Guide: Best Practices for Legacy Data Warehouse **Migration to Amazon Redshift**

By AGILISIUM
BIG ON CLOUD. BIG ON DATA.

aws PARTNER Advanced Tier Services

Contents

Introduction

Legacy enterprise data warehouses were not built to handle the volume, velocity, and variety of data of today's connected world. As a result of such limitations, enterprises today look to migrate to the Cloud from legacy enterprise data warehouses (EDWs). However – a recent Gartner report says, "83% of data migrations either fail outright or exceed their allotted budgets and implementation schedules". Typically, this is because organizations fail to invest time and money in the planning and assessment stages.

One of the popular legacy MPP databases, Netezza, was a popular choice for rapidly analyzing petabyte-scale data volumes. It came with its hardware ecosystem and proprietary data formats and reached end-of-support in June 2019. This forced Netezza customers to migrate their data platform. A top modern EDW considered by such organizations is Amazon Redshift.

Amazon Web Services (AWS) is the world's leading provider of cloud infrastructure services with organizations of all sizes – from start-up to Fortune 500 companies as its customers. Redshift is AWS's oldest data warehouse offering and the top cloud data warehouse service worldwide. Some of the world's biggest organizations – Intuit, Johnson & Johnson, Yelp, and even McDonald's – use Redshift.

Considered alone, Redshift solves only part of the puzzle. However, when considering the entire Amazon ecosystem, the advantages are clear. AWS allows innovation, superior execution of existing use cases, and offers additional benefits. Because of these reasons, many of them choose to migrate to Redshift.

Let’s look at how to mitigate risks during a migration from a legacy MPP database like Netezza to AWS Redshift.

Why move to cloud ?

An organization can choose to migrate data for any of the following reasons – data platform modernization, cloud transformation and end of support and licensing.

In today's increasingly connected world, the explosion of Big data means that pretty much any organization collects a vast array of data that are worlds away from the kind of data that legacy databases were equipped to handle. Besides, this data is further analyzed in near real-time to enable data-driven decision making, another use case the legacy databases were simply not built to fulfill.

Furthermore, legacy databases operate in the pre-subscription, pre-cloud era, and therefore have all the associated disadvantages:

High Costs

- Complex and expensive licensing terms
- Proprietary data format
- CAP-EX vs OP-EX/upfront investment vs. pay-as-you-go (cloud)

Lack of adaptability

- Rigid—can't adapt new technologies to keep up with innovation (unable to query open formats like Parquet, ORC and JSON, cannot query directly from data lake)
- Scalability only at steep cost
  Inability to handle a variety of data formats
  Lack of data lake support
- New projects have long implementation
- cycles and high-failure rates

Operational inefficiency

- Limited or no self-service data availability for a business user
- Exorbitant cost of maintenance – resources, hardware, and extended timelines

For organizations that built their EDW on Netezza specifically, in addition to all the above, end-of-support is a key reason to migrate.

Decoding your Current Environment

Typically, once a firm decision to migrate to a new platform is made, the next question arises - How exactly is modernizing an EDW done? A well-thought-out requirement gathering phase is crucial. We can break it down into four stages:

Understanding existing data:

Often the knowledge of a data ecosystem is spread out across multiple resources within an organization. Therefore, organizations need to identify key stakeholders, talk to the owners of different pieces of their data platform and thoroughly understand how it all works together and how a piece of data gets from point a to point b.

Building a holistic data ecosystem picture

As the current system information is gathered, attention must be paid to collect the information listed in the grid below to paint a holistic picture of the existing data ecosystem. This information is essential to ensure that the new data platform can include existing compliance, regulatory requirements, and unique use-cases.

Data	Data Platform	ETL & Data Ingestion	Security & Governance

• Data Volume
• Variety Velocity
• Data growth (%)
• Data lifecycle
• Data Lineage
	• Types
	  • EDW
	  • Data Mart
	  • ODS
	• DB objects
		• Number of Processes & Complexity
		• Process dependencies
		• ETL Tools & Connectors
		• Real-time requirements
		• Data Science Support
			• Encryption
			• ACL - Roles & Permissions
			• Regulatory & Compliance
			• Audit & Security Monitoring

Consumers(BI & ML)	**Ops Requirements**	**DevOps**	**Other factors**
• BI & Insights • Interactive access • Data Science • Data as a Service	• Monitoring • Workload Management • Performance & Scalability • Availability, Backup, & DR • SLA Guarantees	• Continuous Integration & Continuous Delivery • Infrastructure Automation • Test Automation	• New Business requirements • Technical Debt • Time & Cost

Business and functional use cases

Today, most organizations don't just connect their data platforms to a business intelligence (BI) tool; often, they have data science and real-time analytics requirements. Therefore, gaining a good understanding of the requirements for ad-hoc querying, data science workloads in addition to the BI insights needed. There's also a possibility of data-as-a-service that is offered via microservices-based architecture or Mulesoft integration. Information on such integrations is necessary to provide a higher quality fitment to the cloud provider under consideration, i.e., AWS.

Operational Requirements:

Here's a concise list of the kind of decisions to ensure that the new data platform is run and maintained smoothly:

1. Monitoring of the new data platform (third party vs. AWS's own)
2. SLA for various operational and business tickets
3. Backup schedule for stored data

Besides, on moving to Cloud, implementing DevOps methodology is imperative even for your data ecosystem. DevOps will enable the platform to be nimble and pivot quickly to provide faster end-user value.

Netezza specific questions to ask your service provider before migration

We've seen how to go about migration requirements for any legacy MPP database migration. Here's what an organization using Netezza needs to ask and answer

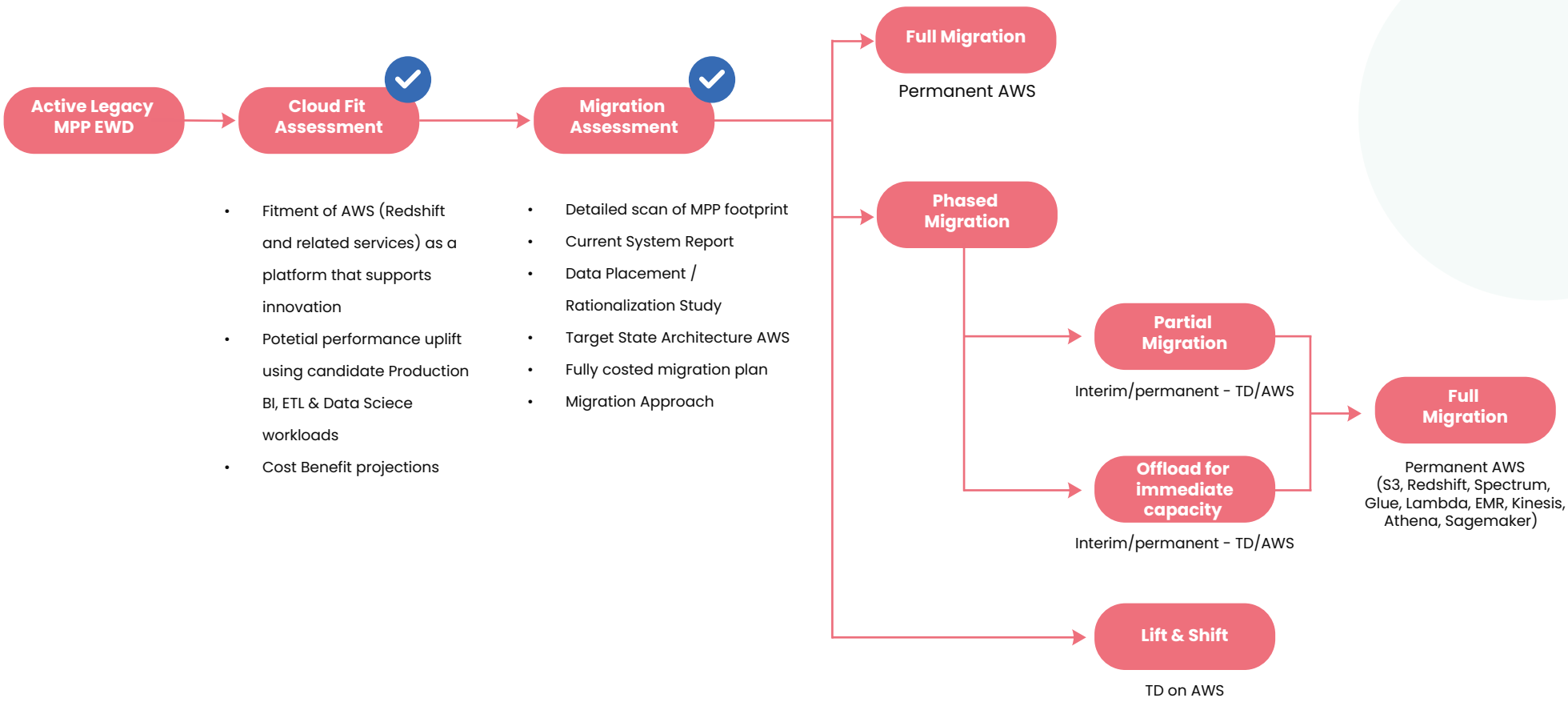
CPU	I/OI	DISK
• I have capacity problems: CPUspecifically, how can I manage CPU to accommodate future growth? • What is my cost for accessing the data? • Where are the resources invested? • Highest consumers of CPU & IO • Complexity of workload • Query Concurrency	• What ETL Processes can be offloaded or retired? • What is the cost of loading this data?	• Which data can I offload to other platforms on AWS? • Are there any datasets I can archive? • How much history should I keep in the warehouse? • How is my storage used: nactive/Cold/Warm/Hot data?

Planning and Budgeting the Migration

On identifying pain points, documentation of new needs, and the thorough analysis of the existing ecosystem, we move on to the all-important planning phase and budgeting for the migration. Platform-wide migration often requires a cloud systems integrator's skilled expertise in both the legacy and target ecosystem and an in-depth and varied strategy to tackle any complexities during migration successfully.

At Agilisium, this is called the migration pathway, and it is broken into two distinct phases – a cloud fit assessment followed by the migration assessment. Let's see in detail what happens in these phases below.

Migration Pathway



Cloud fit assessment

The Cloud fit assessment is the first stage where Agilisium takes your requirements and thoughtfully maps it out to a cloud-based service i.e., AWS, to clearly represent that the cloud provider can provide or support all stated requirements. E.g., If there were a need to support existing ETL pipelines running on top of Informatica, Agilisium's experts would map it to Redshift. For data science requirements, Amazon Sagemaker would be the AWS service that fulfilled those needs.

This process is repeated until all services are mapped out to their cloud equivalent. Typically, AWS can provide all the services needed. However, the conscious process of mapping confirms and documents that the right tools are available for each job.

On validating that the cloud provider (AWS) in this the manner detailed above, we move on to the migration assessment

The migration assessment

This step is where the organization and Agilisium collaborate to deep dive into requirements and map out how exactly the migration will proceed. Here are some of the key questions that can come up during this stage:

a. Is there a need to expire or archive some data, or can all data be migrated?

b. How will each of the services identified in the fit assessment be used?

c. How will the re-architected framework look?

d. What is the TCO and timeline for this migration?

Once this stage is complete, the migration follows one of the two pathways detailed below

a. Full-scale migration – Chosen when the requirements are straightforward, with fully contained systems and with a fair idea of data storage and dependencies. The service provider comes back with a timeline of the migration that is the right fit for the stated business needs.

b. Phased migration: A large, complex implementation usually proceeds in a phased manner. Typically, the service provider takes the fringe or complicated requirements, move it to the Cloud first and build a system around it.

As each workload migrates and the client enjoys real-world benefits, many organizations come back with requests for the service provider to rethink their data platform strategy. A top-notch service provider with deep expertise in AWS would be invaluable here and help the organization extract even more value out of their AWS investment.

This is a broad overview of Agilisium's approach to migration. So how can Agilisium make such a complex process involving multiple stakeholders be rapid, precise, and predictable?

Agilisium advantage

Agilisium's collective experience gained by working on multiple projects has resulted in an exclusive set of toolkits and setup accelerators that ensure that these assessments are thoroughly right and the migration is rapid.
Each toolkit covers key facets uniquely applicable to Redshift and includes expanded questions added by our experienced and certified experts. Here's a list of some of these toolkits.

ASSESSMENT TOOLKIT

DISCOVERY QUESTIONNAIRE: Short and complete questionnaire to capture innovation requirements, current system details, and pain points

DATABASE PARSER: Analyse DB objects (Tables, SPs, Views, Functions, Join Indexes, etc.) and complexity levels from MPP DB's
metadata

ETL / XML PARSER (Informatica, Data Stage, SSIS, Ab initio): Identify map count, target schemas and tables impacted, insights on ETL transformations used, any in-DB optimization, and external scripts used (BTEQ, NZLOAD, FastLoad, PLSQL, etc.)

HISTORY LOAD SCRIPT GENERATOR (Informatica, DataStage, Wherescape): Autogenerate scripts to load historical data from source MPP DB for assessment

DLL Converter: Converts DDL for Tables and Views from any DB to Redshift

MIGRATION FACTORY

DATA MIGRATION TOOLKIT: Python-based toolkit to automatically retrieve, load S3 data into Redshift and optimize it (Encoding & WLM)

METADATA-WISE ETL FRAMEWORK: Control table-driven generic ETL framework to modernize legacy ETL pipelines

POST-MIGRATION VALIDATOR: Custom utility for integrity checks, validation, and audit balance control post-migration

DE-DUPLICATOR: Fuzzy logic-based de-duplication tool to smartly find duplicates and harmonize them

AUTOMATION: CI/CD for infra and Data Ingestion pipelines and Test Automation

Key steps in Netezza to Redshift migration

See below a thorough list of the critical steps specific to migrating from Netezza to Redshift

Pre-Migration Planning	Migration Assessment (Decoding your Current Environment)	Migration Assessment – Phase 0	Development and Testing (Database Migration)
• Cloud Readiness			
• TCO Analysis	• CPU/IO/DISK Analysis	• Migration Planning & Strategy	• DDL Conversion (Tables)
• Data Placement & Rationalization	• Current Workload Analysis	• Future State Architecture	• History Load Scripts
• Workload Offload Strategy	• Daily Load Process Analysis	• Data Movement Approach	• History Load 1
• Cloud Connectivity/ Network Bandwidth	• Current System Analysis	• Conversion Strategy	• DDL Conversion (Views)
	• Current Database Analysis: All Objects	• Project Plan	• Stored Procedure Conversion
• Security		• Timeline	• Functions
• Latency and SLA's	• Current ETL Process Analysis		• ETL Conversion/Repoint
• Cloud Migration Strategy	• Current BI Process Analysis		• BI Conversion/Repoint
• ETL Server Migration Strategy	• Size and Volume		• Optimization
• Data Movement Strategy	• Backup and Restore		• Tuning
• Backup Strategy			• Unit Testing
			• System Integration Testing
			• History Load 2
			• User Acceptance Testing
			• Breaks/Fixes
			• Parallel Run Testing

Where to start and how to evalute a migration partner?

Any organization looking to engage a service provider for a Redshift migration assessment can use the checklist below to evaluate whether they offer genuine value and expertise,

- Do they have AWS certified experts, and have they successfully migrated significant amounts of data to Amazon Redshift?
- Will the stated outcomes/assets from the assessment give you an understanding of the Current State environment and the desiredfuture state outcomes?
- Do they offer a detailed analysis of the current DBMS environment, including DDL dump and XML Dump?
- Do they have an arsenal of best practices, architectural designs, migration patterns, and customer references designed to expand the customer knowledge of Amazon Redshift and other related AWS tools?
- Can they recommend a high-level schema /ETL / application migration architecture and plan to facilitate delivery: (Lift and Shift/Forklift with enhancements)
- Can they build and deliver a Proof of Concept that gives you clarity on how the new Redshift architecture will function?
- Are they able to offer a strategic (Future state in AWS) and tactical roadmap for a Full or Partial Migration?

If the service provider checks most of the above boxes, it's reasonable to assume that they can work with you to deliver a joint vision around cost savings, migration strategy, and future state. This joint vision could take the form of the following:

- Current System Report
- Data Placement/Rationalization Study
- Future State Architecture AWS
- Migration Approach, Tasks, Timeline
- Detailed Cost, Resource Plan

When choosing your migration partner, they must offer most, if not all, services on the checklist above

Conclusion

Moving from a legacy DB infrastructure to the Cloud is a complex task. When it is an out-of-support EDW like Netezza, there is no rollback possible if things go wrong. Hence, organizations must invest in the early planning and assessment with the right talent and platform for a successful migration. In this guide, we've laid out the various strategies that can be adopted to minimize the risk. By investing in thorough assessment and with the appropriate use of proven accelerators, such as the ones offered by Agilisium, a Netezza to Redshift EDW migration can be precise, predictable, and rapid.

sales@agilisium.com