

Choosing your data service in cloud for migrating datasets

Pritpal Singh Khokhar



Introduction

Migrating large data to cloud often is a challenge due to various factors. This process may become cumbersome as larger data migrations can include terabytes of information to be moved leading to network bandwidth limits or other constraints. The planning must be effective during large transformation projects to bring up your applications on time in the given window. This also must go along with the fact that the transformation does not disrupt the normal business operations and the end user experience. This whitepaper is designed to emphasize some of the common challenges in moving large enterprise data sets and how to take a pragmatic approach while moving data from on premise to the cloud. While having knowledge of cloud tool is essential it is also required to have some native tooling knowledge to move data between source and target platforms. The variety of storage options Cloud offers at an economical rate is infinite and comes with the assurance of security and persistent with availability.

In this whitepaper, we present some best practices that can be applied to various Data Migration scenarios to avoid some common pitfalls. The definition of Data Migration would be limited to transfer of data between the platform (source and target) storage devices. While the whitepaper covers most general scenarios for Data Migrations, a prior familiarity with Data Migration and ETL concepts would help the reader in understanding the ideas better.

Organizations today are switching platforms or upgrading them to new versions which brings in the need for data transfer.

Some of the typical migration patterns are as follows:

Rehost – Most organizations prefer this strategy to make initial migrations at scale and support the business case to move to cloud. This migration pattern covers use cases with Lift and Shift strategies for applications.

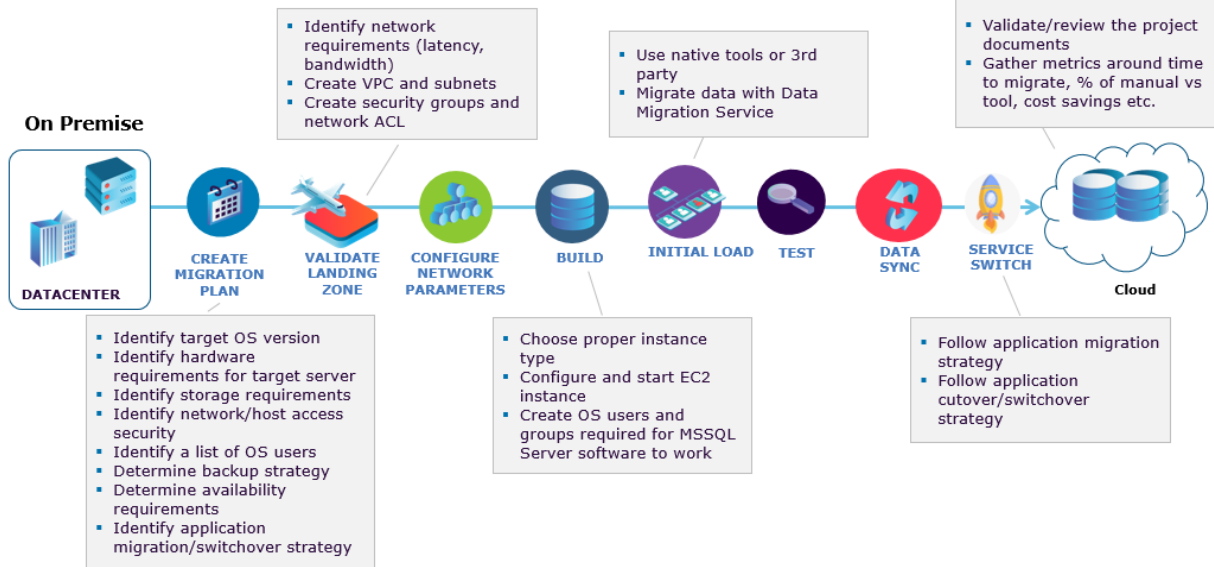
Redeploy – This strategy is followed by organizations to achieve some tangible benefits by making minor cloud optimizations without changing the overall architecture of the application. A typical use case for this pattern requires deploying application binaries on a new OS version with autoscaling.

Refactor or Replatform - Refactor describes running your applications (usually Web applications) on the cloud provider's infrastructure. The applications that are migration with this pattern requires moving to PaaS services of the hyperscaler.

Rearchitect or Redesign – In this pattern organizations make larger reimagination on the way the application should work using mainly cloud native features to achieve performance and scalability.

Retain – This pattern is applicable for the application candidates who continue to remain in the data centre or are shifting to some other base rather than cloud a nearby data centre or same location.

The below diagram describes the stages in a migration journey for a redeploy scenario of data migration:



It is important that organizations pay attention to making large scale dataset migrations in a secured manner when they are planning their Cloud move so that the required services can be identified.

Challenges in a successful large data migration journey include:

1. **Planning** – Inadequate planning during migration may prove disastrous if data migration is not planned. The identification and availability of its data transfer service in regions for the cloud provider where the application is being moved is imperative. Other factors include the amount of data to be moved, offline or online transfer etc.
2. **Architecture** – The migration plan must include the right architecture for the target cloud environment. Failure to envision the migration environment for any of the services might lead to stoppage in the migration process. This also includes the correct service identification process for the target cloud environment.
3. **Resource Skills** – If the team lacks the proper specialized technical skills to carry the migration and the data transfer, it would cause delay in the execution of the project.
4. **Data Security** – Security is one of the most important aspects while making data transfer. Any weak link in the security might lead to severe implications due to data loss. To protect all aspects of a data transfer workflow there is right combination of techniques and decisions required to be implemented for both data in transit and at rest. Encryption standards, Authentication and Authorization solutions are useful techniques to maintain data sovereignty while moving data or protecting data at rest.
5. **Leadership and Project Management** – Inefficiency in project management will bring failure to scope out the required budget properly and the schedule to deliver the project. The leadership team must also be prepared for a cultural change in the organization.

Assessing the Data Migration


Assessing the data migration patterns requires decision making on the following factors:

- Cost** – Data transfer costs, storage costs
- Time** – The time taken to migrate the data from one point to another and to check how much data can be moved in how much time, one can use the below AWS prescribed formula:

$$\text{Number of days} = \frac{\text{Total Bytes}}{(\text{Megabits per second} * 125 * 1000 * \text{Network Utilization} * 60 \text{ seconds} * 60 \text{ minutes} * 24 \text{ hours})}$$

For example, if you have a T1 connection (1.544Mbps) and 1TB (1024 * 1024 * 1024 * 1024 bytes) to move in or out of AWS the theoretical minimum time it would take to load over your network connection at 80% network utilization is 82 days.

Close Far

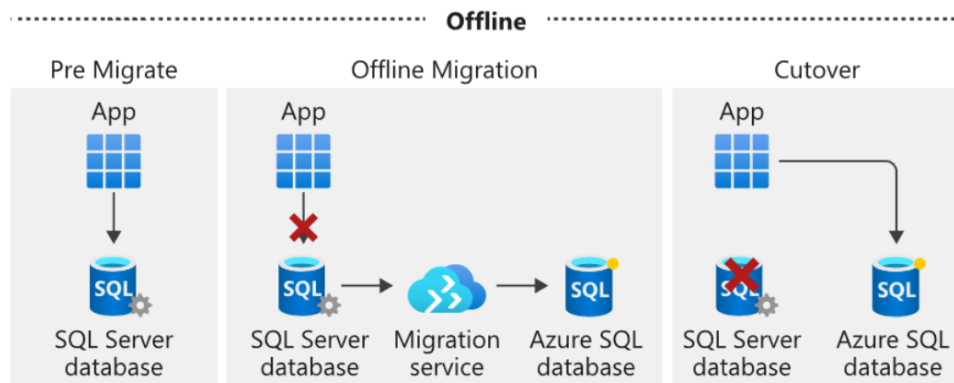


Data Size	100 PB	124 days	3 years	34 years	340 years	3404 years	34043 years
	55 PB	12 days	124 days	3 years	34 years	340 years	3404 years
	1 PB	30 hours	12 days	124 days	3 years	34 years	340 years
	100 TB	3 hours	30 hours	12 days	124 days	3 years	34 years
	10 TB	18 minutes	3 hours	30 hours	12 days	124 days	3 years
	1 TB	2 minutes	18 minutes	3 hours	30 hours	12 days	124 days
	100 GB	11 seconds	2 minutes	18 minutes	3 hours	30 hours	12 days
	10 GB	1 seconds	11 seconds	2 minutes	18 minutes	3 hours	30 hours
	1 GB	0.1 seconds	1 seconds	11 seconds	2 minutes	18 minutes	3 hours
		100 GBPS	10 GBPS	1 GBPS	100 MBPS	10 MBPS	1 MBPS
	Network Bandwidth						

3. Online Vs Offline

Online data transfer refers to exercising online tools against the offline hardware storage devices to move data. Replication sync and cutover can be performed without any downtime with the online data transfer method.

Offline Migration



Online Migration

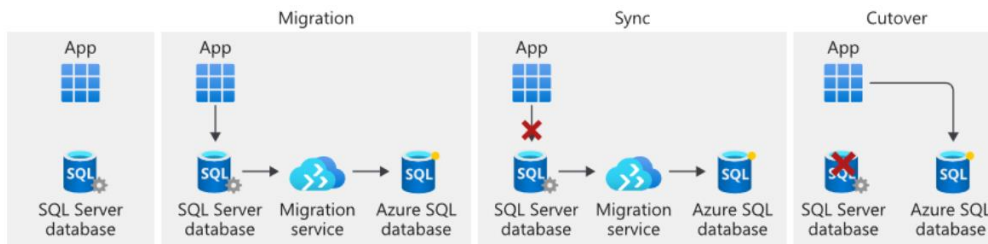


Table depicting transfer speeds for various data sizes and bandwidth – Source Google Cloud

4. Managed Vs Unmanaged data transfer technologies

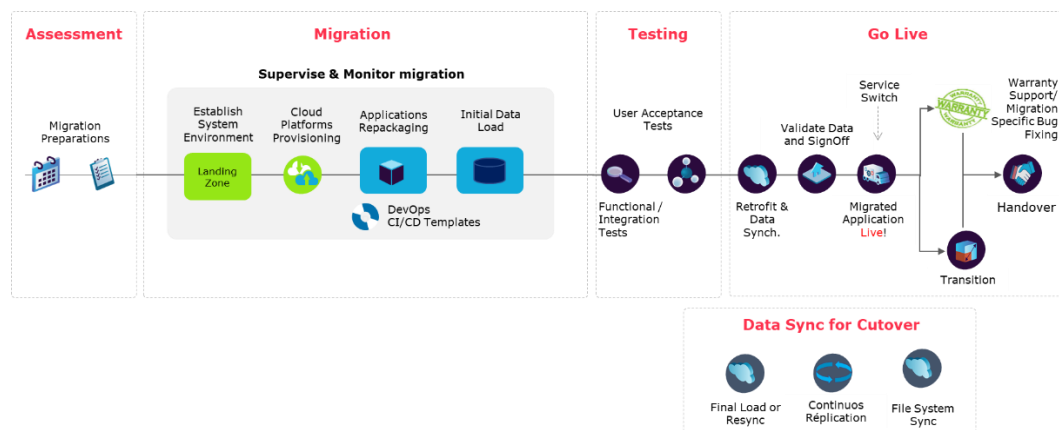
Managed Vs unmanaged services helps to address data transfer challenge on your new cloud, with minimal disruption, cost and time, by providing the smartest way to move your GB, TB, or PB of data.

Below is the suggested approach by Public Cloud provider, AWS for making decision on the data migration tool.

Connection	Data Scale	Method
Less than 10 Mbps	Less than 500 GB	Unmanaged
More than 10 Mbps	More than 500 GB	Managed Service

5. Data Transfer or Replication Tools and Technologies

Depending on the source and destination of the environments, we can decide on the tools for data transfer and replication. For databases, if you can leverage native tooling then it can work well provided there is expertise and good bandwidth available. The following is a typical data migration process used in application and database migrations.



Data Migration and Cutover Process

Depending on the building blocks, the right kind of data migration and replication pattern can be picked up. Further down this paper is a list of some of the data migration patterns that have been implemented in organizations extensively.

6. Data Classification and Security constraints

The more sensitive the information, the more is the reluctance in moving the data to cloud. The doubts or questions create a notion of insecurity while migrating data for business-centric critical data applications. It is thus very crucial to handle objections towards the compromise or threat to data security. Cloud infrastructure also comprises of open source codes and same vulnerabilities exists for them. An assessment pattern for data classification helps in clustering the application with sensitive data together or to identify vulnerable candidates. The encryption requirements for data in transit and data at rest should be captured during the assessment or planning phase.

Data Transfer Patterns:

The data transfer pattern is identified during the migration planning and should be prepared before the migration kicks off. One of the prerequisites for majority of the data transfer pattern is the connectivity between source and target environments or on-premise and cloud provider. Some popular services managed by cloud providers is AWS Direct Connect, Azure Site to Site connectivity.

	DATA SIZE RANGE	AWS	AZURE	GOOGLE CLOUD
MIGRATION PLANNING		<ul style="list-style-type: none"> CLOUDSCAPE TSO LOGIC CLOUD APPLICATION READINESS TOOL 	<ul style="list-style-type: none"> AZURE MIGRATE CLOUD ADOPTION FRAMEWORK APP SERVICE MIGRATION ASSISTANCE 	<ul style="list-style-type: none"> CLOUD ADOPTION FRAMEWORK CLOUD MATURITY ASSESSMENT
BULK DATA MIGRATION	TERRABYTES TO PETABYTES	<ul style="list-style-type: none"> SNOWBALL SNOWBALL EDGE SNOWMOBILE 	<ul style="list-style-type: none"> DATABOX DATABOX HEAVY DATABOX DISK 	<ul style="list-style-type: none"> TRANSFER APPLIANCE
DATA MIGRATION OVER NETWORK	UPTO 10 – 20 TB	<ul style="list-style-type: none"> DATA SYNC TRANSFER FOR SECURE FILE TRANSFER PROTOCOL STORAGE GATEWAY 	<ul style="list-style-type: none"> AZURE STACK EDGE DATABOX GATEWAY 	<ul style="list-style-type: none"> CLOUD ONLINE DATA TRANSFER STORAGE TRANSFER SERVICE
SERVER MIGRATION	UPTO 5 TB	<ul style="list-style-type: none"> SERVER MIGRATION SERVICE CLOUDENDURE MIGRATION 	<ul style="list-style-type: none"> SITE RECOVERY 	<ul style="list-style-type: none"> MIGRATE FOR COMPUTE ENGINE ANTHOS
DATABASE MIGRATION	UPTO 5 TB	<ul style="list-style-type: none"> DATABASE MIGRATION SERVICE SCHEMA CONVERSION TOOL 	<ul style="list-style-type: none"> DATABASE MIGRATION SERVICE 	<ul style="list-style-type: none"> BIGQUERY DATA TRANSFER SERVICE
STORAGE	UPTO 2-5 TB	<ul style="list-style-type: none"> NFS FSX EBS S3 	<ul style="list-style-type: none"> AZURE FILE STORAGE BLOB STORAGE MANAGED DISKS 	<ul style="list-style-type: none"> GOOGLE CLOUD STORAGE ZFS/AVERE GOOGLE COMPUTE ENGINE PERSISTENT DISKS

Table for ideal data handling capacity based on our migration experience

The following are some of the common data transfer patterns preferred during migration executions:

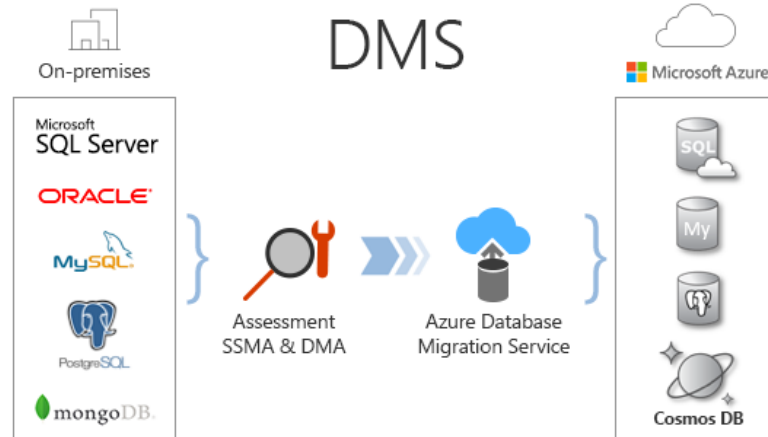
1. Patterns for Database Data Migration

- a. **Cloud Native Services** – If you have a direct connection between the source and target environments then using Cloud Native Services like AWS DMS and Microsoft DMS is a good way to move data for both Heterogenous and Homogenous migrations.

In Heterogenous Migration, the database engine or the database type is different in nature (example MS SQL to MySQL). In Homogenous migrations the database type remains the same (example Oracle to Oracle or MSSql to MSSql) These cloud native tools also provide successful transfer schemas between source and target databases and provide a means for online replication to maintain continuous data sync.

Cloud providers like AWS provide tools like Schema Conversion Tool that helps in converting schema objects and produce reports in case of heterogenous migrations like Oracle to PostgreSQL. Reports like these help developers in

understanding the risks during migration of DB objects by identifying how much code is automatically converted and how much would require manual efforts. Thus, with schema validations missing objects can be determined and can be created at target to demonstrate the success of a migration.



- b. **Database Native Solutions** - You can use some existing Database Native tooling to move data from source to target databases. Database provided patterns like 'Always On with Peer to Peer' replication for MS SQL is also a good way to replicate data. The initial load can be done using object storage or large data transfer service and then replication can be continued with one primary node in source datacentre peered to a secondary in the target datacentre.
2. **Using file transfer utility** – An organization's internal storage utility could be an effective tool for data migration if the data size is limited to < 2GB.
3. **Cloud Managed Data transfer solution** – Hyperscaler like AWS or Azure provide a variety of services to load data and store objects for applications that are migrated to cloud. These services offer different storage options depending on the use case and longevity of data.
For instance, in case of AWS, the FSX share can be used to mount storage and copy data. Similarly, the NFS devices from hyperscaler can be used to mount/unmount data and attach to the computing resources.
4. **Migration for Big Data Objects** – To transfer Big Data the transfer solution must be decided by the size of data and the transfer rate. For instance, 10TB data transfer over wire of 1Gbps would require approximately a day. So large data objects are preferred to be moved with exabyte-scale data migration service of hyperscaler. For example, services such as AWS Snowmobile or Snowball are preferred options to migrate large data sets that have more than 200TB of objects. Both the services have different features based on availability and volume. To migrate large datasets of 10PB or more in a single location Snowmobile is preferred. A good use case for snowmobile is a particular data centre exit scenario. For datasets less than 10PB or distributed in multiple locations, the preferred option is Snowball.

Conclusion

Data Migration is a crucial process during the cloud migration journey of an organization. It has to be treated as a separate process and efforts have to be calculated based on the parameters discussed in this whitepaper. The Cloud COE team can provision a data services team to provide the vital views

for planning architecture and security in the data migration patterns. As a good practice all migration related patterns must be engineered and tested well in advance before the execution. This helps in replicating various scenarios quickly and empowering application teams to expedite the migration journey.

References

Google Cloud - <https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets>

Alibaba Cloud - [http://alicloud-common.oss-ap-southeast-](http://alicloud-common.oss-ap-southeast-1.aliyuncs.com/Updated_Materials/Solution%20Whitepaper_Data%20Migration%20Methodology_V3.pdf)

[1.aliyuncs.com/Updated_Materials/Solution%20Whitepaper_Data%20Migration%20Methodology_V3.pdf](http://alicloud-common.oss-ap-southeast-1.aliyuncs.com/Updated_Materials/Solution%20Whitepaper_Data%20Migration%20Methodology_V3.pdf)

Amazon Web Service - <https://aws.amazon.com/cloud-data-migration/>

About Sogeti

Sogeti is a leading provider of technology and engineering services. Sogeti delivers solutions that enable digital transformation and offers cutting-edge expertise in Cloud, Cybersecurity, Digital Manufacturing, Digital Assurance & Testing, and emerging technologies. Sogeti combines agility and speed of implementation with strong technology supplier partnerships, world class methodologies and its global delivery model, Rightshore®. Sogeti brings together more than 25,000 professionals in 15 countries, based in over 100 locations in Europe, USA and India. Sogeti is a wholly-owned subsidiary of Capgemini SE, listed on the Paris Stock Exchange.

Learn more about us at
www.sogeti.com