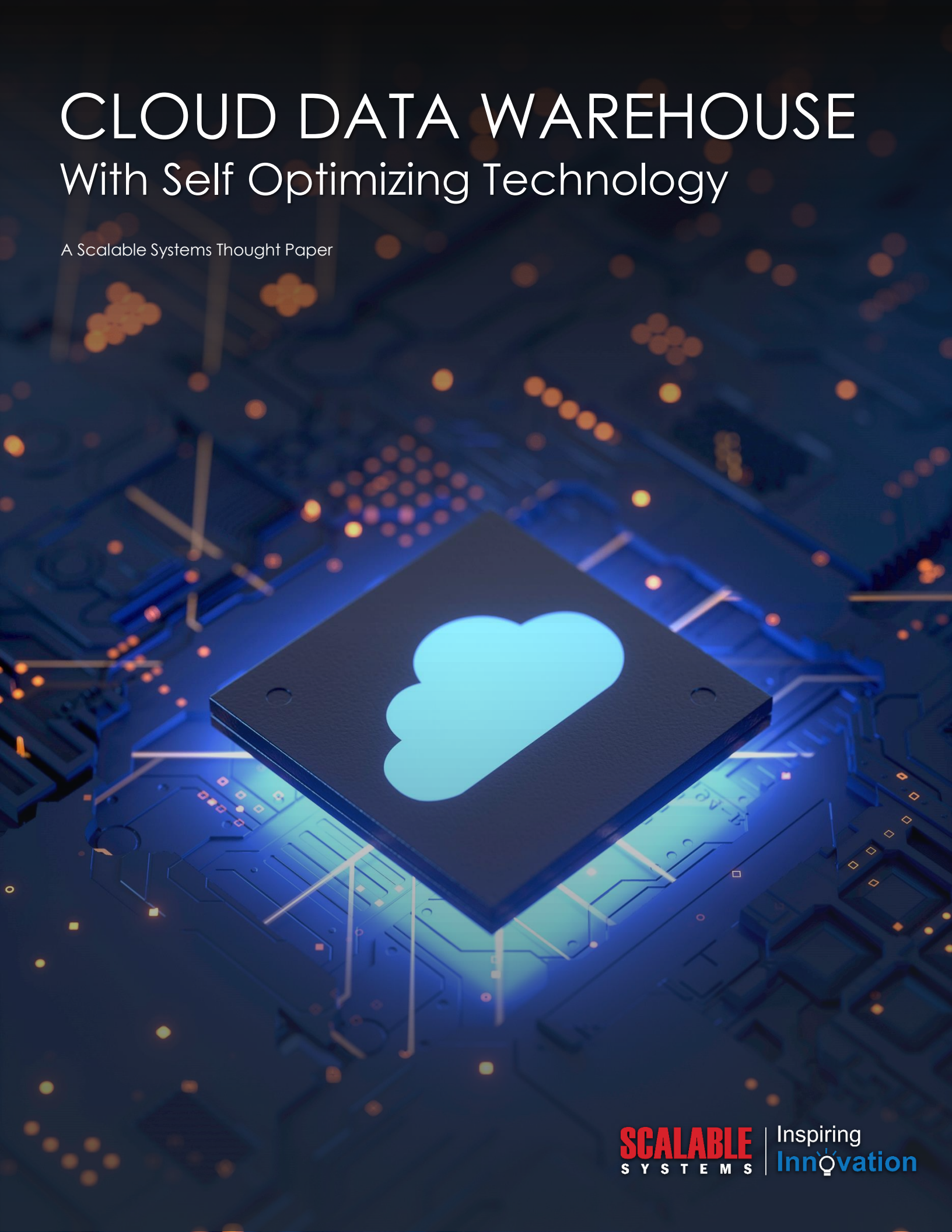
CLOUD DATA WAREHOUSE

With Self Optimizing Technology

A Scalable Systems Thought Paper



SCALABLE SYSTEMS | Inspiring Innovation

TABLE OF CONTENT

INTRODUCTION

Data warehouse is an integral source of information for companies for years and will continue to play a pivotal role in providing insights for decision making. But they are undergoing massive change in terms of technology and adoption. New players are driving the market with innovative solutions that best suits for the customers in terms of scalability, performance, accessibility with less cost compared to on-premise Data warehouse solutions.

After the successful deployment of the warehouse, when the users are able to utilize the data, the focus shifts to the maintenance and growth part. The data warehouse team becomes responsible for providing the users with the required data, views and processing capabilities.

But as the scale and the size of the data warehouse increase the complexity in handling and managing will also increase exponentially. Companies must constantly monitor the connections between the data sources, architecture maintenances, updates and patches, manage security challenges, access management and more.

Cloud-Based Data warehouse solutions solve these issues to a great extent, since companies can shift the responsibility of managing the resources to cloud providers and focus on their core business.

For instance, AWS Redshift and Google Cloud Platform (GCP) are highly scalable cloud-based data warehouse solutions with managed service offerings. In addition, players such as snowflakes offer an innovative cloud solution by understanding customer needs and challenges. But an increasing number of customers are looking for solutions that can provide collaborative features without compromising on security. As their demand increase, we need to investigate the technology frameworks that can self-adapt and self-optimize based on the workload and resource requirements.

SELF-OPTIMIZATION: THE KEY TO SUCCESS

Self-optimization is the key for successful implementation and management of Data Warehouse solutions, given the dynamic nature of the requirements based on business needs. Cloud-based Data warehouse solutions are designed to utilize the self-optimization features in terms of mapping virtual to physical resources and managing data loads to applications based on usage pattern. As technology evolves to meet new data demands, it also provides new areas of opportunity for business growth and operational efficiency.

Today we can observe how automation and machine learning plays a great role in creating a self-optimization system that can automate and make a decision on a variety of functions which as manual in nature. This can be extended to Data Warehouse systems which enable them to tune itself based on the requirements. This capability is accelerating the speed at which data warehouses deliver value to businesses.

Cloud-based Data warehouse systems can employ and exploit the self-optimization features for better user experience.

Areas where Self optimization can be employed

OPTIMIZATION OF DATA QUERIES

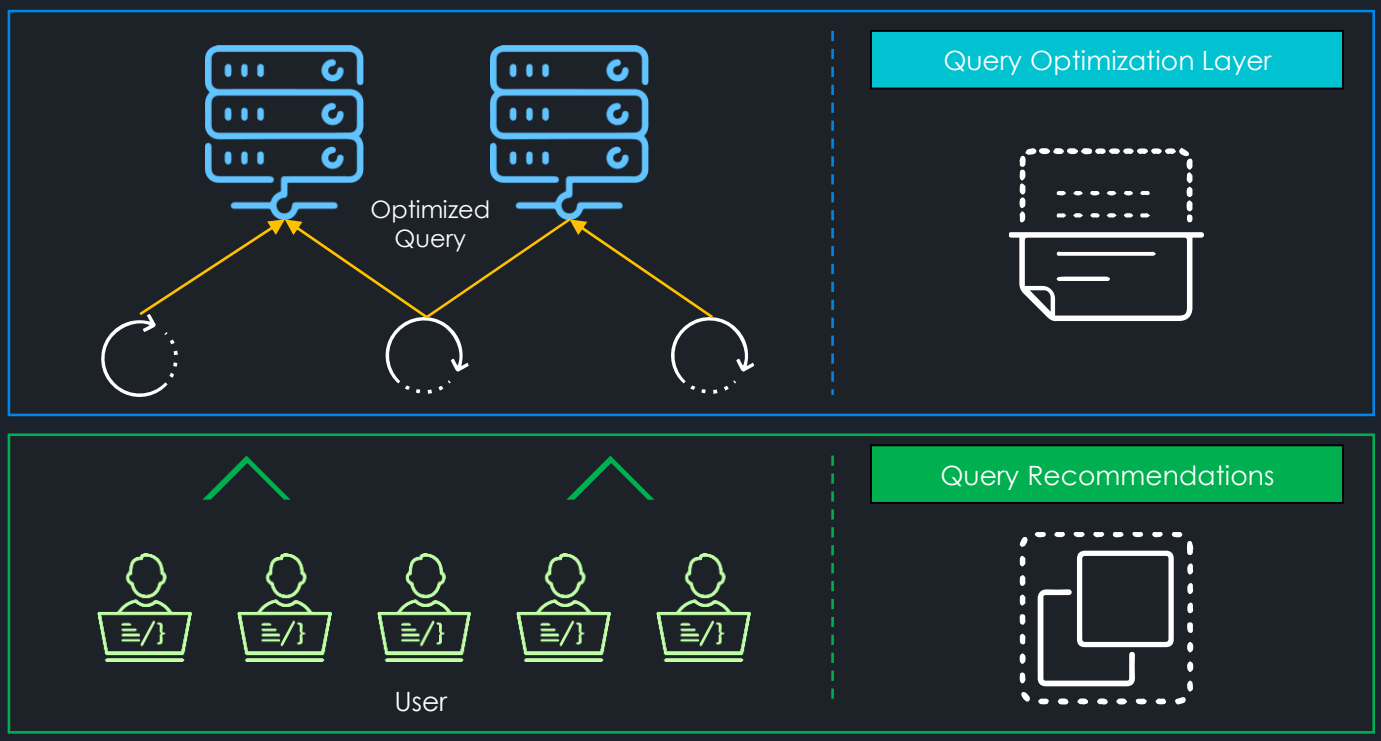
Being the central hub of data for an organization, data warehouse handles millions of queries from the users for extracting data for the business purpose. Some of these queries follow best practice whereas most of them will be ad-hoc in nature which can cause an increased load to the system. An intelligent cloud data warehouse system can analyze these query pattern and optimize them or provide query recommendations to the users that they can use for extraction. This can significantly improve the performance of the Data Warehouse system.

Advanced Machine learning algorithms combined with automation features, analyze the historical pattern of the user queries and identifies the best possible recommendations to the user to efficiently extract the data for their business need.

For example, performing a join on multiple tables are resource intensive and slow. In rare scenarios, poorly written queries based on joins can lead to a crash of the system. To avoid these situations, there should be a mechanism in place which can identify these issues and provide a better query recommendation to the user. ML-based optimization techniques solve this purpose by streamlining the joins between the tables to extract the data, thus reducing the load and increasing the speed of execution.

Many cloud players are started to implement ML-based query optimization techniques to predict and recommend queries to the users based on the activities. These solutions are incorporated into the cloud and be made available to the users based on the subscriptions.

Query Optimization for Cloud Data Warehouse



AUTOMATED RESOURCE ALLOCATION

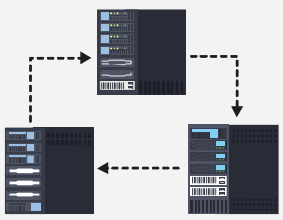
Cloud-based Data warehouse systems offer a flexible and efficient approach to automate resource allocation based on workload demands. This ensures high availability of data source to the consuming applications in downstream. Cloud computing shifts computing from local dedicated resources to distributed, virtual, elastic, multi-tenant resources. This paradigm provides end-users with On Demand access to computing, storage, and software services.

Quality of service (QoS) requirements is an important parameter when allocating resources to the end user. The resource allocation should support quick response time and throughput of a variety of downstream applications. Cloud providers offer a menu of server types with different configurations of CPU capacity, memory, network capacity, disk I/O performance, and disk storage size. An important objective of cloud platform providers is minimizing operational costs while guaranteeing the best resource allocation that meets their client QoS requirements.

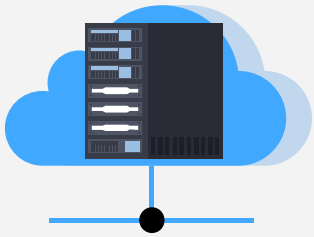
As the demand increases the cloud-based Data warehouse solution should be able to support the client requirements. Auto-scaling is one of the solutions where the cloud platforms can provide necessary resources to the demand, but this is not a viable solution as is may not be cost effective. The solution should focus on self-optimization when allocating resources thus enabling a better resource utilization that can provide better cost benefits as part of the QoS requirement.

Like query optimization, we can enable machine learning based algorithms for resource optimizations. This provides an effective way to analyze the QoS requirement and efficiently optimizes cloud resource allocation.

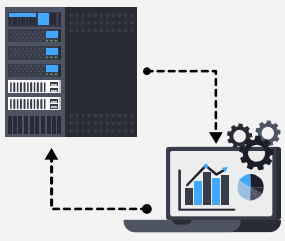
A self-adapting data warehouse will also optimize workloads, where legacy data warehouses could not easily optimize all queries and response times. A self-adapting data warehouse can also streamline and organize data with automatic adjustment of computing power so that it provides the best performance for each user and maximizes the ability to handle concurrent load execution.



Cloud Data Warehouse Service Optimization



Automated Resource Pooling and infrastructure Optimization



End User Optimization

SELF-AWARENESS ON DATA SECURITY

Numerous security breaches are reported across industries, where critical data being compromised. To avoid these situations, a data warehouse system should be able to self-analyze in terms of data security threats and provide early warning to the necessary stakeholders or enable defense mechanisms. Every day a cloud data warehouse will handle millions of transactions which provides a huge resource base for itself to identify the genuine and fraudulent transactions.

We can enable the cloud solutions to learn every day to detect and profile the transactions based on the origin, query request, transaction type and more. When a critical parameter is breached, an automated response should be in place to analyze the type of transaction request in real-time to identify the threat. Customer can avail the services as part of their subscriptions and can enable an On-premises monitoring solution that can work in sync with Cloud data warehouse to identify the threats in transactions.

ML Enabled Data Protection to prevent threats from various sources

ROADMAP FOR THE DATA SUSTAINABILITY

As the cloud technology evolves to handle various customer requirements, self-optimization is a key factor that can enable a self-sustaining data warehouse ecosystem. Customer should focus on the solutions that can provide this important feature as part of their overall cloud data warehouse architecture.

We at scalable solutions are aware of this critical requirement and have developed a centre of excellence in this practice. We work closely with our vendor partner and provide necessary inputs to align the solution that can self-optimize and self-sustain in the long run. Cloud Data warehouse strategy of an organization should focus on a long-term road map to ensure business growth and success. In the age of automation and machine learning, we need to fully capitalize on the opportunities these technologies bring to the table and implement a solution that sustains itself with minimal manual intervention.

About Scalable Systems

Scalable Systems is a Data, Analytics & Digital Transformation Company focused on vertical specific innovative solutions. By providing next generation technology solutions and services, we help organizations to identify risks & opportunities, achieve operational excellence and to gain an innovative edge.
www.scalable-systems.com/

Scalable Systems
Email: info@scalable-systems.com
Web: www. scalable-systems.com