Migration Strategies for Legacy Data Warehousing Systems to Cloud Platforms

Santosh Kumar Singu

Senior Solution Specialist, Deloitte, 338 Autumn Sage Dr, Indian Trail, NC – 28079, USA
Email: *santoshsingu7[at]gmail.com*

Abstract: ***This article discusses one of the most pressing tasks of transferring traditional data warehousing platforms to modern cloud versions in the context of the critical demand for fast, reasonably priced, innovative data processing systems. They include scalability, cost, and analysis features to understand the key motivators for migration from the above outline. This paper aims to provide the best practice guidelines in assessment and planning, choosing the cloud platform, migration approaches, schema, testing and performance, security, and change management. Setting up the article as a case study, the author demonstrates real - life examples of migrations from Teradata to Amazon Redshift, Netezza to Google BigQuery, and Oracle to Azure Synapse Analytics. Migration After - research also covers the typical issues that repeatedly recur during the migration process and possible ways of avoiding them. The conclusion then contemplates the potential of emerging technologies in data warehouse computing. It focuses on cloud migration as the key to sustaining the edge amid the fast - changing data environment.***

Keywords: warehousing, scalability, cost, migration, technologies, and cloud

1. Introduction

Most organizations realize the need to update their structures in a fast - paced environment. This year's major shift in this field is the transition of old - styled data warehousing systems from physical to cloud. This article tackles how one can migrate traditional on - premise Data warehouses (Teradata, Netezza, and the like) to Cloud solutions and iterates with relevant examples of new - age cloud - based solutions.

The Need for Migration

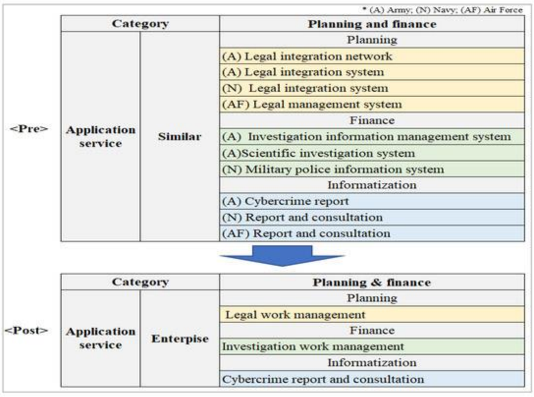
There are many reasons as to why migration from traditional data warehouses to the cloud is occurring. One of the main benefits of cloud platforms is that they are virtually infinitely scalable, so an organization can easily modify the amount of IT resources it provisions as needed. In many cases, they are becoming more cost - efficient, freeing the need for mass investment in physical gear and low maintenance expenses. Today's cloud data warehouses offer Machine Learning and AI features that make analytics more accessible and efficient. These platforms use distributed computing and sophisticated optimization techniques to execute queries faster. Security is of great concern to cloud providers because they dedicate resources to security to offer clients more than some corporate data centres might be able to offer [8]. Furthermore, cloud solutions are flexible in some regards, such as data integration, tool compatibility, and geographical location.

Best Practices for Data Warehouse Migration

Assessment and Planning

Therefore, before engaging in a migration project, a complete evaluation of the current state of the data warehouse should be made. This includes listing all data sources, tables, views, Stored Procedures, etc. As for the organizations, it will be reasonable to analyze query patterns to define which queries are the most frequently used ones, when precisely the traffic is at its highest, and what performance issues may occur. One should evaluate existing volumes of data and predict their growth in the future. All the applications and processes that rely on the data warehouse should also be identified [1]. The migration objectives should be set concerning performance standards and cost constraints to be met.

The figure below is an example of integrating similar services related to the planning and finance function to create a new enterprise service.

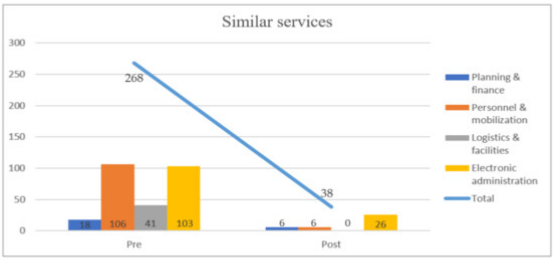


* (A) Army; (N) Navy; (AF) Air Force

	Category		Planning and finance
<Pre>	Application service	Similar	Planning
			(A) Legal integration network
			(A) Legal integration system
			(N) Legal integration system
			(AF) Legal management system
			Finance
			(A) Investigation information management system
			(A)Scientific investigation system
			(N) Military police information system
			Informatization
			(A) Cybercrime report
			(N) Report and consultation
			(AF) Report and consultation

	Category		Planning & finance
<Post>	Application service	Enterpise	Planning
			Legal work management
			Finance
			Investigation work management
			Informatization
			Cybercrime report and consultation

The bar chart below displays services of two phases: planning & finance, personnel & mobilization, logistics & facilities, and electronic administration; division into pre - and post is also shown. They show a much higher total in terms of the 'Pre' phase, where there were 268 services in personnel & mobilization and electronic administration. In the "Post" phase, the total number of services reduces to 38; a decrease is observed in all the buckets [5].



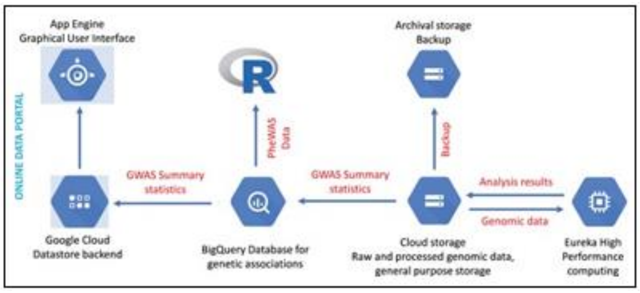
Choosing the Right Cloud Platform

Choosing the right cloud is, therefore, important for the migration process to be successful. Other considerations include compatibility with existing data types and workloads, query performance and data loading, elastic and flexible computational and storage services, and costs [13]. It is thus crucial to consider the availability of tools and services that can link up with the platform and geographical presence. Currently, cloud data warehouses are Amazon Redshift, Google Big Query, Azure Synapse Analytics, and Snowflake.

Data Migration Strategies

Depending on the requirements, organizations can select from several options for the data migration approach. In the "Lift and Shift" strategy, the current DW physical structure is migrated to the cloud with the least changes. Though fast, it can still leave optimization of cloud functionalities suboptimal when re - platforming actual data. Modifying the existing data model occurs to align it to the target cloud platform better, leveraging cloud peculiarities. Refactoring, on the other hand, entails redesigning the data warehouse afresh to optimize it for cloud - native resources and advanced data organizations. A migration strategy can also be gradual, with some tasks being shifted to the cloud while others are left locally for a while [2].

The figure below depicts a flow of data for genomic analysis and storage that connects several technologies. Sequencing data is archived in scalable, redundant cloud storage and backed up in tape storage, while Eureka high - performance computing is used to process analysis results. Information and summary statistics are passed between BigQuery databases, Google Cloud Datastore, and an app engine graphical user interface for frequency analysis and genomic information [9].



Schema and Query Optimization

More importantly, specific attention should be paid to the fact that the data model and the queries must be optimized for the cloud environment. This will involve the depopulation of standard tables when it encourages query optimality, spread and clustering of data following cloud - preferred techniques when necessary, creation of materialized views for often accessed partial datasets, and optimization of SQL queries for cloud platforms customarily [3].

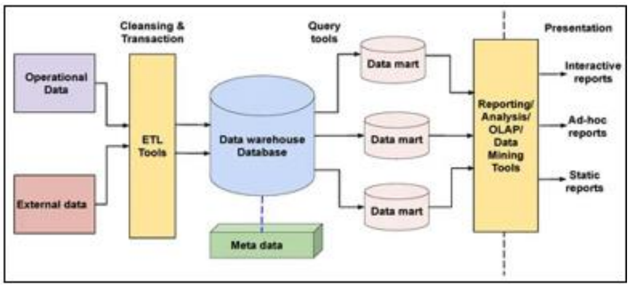
Testing and Validation

It is imperative to stress that the mentioned issues that may occur during the migration stress the importance of deep testing. This may consist of data validation, where the validation ensures that all data was exported and imported appropriately and in its entirety; query performance test, where query performance between the successive system and the cloud operation is compared; function test, where all existing reports and applications run correctly in the new environment, load test where loading is done to put the system under extreme pressure and fail over and recovery test where testing is done in terms of disaster recovery and high available scenarios [4].

Performance Tuning

When migration is done, attention should be directed towards improving performance within the cloud platform. This incorporates reasonable resource allocation and query optimization strategies, querying resource caching mechanisms inherent in the cloud platform, constant querying for slow query optimization, and efficient distribution of queries in nodes for parallelism [9].

The image depicted below is a data warehouse architecture. Real - time and transactional information passes through operational and external data ETL tools, undergoes cleansing, and is placed in a data warehouse database along with metadata. Data marts and query tools exist to report and analyze data, enabling users to produce interactive, ad - hoc, or static reports [11].



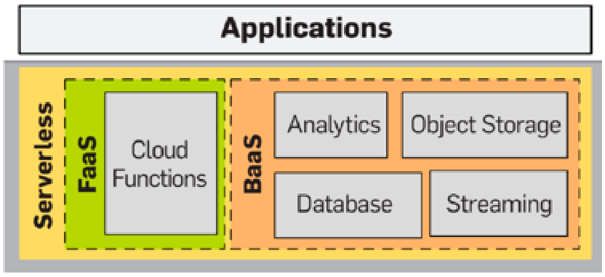
Security and Compliance

It is important to emphasize the security protocols within the cloud space strongly. Encryption should be applied when data is at rest and in motion, access controls and roles should be implemented, and active and detailed auditing must be performed. The cloud solution must also meet compliance standards from the country or jurisdictions where operations are conducted (e. g., GDPR, HIPAA).

Training and Change Management

Key stakeholders should ensure that the organization ushers in the change successfully for the change to be effective. This involves offering orientations on the new cloud platform and any associated changes to procedures, acquiring all necessary written material for the new system and migration process, developing a mechanism for handling users throughout the migration process after the alteration has taken place, and explaining the reasons for changing the system [6].

The figure below shows two technologies that emerged in 1950 and have shaped the computing evolution: containerization in shipping and time - sharing in computing, which brought automation and multitenancy, respectively.



2. Case Studies

Teradata to Amazon Redshift Migration

Big retail firms often encountered troubles with their Teradata system; the problem began to arise with the increasing maintenance cost and architectural immobility. They agreed to shift to Amazon Redshift. During the migration process, the Teradata environment was first evaluated; second, the Schema Conversion Tool for Teradata schemas conversion to Redshift schemas was used; and third, migration was carried out in stages. AWS Data Pipeline was used for the initial data loading, and AWS Database Migration Service was used for replication. For efficiency, the team maximized data distribution and sort keys in Redshift and self - hosted Redshift Spectrum to query data in S3 so that warehousing costs do not occur. Consequently, the company improved the total cost of ownership by 40%, increased the performance of queries up to 3 times for highly analytic workloads, and the ability to quickly match the resource in high traffic periods.

Netezza to Google Big Query Migration

A North American financial services giant required support for high - volume workloads beyond the capabilities of the present Netezza appliance. Google BigQuery became their option, and they decided to migrate. Data type conversions were made from Netezza to BigQuery compatible ones, utilizing Cloud Storage as a temporary extraction layer from Netezza and the BigQuery Data Transfer Service for scheduled data ingesting [15]. The team opted for federated queries and in - database machine learning using Big Query ML to work with the external data. This migration led to reduced cost of managing hardware, live data availability every 15 minutes for critical reports, five - fold improvements in query response times for complex queries, and adoption of new advanced analytics capability, which were not possible on Netezza.

Oracle to Azure Synapse Analytics Migration

The healthcare provider experiences rising expenses and growing challenges in managing an on - premise Oracle DW environment. They decided that they needed to move to Azure Synapse Analytics. The migration process operates azimuth Data Factory, then moves data from Oracle to Azure Blob Storage, employs Azure Synapse SQL pool for massive parallelism, and uses PolyBase to import data from Blober Storage of Azure. For the ETL requirements and the machine learning task, the team leveraged Azure Databricks, while for identity management, the team used Azure Active Directory. The migration reduced the general infrastructure cost of the data warehouse by 60%, increased the frequency of data updates from daily to hourly, improved security with built - in Azure features, and made self - serve analytics possible through the integration of Power BI.

3. Challenges and Mitigation Strategies

Although there is much to gain when migrating to the cloud, organizations may experience several issues. If large volumes of data are transferred, the initial transfer time is long and expensive. This challenge can be addressed by data compression, partial data loading, and data transfer services provided by cloud providers [10]. Indeed, the architecture and structures of legacy systems involve complex interconnections that are hard to understand. One oversight is that implementation dependencies may not correspond to domain dependencies clearly, which creates issues in following them or mapping them back to the domain model; furthermore, phasing may introduce dependencies that are not present in the 'all at once' path or remove ones present in the ‘all at once’ path Depending on the domain model reached and the application of the implementation style, phasing and dependency mapping may help address this problem [14]. There may be differences in query performance, depending on whether the system is on - premise or cloud. To address this, there is a need to adopt best practices on data models and queries to enhance cloud use out of native features. Teams may also be inexperienced in this area, which can be countered by the organization's willingness to establish training or alliances with cloud migration specialists [13]. Several strategies require implementation to support business continuity during migration, including having a good and effective cutover process and perhaps running two sets of systems during migration.

4. Future Trends in Data Warehousing

As organizations move more of their operations to the cloud, several factors that define the future of data warehousing are emerging. Bernstein notes that serverless data warehouses are also increasingly common since these are fully managed and offer variable resources based on usage requirements. Multi - cloud and HYBRID CLOUDILITY Are Becoming Popular As a Result of Strategies Where Different Clouds from Different Providers Are Being Used, or a Combination of Clouds and LOCAL environments Are Employed. One emerging approach is data mesh, which provides a way to organize data as a product created and controlled by domain teams. Thus, AI is applied to query optimization, resource distribution, and predictive maintenance [7]. The need for operations in real - time or near real - time data processing and analysis remains high among organizations. Furthermore, there are features of data governance, lineage, and self - service for finding data assets.

5. Conclusion

Porting existing data - warehousing platforms to cloud infrastructure is not easy yet highly profitable. Following best practices and good practices from different and similar cases could help unleash the full potential of the data in the cloud. It, therefore, needs proper planning and an excellent approach to migration and optimization processes. With the data landscape increasingly complex, flexible, and scalable cloud - based data warehouses are essential to underpin business innovation and competitive advantage. These organizations will thus be better placed to use their data effectively to support strategic management and organizational improvement in the digital business environment.

References

[1] Abdou Hussein, A. (2021). Data Migration Need, Strategy, Challenges, Methodology, Categories, Risks, Uses with Cloud Computing, and Improvements in Its Using with Cloud Using Suggested Proposed Model (DMig 1). *Journal of Information Security*, *12* (01), 79–103. https: //doi. org/10.4236/jis.2021.121004

[2] Alice Elizabeth Matenga, & Khumbulani Mpofu. (2022). Blockchain - Based Cloud Manufacturing SCM System for Collaborative Enterprise Manufacturing: A Case Study of Transport Manufacturing. *Applied Sciences*, *12* (17), 8664–8664. https: //doi. org/10.3390/app12178664

[3] Aljaloud, A., & Razzaq, A. (2023). Modernizing the Legacy Healthcare System to Decentralize Platform Using Blockchain Technology. *Technologies*, *11* (4), 84. https: //doi. org/10.3390/technologies11040084

[4] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). *Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics*. https: //15721. courses. cs. cmu. edu/spring2023/papers/02 - modern/armbrust - cidr21. pdf

[5] Cho, S., Hwang, S., Shin, W., Kim, N., & Hoh Peter In. (2021). Design of Military Service Framework for Enabling Migration to Military SaaS Cloud Environment. *Electronics*, *10* (5), 572–572. https: //doi. org/10.3390/electronics10050572

[6] Debauche, O., Mahmoudi, S., Manneback, P., & Lebeau, F. (2021). Cloud and distributed architectures for data management in agriculture 4.0: Review and future trends. *Journal of King Saud University - Computer and Information Sciences*, *34* (9), 7494–7514. https: //doi. org/10.1016/j. jksuci.2021.09.015

[7] Devan, M., Shanmugam, L., & Tomar, M. (2021). AI - Powered Data Migration Strategies for Cloud Environments: Techniques, Frameworks, and Real - World Applications. *Australian Journal of Machine Learning Research & Applications*, *1* (2), 79–111. https: //sydneyacademics. com/index. php/ajmlra/article/view/80

[8] Dong, H., Yang, X., Wang, W., Cao, Y., Wu, K., Hu, H., Rao, G., Srinivas, K., Samee, S., Venkatesh, K., Dadheech, P., Raja, L., & Yagnik, G. (2021). *A Secure and Efficient Data Migration Over Cloud Computing You may also like Influencing factor analysis of car - sharing demand based on point of interest data RETRACTED: Application and Management of Financial Sharing Under the Background of Big Data Era Lingqi Xue - The Value of Hybrid MTS/MTO Supply Chain Sharing Demand Forecasts under Big Data A Secure and Efficient Data Migration Over Cloud Computing*. https: //doi. org/10.1088/1757 - 899X/1099/1/012082

[9] Kahn, M. G., Mui, J. Y., Ames, M. J., Yamsani, A. K., Pozdeyev, N., Rafaels, N., & Brooks, I. M. (2021). Migrating a research data warehouse to a public cloud: challenges and opportunities. *Journal of the American Medical Informatics Association*, *29* (4), 592–600. https: //doi. org/10.1093/jamia/ocab278

[10] Loukiala, A., Joutsenlahti, J. - P., Raatikainen, M., Mikkonen, T., & Lehtonen, T. (2021). Migrating from a Centralized Data Warehouse to a Decentralized Data Platform Architecture. *Lecture Notes in Computer Science*, 36–48. https: //doi. org/10.1007/978 - 3 - 030 - 91452 - 3_3

[11] Nambiar, A., & Divyansh Mundra. (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data and Cognitive Computing*, *6* (4), 132–132. https: //doi. org/10.3390/bdcc6040132

[12] Ponnusamy, S., & Eswararaj, D. (2023). Navigating the Modernization of Legacy Applications and Data: Effective Strategies and Best Practices. *Asian Journal of Research in Computer Science*, *16* (4), 239–256. https: //doi. org/10.9734/ajrcos/2023/v16i4386

[13] Quezada - Gaibor, D., Joaquín Torres - Sospedra, Nurmi, J., Yevgeni Koucheryavy, & Huerta, J. (2021). Cloud Platforms for Context - Adaptive Positioning and Localisation in GNSS - Denied Scenarios—A Systematic Review. *Sensors*, *22* (1), 110–110. https: //doi. org/10.3390/s22010110

[14] Ramalingam, C., & Mohan, P. (2021). Addressing Semantics Standards for Cloud Portability and Interoperability in Multi Cloud Environment. *Symmetry*, *13* (2), 317–317. https: //doi. org/10.3390/sym13020317

[15] Sun, L., & Jin, B. (2023). Improving NoSQL Spatial - Query Processing with Server - Side In - Memory R* - Tree Indexes for Spatial Vector Data. Sustainability, 15 (3), 2442. https: //doi. org/10.3390/su15032442