International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Optimizing Costs in Synapse Cloud Analytics: Techniques and Best Practices

Hari Prasad Bomma
Data Engineer, USA
haribomma2007@gmail.com

Abstract: Cloud based analytics platforms, such as Microsoft Azure Synapse Analytics; provide a scalable and flexible solution for data processing and business intelligence. However, the cost of running these analytics workloads can be a concern for most organizations. Inefficient resource utilization, idle clusters, and suboptimal configuration can lead to unnecessary expenses. In this article will see few of the best practices that help in cost optimization. We will also look at few of the case studies that provided better results.

Keywords: Azure Synapse, Azure Data Lake Storage (ADLS), Cost Optimization, Data Warehouse Units (DWU), ETL optimization

Introduction:

Azure Synapse Analytics is a powerful tool that integrates various data storage and analysis capabilities. It combines Azure Data Lake Storage (ADLS) and Blob Storage, allowing storage of huge amounts of both structured and unstructured data. This means one can keep all the data in one place and access it as required. Synapse also offers data warehousing capabilities. This enables to run complex queries and generate insights. It provides integration with popular Business Intelligence (BI) tools, making it easier to visualize and share the findings.

In essence, Azure Synapse acts as a one-stop shop for all the data needs. With its ability to store different types of data and perform analytics, one can streamline the workflows and make data-driven decisions more efficiently. Whether dealing with vast amounts of historical data in ADLS**,** performing real-time analytics on data in Blob Storage, or using BI tools for reporting, Synapse offers the versatility and power needed to handle it all. However if not utilized in a proper fashion, cloud based integration technologies become cost inefficient, and Synapse analytics in no exception. Few common reasons that lead to the higher cost accumulation are over provisioning Data Warehouse Units (DWU) and poor optimization of queries.

Optimization Techniques and Best Practices:

Sometime small insignificant changes also lead to considerable efficient results. Things that are over looked can be corrected to yield better results. Few of the optimization techniques and best practices in Synapse Cloud Analytics that data engineers should consider to increase cost efficiency are listed below:

Right sizing Resources: Over allocation of DWUs can lead to unnecessary costs, where as allocating fewer units will result in performance issues. Monitor and adjust DWUs based on the requirements and workload. Bottom to top approach is best to start with.

Storage Optimization: Choosing a right storage tier is very important in order to achieve cost efficiency. Tiered storage options can be utilized based on the data storage and access requirements. Implement techniques such as indexing, compression, and data partitioning to reduce the storage requirements.

Query Optimization and Tuning: Optimize queries and tune the performance to reduce runtime and

utilization of resources. Well structured quires can have a lot of impact on cost savings. All the best practices of sql query optimization techniques are applicable here.

Server less and Auto pause: Utilize server less options for on-demand query processing. This provides the facility to pay for the compute resources as needed. Implement auto pause to pause compute resources during idle periods.

Monitoring and Anomaly Detection: Continuous monitoring o detect any anomalies is also an efficient way. Utilize Azure Cost Management and Analysis tools to monitor spending trends. Use tool such as Azure Advisor to get personalized best practice recommendations based on cost usage patterns.

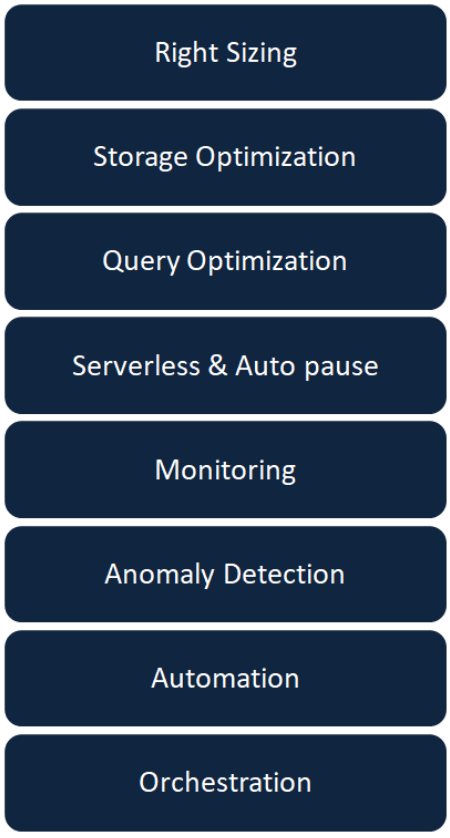

Figure 1: Optimization Techniques

Automation and Orchestration: Utilize automation and orchestration tools to streamline the pipelines for optimization of Synapse Cloud Analytics environment.

Use case scenarios:

Case Study 1: Challenges faced due to long running jobs and delayed reporting. Jobs that usually run in a specific time are taking longer hours.

Job Diagnosis: Azure Synapse has logging feature that helps in tracking performance of jobs and data pipelines. Alert notification mechanism is setup to trigger emails as and when a job runs for more than specific timelines.

Results: Reduced the execution times by identifying bottleneck scenarios leading to efficient utilization of resources.

Case Study 2: Issues were notified regarding speed of data ingestions and processing. Delayed ingestion of data has cascading effects on the underlying time constrained dependencies.

Parallel Processing: Copy activity is one of the most used features. Parallel threads were adjusted to effectively use resources.

Results: Multiple threads processed at the same time decreased ingestion timelines and increased speed helps underlying dependencies.

Case Study 3: Performance issues were faced due to long running queries, data skewness was identified in the logs.

Efficient Indexing and Table distribution methods: Clustered column indexes are used for large Fact tables. Heap tables are used for temporary data loading without indexes to speed up the process, and indexes are applied afterward to improve query performance. Hash distribution is used for large tables, round robin distribution for small or unpredictable tables, and replicated tables for frequently joined small tables to speed up data processing.

Results: Query performances improved and lead to improved speeds.

Case Study 4: Reports require joining large tables (e.g., sales, inventory, and customer data) and aggregating data over multiple dimensions were hampered due to increasing volumes.

Materialized Views: Existing views that were used for report generation were becoming inefficient due to the frequent and resource-intensive data extraction processes. Materialized views in Azure Synapse Analytics are used to improve query performance and reduce the load on the data warehouse. These queries ensure that the necessary aggregations and joins are precomputed and stored.

Results: The precomputed data in materialized views allows for faster query execution and reduced latency in generating reports. By reducing the need for repetitive and resource-intensive computations, the overall load on the data warehouse is decreased, resulting in better resource utilization.

Case Study 5: The data pipeline often experiences significant delays and increased costs due to data transfers across different regions and availability zones. Frequent inter zone data movement resulted in higher egress costs and increased latency, impacting the timeliness and cost effectiveness of data processing tasks.

Method: Aligned the Azure Data Lake Storage (ADLS) and other storage resources within the same region and availability zone. Evaluated the current data architecture and identified storage resources located in different regions and availability zones. Moved the identified resources to the same region and availability zone to minimize data transfer distances.

Results: Reducing inter-zone and inter-region data transfers, significantly lower their egress costs, resulting in substantial cost savings. Aligning storage resources within the same zone minimized latency and enhanced the performance of data processing tasks, ensuring faster data availability and improved analytics.

Literature Review:

The literature review reveals several ways to save costs on cloud-based analytics platforms. One study discusses adapting shared infrastructure (multi-tenant architectures) specific to cloud services while considering various price factors and billing schemes. Another paper presents a combined approach where pricing plans, schedulers, and billing models work together to offer environmentally friendly and cost-efficient options for users. A study also explores how cloud providers can use short and long-term pricing strategies to maximize profits while maintaining quality service.

These insights provide a comprehensive view of cost-saving techniques and best practices that can be applied to Synapse Cloud Analytics. By leveraging strategies such as adapting multi-tenant architectures, implementing greener cloud pricing plans, and optimizing profit maximization methods, organizations can achieve significant cost savings in cloud environments.

Conclusion:

In conclusion, Azure Synapse Analytics is an advanced platform that integrates various data storage and analysis capabilities. This makes it a go to solution for businesses looking for efficient data management and analytics services. Its ability to handle vast amounts of both structured and unstructured data, combined with its data warehousing and business intelligence integration, allows organizations to streamline workflows. In turn makes data-driven decisions more effective in nature. However, to avoid cost inefficiencies, it's essential to implement best practices such as right-sizing resources, optimizing storage and queries, and using serverless and auto-pause features.

Continuous monitoring and anomaly detection, vendor-specific optimization, and automation tools play crucial roles in maintaining cost efficiency. Real-life case studies have shown that optimizing job performance, utilizing parallel processing, and efficient indexing and table distribution can significantly enhance data processing speed and resource utilization in Azure Synapse Analytics. Adhering to these strategies, organizations can maximize the benefits of Azure Synapse while minimizing costs, ensuring a seamless and cost-

effective data management experience. Adopting these cost optimization techniques and best practices, organizations can benefit the with high performance synapse analytics environment.

References:

[1]. M. Perron, R. C. Fernandez, D. J. DeWitt, and S. Madden, *"Starling: A Scalable Query Engine on Cloud Functions*," May 29, 2020. doi: 10.1145/3318464.3380609.

[2]. M. Perron, R. C. Fernandez, D. J. DeWitt, and S. Madden, "*Starling: A Scalable Query Engine on Cloud Function Services,"* Jan. 01, 2019, Cornell University. doi: 10.48550/arxiv.1911.11727.

[3]. M. D. de Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, *"Big Data computing and clouds: Trends and future directions,"* Aug. 27, 2014, Elsevier BV. doi: 10.1016/j.jpdc.2014.08.003.

[4]. U. Hohenstein, R. Krummenacher, L. Mittermeier, and S. Dippl, *"CHOOSING THE RIGHT CLOUD ARCHITECTURE - A Cost Perspective,"* Jan. 01, 2012. doi: 10.5220/0003918803340344.

[5]. R. McHaney, *"What Are Cloud Business Concerns?"* p. 119, Apr. 30, 2021. doi: 10.1002/9781119769514.ch6.

[6]. R. T. Kaushik, P. Sarkar, and A. Gharaibeh, "*Greening the compute cloud's pricing plans,"* Oct. 30, 2013. doi: 10.1145/2525526.2525855.

[7]. *D. S. Purwanto, "Cost Optimization for Azure Synapse Analytics," International Advanced Research Journal in Science, Engineering and Technology"*, vol. 9, no. 2, pp. 10-15, Feb. 2022. [Online]. Available: https://iarjset.com/upload/2022/2022-02/B-4.pdf.

[8]. I. Ignatov, *"Serverless Data Processing and Cost Reduction with Azure Synapse," International Journal of Cloud Computing and Services Science*", vol. 11, no. 3, pp3. 183- 190, Sept. 2023. [Online]. Available: https://ijcss.com/archive/v11n3/23/.

[9]. N. Y. Kuo, "Optimizing Performance and Cost in Cloud-Based Data Analytics," *Journal of Cloud Computing Research"*, vol. 7, no. 1, pp. 95-105, Jun. 2023. [Online]. Available: https://jccr.com/vol7no1kuo/23/.