen1 and turn off bidirectional replication
- Decommission Gen1

A bidirectional synchronization approach using WANdisco LiveData Plane 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.

PROS

Assuming use of WANdisco LiveData technology

- 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

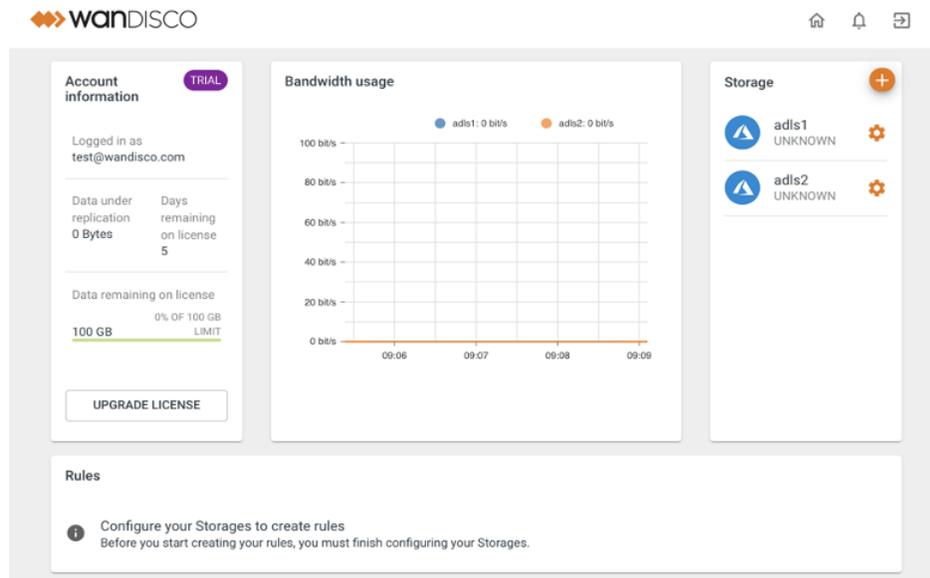
- High complexity if custom developed and not using WANdisco



WANDISCO LIVEDATA PLANE

Keeping data consistent in a distributed environment is a massive challenge. WANDisco LiveData Platform solves the exponentially growing challenge of keeping unstructured data available across diverse IT environments regardless of geographic location or architecture. At the heart of LiveData Plane is WANDisco's patented Distributed Coordination Engine (DConE), which uses consensus to keep Hadoop and object store data accessible, accurate, and consistent in different locations across any environment. WANDisco LiveData Plane is a foundation for a modern cloud data strategy—a LiveData strategy—because it prevents data disasters, de-risks data migration, and ensures data consistency across multiple distributed environments.

Traditional approaches to data replication are batch-based, do not guarantee data consistency, and cannot operate over wide-area networks or the public internet. Unlike other technologies which move data from one location to another, WANDisco LiveData Plane uses DConE to coordinate distributed changes to data, enabling shared access to common data sets. The technology works by applying a mathematically-proven approach to consensus which works regardless of the distance between data sources or types of data stores.



WANDISCO User Interface

Create replication rules

Start migration and perform consistency checks



LIVEDATA PLANE CAPABILITIES

- **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 and replicates data sets at any scale
- **Selective Multi-directional Replication:** Allows selection of which data sets should be replicated, and allows active workloads to continue across all locations
- **Guaranteed Consistency:** Coordinates changes to data across multiple environments achieving consensus and consistency at scale
- **Security:** Compatibility with common security protocols—Kerberos, SSL/TLS, LDAP
- **Self-healing:** Eliminates the need for administrators to respond to system level failures with automated recovery, and near zero RPO/RTO
- **Interfaces and API:** Migration and replication can be managed through a comprehensive and intuitive command-line interface, or from the self-documenting REST API, allowing users to track and monitor progress
- **Browser-Based UI:** Users can also take advantage of the optional management available from WANdisco's browser-based user interface and deep integration with cloud vendors' management interfaces

LIVEDATA PLANE 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

- Patented coordination engine ensures 100% data consistency across multiple distributed environments

Cost Avoidance/IT efficiency

- No code maintenance
- Eliminates the need for administrators to respond to system level failures with automated recovery
- No "Big Bang" cutover for applications during data migrations

About WANdisco

WANdisco is the LiveData company. WANdisco solutions enable enterprises to create an environment where data is always available, accurate, and protected, creating a strong backbone for their IT infrastructure and a bedrock for running consistent, accurate machine learning applications. With zero downtime and zero data loss, WANdisco LiveData Cloud Services keep 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. For more information on WANdisco, visit www.wandisco.com.