ADASTRA

Migrating from Legacy Databases to Amazon Redshift

adastra.digital.com

Introduction

As the reliance of businesses on fresh, timely insights for decisionmaking increases, they have come to recognize the crucial role played by data generated both by their own business processes and their suppliers, partners, and other external sources. Analytics is by no means a new business need, and early adoption of analytics dates back to the late 90s. However, both databases and the analytics tools themselves have undergone a sea change in order to accommodate the massive volumes of data and bring efficiency into the process.

Looking back to the 90s and early 2000s, the earliest database platforms came as a blessing for organizations that were already handling large data volumes. Many of the early adopters picked Teradata as their core database platform. This was an impressive platform, but the costs limited its adoption to mature organizations willing to make large investments and capable of making a business case. However, while customers undoubtedly benefited from even these early solutions, the implementations were not devoid of challenges.

Fast forward two decades and some of these platforms are still around, albeit following multiple upgrades and an overhaul of the solution itself. The digital transformation wave has led to exponential growth in data volumes, and drastically changed how organizations collect, transform, analyze, and visualize data. Business use cases for data are now broader and user needs are evolving at a faster pace than ever before. Data volumes have increased hand-in-hand with data velocity, with data being produced and ingested much faster, at times with impact to value and risk. Data variety, too, has broadened, and modern applications are producing a lot more semi-structured data, in XML and JSON formats. Faced with these challenges, legacy databases can no longer offer the value they once did.

When these legacy platforms were first introduced, business requirements were much simpler, and even though there were limitations associated with their use, they were possibly the best available option. That is no longer the case today, and many organizations find legacy platforms to be a hindrance to business agility and change.

The Evolution of Legacy Databases and their Limitations

While some of the challenges associated with database platforms have changed over time, most remain unaddressed to some degree. Let us look at some of the key limitations and their evolution over the last two decades:

Security:

Data security is top-of-mind for everybody, and organizations seek assurance that their data is protected and accessible by the right users, with mechanisms in place to monitor security threats. THEN: A few decades ago, however, with data on-premise, these concerns were rather low and security requirements were not as stringent as they are now. Data was often not encrypted at rest and security-focused monitoring was relatively basic. NOW: Since then, legacy solutions have evolved to handle more recent security regulations.

Scalability:

As businesses grow, they collect more data and make more complex use of it. The ability to handle more data, users, andqueries, with no changes to the architecture and minimum provisioning time are immensely important.

THEN: The older legacy platforms could scale and handle Terabytes or even Petabytes of data, but all the hardware required space and capable infrastructure, including power, cooling, etc., which were available in purpose-built data centres, but possibly a challenge elsewhere. This came at a cost and required long provision times.

NOW: The newer versions, too, will scale to handle the requirements of typical organizations, but continuous growth adds to provisioning effort and costs. Moreover, with the lack of elasticity, additional limitations emerged, as even implementation of limited time solutions requires the acquisition and retention of additional capacity. This meant that the platform is sized for peak utilization with extra capacity that often lies unutilized.

Provision time:

Time-to-market is a key consideration to keep up with internal demand and remain ahead of the competition.

THEN: When legacy platforms were relatively new, it took over one month to merely procure the hardware, and then weeks to install and configure it. Scaling and additional capability building had to be planned well ahead of time due to the long provision times and the effort it entailed.

NOW: While it may not take as long to provision new hardware, the process still requires frequent, detailed, and ongoing planning to accommodate increasing data volumes and rapidly changing business needs.

Capability and Integration:

Databases are only one component of a solution, albeit an important one. It is important that a database satisfies users' requirements for more, broader capabilities or integrates well with other components that can do that.

THEN: Twenty years ago, data of all kinds was squeezed into relational structures or managed as binary objects, diminishing its usability. Semi-structured data was relatively scarce at that time and the older platforms were not equipped to handle it.

NOW: Although new-age legacy platforms have evolved to handle semi-structured data, its use is not fully transparent to users. They can even integrate with the Data Lake, but their legacy architecture limits their extensibility, and new capabilities typically trail new vendors in time and scope.

Resilience:

Mature organizations make the best of their data, so analytical platforms play a critical role and need to be managed as such. Failures imply downtime and possible data loss, and impact users and decision-making. It is imperative to keep the platform available and resilient to failures.

THEN: In the initial versions of the legacy platforms, data was normally safguarded by storage redundancy, via RAID configurations, but nodes were still single points of failure, that is, the falure of a node meant that

the whole database would be inaccessible. Recovery time, assuming that another node was readily available, often took more than 1 day. Disaster Recovery solutions could help mitigate these challenges, but they were not simple to architect and more importantly, they involved considerable costs.

NOW: Today's legacy databases need to support more data, more users and broader use cases, and the need for resilience has only heightened, putting more pressure on data availability, and lowering the risk of data loss. The cost of Disaster Recovery is still quite large and almost doubles the overall cost.

and MDM templates for financial, retail, automotive, and healthcare industries, among others. They also continue to develop and provide methodologies for iterative, extensive, and scalable MDM solution planning and implementation, as well as accelerators for data quality, matching and merging rules, and AI and ML Augmented Stewardship.

Management:

Monitoring, applying patches, performing backups and performance tuning are some of the tasks required for maintaining a database platform, to keep it healthy, fast, and meet Service Level Agreements. Management requires considerable effort and various people with specialized skills.

THEN: In the earliest legacy platforms, monitoring often leveraged third-party tools, which added to the costs. As one can imagine, patches required planning, downtime, and overtime, and backups took considerable time.

NOW: With increased usage, the need for monitoring and keeping the platform operating at peak performance has only increased.

Cost:

Cost should not be the biggest driver in the selection of a database platform, but higher costs inevitably make it more difficult to create a viable business case. More importantly, it moves the focus away from business value and inhibits innovation. THEN: In the late 90s and early 2000s, the costs entailed in setting up, operating, and managing legacy platforms were very high. The servers often leveraged proprietary

hardware and a complex network architecture, all of which were expensive. SANs with high performance disks had 7-figure price tags, and the costs nearly doubled when involving High Availability/Disaster Recovery. Moreover, there were also management costs involved for all these resources. Additional hardware and software costs were linked to growth and unfortunately, not in a linear fashion.

NOW: In the present times, data growth-related costs with legacy platforms continue to remain non-linear. In order to support their users, many organizations are forced to keep paying high costs, or even worse, compromise on the solution to keep costs manageable. These platforms have a balanced approach to storage and compute, meaning that even if only one aspect requires growth, the customer is forced to pay for an increase in both areas. Due to lack of elastic capabilities, additional costs accrue due to unused capacity, which is established based on peak utilization. Maintenance costs, which typically run up to 20% of the overall cost, really add up, even without much innovation or improvements to the platform.

Migrating Legacy Solutions to the Cloud

As one can imagine, while legacy platforms had limitations to begin with, in today's fast-paced business environment with its complex needs, organizations find it increasingly harder to work around the challenges posed by these platforms and adapt at the pace that is required.

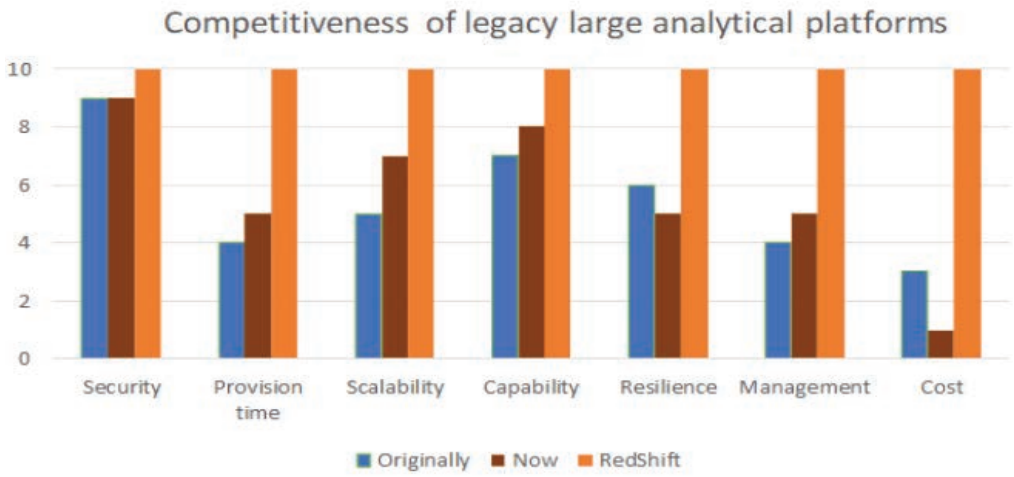
Migration to the cloud has opened a new world of opportunities for such organizations. However, the decision to migrate should not be made on the merit of the database alone, but of the ecosystem as a whole.

With its promise of unprecedented flexibility in terms of platform, infrastructure, processes, etc., the cloud enables organizations to address existing challenges and unlock new opportunities. For instance, it is possible to have the database on the cloud while keeping the remaining solution on-premise or even on another cloud. However, this setup has performance and integration limitations that void any perceived advantages of that cloudspecific database platform. For most organizations, the decision to move to a cloud provider is based on a mix of criteria, including the overall ecosystem offered by the cloud, its integration with various elements, tools offered, ability to innovate, and of course, costs.

Amazon Redshift is a fast, simple, and cost-effective Data Warehousing service. While Redshift's inception was not on the cloud, it was acquired by AWS in 2012, and has evolved dramatically to take full advantage of the benefits associated with the cloud. It is an important service within the vast AWS ecosystem, which is the largest, most mature cloud provider in the world.

How Redshift Compares to Legacy Databases

Now, we review Redshift's capabilities across the parameters we considered for legacy platforms and see how it compares to them. The below graph is based on our analysis, and we detail our rationale for Redshift's comparative competitiveness in this section.



Security:

Operating on the cloud, Redshift has strong security, including encryption at rest and in transit. Clusters can be isolated using Amazon VPC and customers benefit from strong access controls, security threat detection and monitoring services. Default Redshift configurations are compliant with SOC1, SOC2, SOC3, PICC DSS Level 1 requirements, DoD, ISO, FedRAMP, and HIPAA.

Provisioning:

Provisioning can be done easily and in minutes, via the web console, APIs, or automated via CloudFormation.

Scalability:

Clusters can scale from 160GB to 2PB with resizing requiring no downtime. Concurrency scaling adds additional cluster capacity to handle periods of unusually high utilization, via increased number of users or queries.

Capability and Integration:

Redshift clusters can leverage one of three types of nodes, targeted at different scenarios, such as compute or storage dense, and 3rd generation which decouples compute and storage. Redshift is integrated with the Data Lake through Redshift Spectrum, providing access to Exabytes of data. It is also integrated with PostgreSQL through federated queries.

Resilience:

The service offers multiple levels of data redundancy to provide 99.99999999% durability. This includes multiple copies across disks, but also across multiple availability zones through S3. Data can be further replicated across multiple regions. Node failures are automatically monitored and will be replaced by new ones, with data recovered based on most recently queried data.

Management:

The database platform is fully managed by AWS. This includes automated patches, backups, and monitoring, and reduces the need for skilled resources for management and maintenance. Redshift also automatically chooses table distribution styles and optimizes queries.

Cost:

Unlike legacy platforms, Redshift is a service, and the costs are based on the evolving required capacity, which eliminates the wastage associated with unused capacity driven by peakutilization. With costs as low as $1000/Terabyte per year, Redshift's costs and capabilities are hard to match for organizations looking to move their data to the cloud. Redshift scores better than legacy database platforms on all parameters, and addresses some of the key challenges that have been associated with legacy platforms since their inception. Moreover, the service has been evolving at an astounding rate, with speed, performance, and tools improving every year.

hardware and a complex network architecture, all of which were expensive. SANs with high performance disks had 7-figure price tags, and the costs nearly doubled when involving High Availability/Disaster Recovery. Moreover, there were also management costs involved for all these resources. Additional hardware and software costs were linked to growth and unfortunately, not in a linear fashion.

NOW: In the present times, data growth-related costs with legacy platforms continue to remain non-linear. In order to support their users, many organizations are forced to keep paying high costs, or even worse, compromise on the solution to keep costs manageable. These platforms have a balanced approach to storage and compute, meaning that even if only one aspect requires growth, the customer is forced to pay for an increase in both areas. Due to lack of elastic capabilities, additional costs accrue due to unused capacity, which is established based on peak utilization. Maintenance costs, which typically run up to 20% of the overall cost, really add up, even without much innovation or improvements to the platform.

Adastra's Approach for Migrating to Redshift

Migration Assessment and Roadmap

There are multiple strategies to approach such a migration but with any of them, it is critically important to define an accurate and detailed plan that will provide quick value and minimize impact on the business.

The strategies can be characterized as:

Lift and Shift: This approach focuses on the data itself, whereas the data processing is done externally to the database, traditionally by an ETL tool. Users will continue to get the same business capabilities, meaning that their needs are stable.

Re-factoring: This approach will also reproduce the same business capabilities but in addition to the data migration, it will also require a data processing rewrite, since the previous solution has leveraged an ELT approach and/or leveraged native utilities.

Re-architecture: This scenario is enacted when the previous solution is outdated and falls short of meeting the current requirements. This can touch on data scope, data processing scope and definition as well as support to new use cases.

While some organizations immediately recognize the need for a rearchitecture, a detailed analysis of the current solution is always required. After all, much of the data scope and a considerable portion of data processing and business requirements are still valid. When the migration approach is not pre-determined, the analysis is complemented by an assessment to ascertain strengths and weaknesses of the solution, to help inform which aspects can be preserved or otherwise, require more rework and effort.

We also challenge requirements because the market is continually evolving, and organizations need to adapt their services and required capabilities. As a result, we challenge the data scope and inquire about any additional data sources of relevance. We also revisit data processing, including business rules. Finally, we hold discussions about the relevant use cases for the organization and what additional ones, if any, are of interest. All of this will define the type of migration, duration, effort, and changes. Changes are important as they will require additional testing, whereas in absence of that, we know what outcome to work towards.

While larger changes and re-architecture may at times be required, it poses a challenge to organizations that want to continue realizing some benefits. Typical scenarios include:

- **Parallel platforms:** Users will continue to leverage the current platform while the new one is being worked on followed by a transition period.

- **Quick transition followed by evolution:** This can be dictated by an end of life of the current database infrastructure and/or by risk minimization. In this case, existing capabilities are migrated and reproduced, providing users with similar capabilities. Once the solution running on the cloud and Redshift is stable, additional changes can be planned and incrementally deployed.

Migration Assessment and Roadmap

Before migrating the data, it is important to make any necessary schema adjustments. Most data types can be migrated as is, but the use of AWS SCT (Schema Conversion Tool) significantly facilitates this verification and migration. SCT also migrates most of the database views.

These solutions commonly hold Terabytes of data that are migrated in two phases:

- **Initial load:** At this time, we migrate the bulk of the data, which is no longer changed, e.g., up to the end of the last month. We reconcile all the data across both database platforms and assure organizations that there are no data loss or integrity issues.

- **Delta load and final catch-up:** Since the transition from the old to new platform may take weeks, it is important to keep them in sync (as much as possible), when it comes to new data. For example, we can execute this on a daily or weekly basis. At the very end, just before users are transitioned, we migrate the last set of changes, which may account for a few days or hours of changes.

The data migration commonly leverages AWS Database Migration Service, a hugely versatile and capable service which enables all the above actions (and more), in batches or continually.

Data Processing Migration

Teradata-specific data processing: In cases where organizations leveraged Teradata capabilities for data processing, such as stored procedures or utilities, we can also leverage AWS SCT which can automatically convert much of the code.

ETL data processing: In cases where an ETL tool is used to perform the logic, e.g., Informatica, there are two alternatives:

- **Continue leveraging the same tool:** We check the compatibility between the tool and Amazon Redshift, which is hugely common. Typically, changes are limited to JDBC connectivity.

- **Re-factor the code and adopt AWS services:** Processes can be rewritten in AWS Glue, which leverages PySpark, thus providing high performance and scalability. This solution is not proprietary to AWS and can be easily ported to other platforms, such as Databricks.

Conclusion

While many organizations are moving their platforms to the cloud, the adoption of services like Amazon Redshift to support their analytical platform is not done because it is mandatory. Many solutions can, in fact, be ported with no or little changes. The adoption of Redshift is justified due to diminished value on return of the old solutions, which have become outdated in their ability to fully meet the organization's needs and more importantly, in their ability to do so quickly. Time to Market is of critical value to organizations in today's dynamic, fast moving market.

Amazon Redshift's high integration with other AWS services, its security, high performance, scalability, continuous innovation, ease of management and low costs further help to make the case for such transition.

ADASTRA

About Adastra

Adastra transforms businesses into digital leaders. Since 2000, Adastra has been helping global organizations accelerate innovation, improve operational excellence, and create unforgettable customer experiences, all with the power of their data. By providing cutting-edge Artificial Intelligence, Big Data, Cloud, Digital and Governance services and solutions, Adastra helps enterprises leverage data that they can control and trust, connecting them to their customers – and their customers to the world.

Adastra has been helping companies for the past 20 years, across various industries in multiple lines of business realize value in their data, with our award-winning expertise, proven methodologies, and highly qualified team. Let Adastra help your company achieve data quality excellence.

Contact infosk@adastragrp.com to schedule a free consultation.

adastra.digital.com