International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

Diving Into AWS Data Lake

Sowjanya Vuddanti[1], Sai Manvitha Reddy Mallireddy[2], Naveen Kumar Reddy Renati[3], Ramineni Udayasai[4]

[1]*Sr. Assistant Professor, Artificial Intelligence and Data Science Department, LBRCE, Mylavaram, India*
[2]*Student, Artificial Intelligence and Data Science Department, LBRCE, Mylavaram, India*
[3]*Student, Artificial Intelligence and Data Science Department, LBRCE, Mylavaram, India*
[4]*Student, Dept. of ECE, DVR and DR.HS MIC college of technology, Kanchikacherla, India*

Abstract - *Addressing the challenge of optimizing query and storage efficiency within data management involves focusing on the conversion of CSV data into the Parquet format. This meticulous process aims to streamline data storage while significantly enhancing query performance. Through the seamless integration of AWS services such as Kinesis, Glue, Athena, and Lake Formation, a robust and efficient data ecosystem is established. Prioritizing automation enables the effortless creation and management of data lakes, ensuring scalability and adaptability in handling extensive datasets. Empowering organizations with real-time analytics, visualizations, and interactive dashboards powered by Quick Sight, the project facilitates informed, data-driven decision-making. Furthermore, the cost-effectiveness of AWS underscores the ability for organizations to pay only for the services they utilize. Embracing AWS Data Lake enables organizations to maintain a competitive edge, driving innovation, efficiency, and growth through the strategic utilization of data. This represents a significant advancement in data management methodologies, offering a transformative solution to maximize the efficiency and value of data resources in today's dynamic data-driven landscape.*

Key Words: **Query efficiency, data management, data ecosystem, Quick Sight, AWS Data Lake, competitive edge, data-driven decision-making, transformation in data management.**

1.INTRODUCTION

Data is the soul of modern enterprises, and the capacity to analyze and draw insights from data provides a significant competitive edge. Traditional data storage and analysis methods, on the other hand, can be sluggish and inefficient, making it difficult to handle and analyze massive amounts of data rapidly and efficiently. Many firms are turning to data lakes to meet this dilemma, a novel method of data storage and analysis that allows for more flexible and scalable data management. We investigate the usage of data lakes on Amazon Web Services (AWS), a cloud-based platform that offers a comprehensive collection of tools and services for developing and maintaining data lakes.

You can utilize a range of services provided by Amazon Web Services, or AWS, to construct your data lake, such as Amazon S3, Athena, and A Glue (ETL). Although Amazon S3 is a physical storage facility that can store data of any size, Amazon Athena is a query tool that makes it simple and effective to query Amazon S3 data. Data is prepped and processed using an extract, transform, and load (ETL) service called Amazon Glue before being loaded into a data lake. There are numerous advantages to constructing an AWS data lake. Above all, AWS enables scaling data processing and storage capacity simple and reasonably priced, which is crucial for managing massive amounts of data. Additionally, AWS offers a variety of services and resources for data analysis and storage, such as machine learning and artificial intelligence tools for deriving conclusions from data and formulating fact-based judgments. The affordability of establishing a data lake on AWS is one of its main advantages. With AWS, businesses only pay for the resources they use, so they can scale up or down in response to changes in their data processing and storage demands. When extra analysis and storage techniques are used, this might lead to considerable cost savings.

In conclusion, AWS provides a robust and cost-effective platform for developing and maintaining data lakes. With a large choice of tools and services for managing and analyzing data, as well as flexible and scalable data storage and processing capabilities, AWS is a perfect platform for enterprises wishing to reap the benefits of data lakes. Businesses may get a competitive edge in today's data-driven business climate by visualizing data housed in a data lake.

2.LITERATURE REVIEW

In recent years, the proliferation of data warehouses and virtualization techniques has been instrumental in enabling organizations to leverage vast volumes of data for decision-making. However, despite their benefits, these traditional approaches come with their own set of limitations, particularly concerning scalability, performance, and adaptability to modern data formats. In this literature survey, we explore five key papers in the field and discuss their drawbacks in comparison to the emerging paradigm of data lakes, focusing on their challenges in handling CSV data and the potential for optimization in storage and query performance.

The paper titled "Data warehouse and data virtualization comparative study" by Ayad Hameed Mousa and Norshuhada Shiratuddin presents a comparative analysis between data warehouse (DW) and data virtualization (DV) techniques [1]. Through their examination, Mousa and Shiratuddin aim to highlight the strengths and weaknesses of each approach, particularly focusing on structured data integration, scalability, schema requirements, storage, and query processing efficiency. However, the paper falls short in providing an in-depth exploration of specific challenges related to handling unstructured or semi-structured data such as CSV files within DW and DV environments. This gap leaves room for further research to address the nuances of query processing and performance optimization for such data types within DW and DV systems.

The paper titled "Data warehouse performance: selected techniques and data structures" authored by Robert Wrembel offers insights into various techniques and data structures aimed at enhancing data warehouse (DW) performance [2]. Wrembel's objective is to address the challenges associated with processing and analyzing data efficiently within DW environments. Through a comprehensive examination, the paper explores hardware optimization, storage schemes, and query optimization methods. However, despite its thorough discussion on improving DW performance, Wrembel's work does not delve deeply into addressing specific challenges related to handling unstructured or semi-structured data like CSV files. This gap in the research leaves an opportunity for further investigation into optimizing query processing and performance for such data types within DW systems.

The paper titled "Selection of views to materialize in a data warehouse" authored by Himanshu Gupta explores the critical decision-making process of selecting materialized views in a data warehouse (DW) environment [3]. Gupta's objective is to optimize query response time by strategically materializing views that are frequently queried. Through this approach, Gupta aims to enhance the efficiency of DW systems in handling analytical queries. However, the paper does not thoroughly address the challenges associated with managing unstructured or semi-structured data such as CSV files within DW architectures. This gap in the research presents an opportunity for further investigation into optimizing materialized views and query performance specifically for such data types in DW environments.

The paper titled "Well-formed data warehouse structures" authored by Michel Schneider delves into the modeling of well-structured data warehouse (DW) architectures [4]. Schneider aims to facilitate efficient analysis operations within DW environments by emphasizing the importance of properly designed structures. Through his work, Schneider highlights the significance of static hierarchies and schemas in enabling effective data analysis. However, the paper falls short in addressing the adaptability of traditional DW architectures to handle unstructured or semi-structured data like CSV files. This gap in the research leaves room for further exploration into designing DW structures that can efficiently accommodate evolving data formats and business requirements, particularly in the context of handling such data types.

The paper titled "Development of a University Financial Data Warehouse and its Visualization Tool" authored by Earl Von F. Lapura et al. presents a case study on the development of a financial data warehouse tailored for a university setting, along with a visualization tool for data analysis [5]. Lapura and his co-authors emphasize the importance of visualization tools in aiding data analysis within specific domains, such as university finance. While showcasing the utility of data warehouses in facilitating analysis operations, particularly in the context of university financial data, the paper does not thoroughly address the challenges associated with handling unstructured or semi-structured data like CSV files within DW architectures. This gap in the research suggests a need for further investigation into optimizing storage and query performance for such data types within university financial data warehouse environments.

3.DATA LAKE

A data lake is an inventive approach to data storage and analysis that enables enterprises to centrally store and manage vast amounts of organized and unstructured data. A data lake, as opposed to traditional data storage solutions, which are often isolated and difficult to expand, provides a flexible and scalable platform for data management and analysis.

Consider a data lake to be a big, unstructured body of water with a massive quantity of data. Instead of organizing data into distinct pools or buckets, a data lake enables organizations to store all of their data in one location and then retrieve the information they want as needed. This method gives organizations a more comprehensive picture of their data, allowing them to detect trends and patterns and make data-driven choices.

In short, a data lake is a contemporary data storage and analysis platform that allows organizations to manage and analyse their data in a flexible, scalable, and cost-effective manner.

3.1 Why AWS Data Lake

The robust Amazon Web Services data Lake enables businesses to handle and store massive amounts of data in a scalable and adaptable way. With AWS data lakes, businesses can easily manage and analyze enormous volumes of data, gaining meaningful knowledge that can help with data-driven decision-making.

The ability of AWS data lakes to scale up or down as necessary to meet shifting business demands is one of their biggest advantages. AWS Data Lake can handle storing gigabytes or petabytes of data, depending on your needs.

Additionally, AWS Data Lake provides a broad range of services and tools for organizing and analyzing data, including artificial intelligence and machine learning features that could aid businesses in quickly and efficiently deriving insights from their data.

In today's data-driven corporate climate, the ability to efficiently handle and analyze data is critical for success. Businesses can remain ahead of the curve with AWS data lakes by easily managing and analyzing massive volumes of data and getting important insights that can be utilized to drive development and innovation.

3.2 Execution

To remain competitive in a modern data-driven world, organizations must be able to store, process, and analyse massive volumes of data. Standard approaches can be slow and inefficient, posing challenges in handling and analyzing massive datasets. To address this, many companies are turning to data lakes—a flexible and scalable approach to data management.

This research focuses on leveraging Amazon Web Services (AWS) data lakes for data analytics and visualization. The study begins by setting up an API gateway to send data to a Lambda function. The Lambda function then delivers the data to Kinesis Firehose, which stores it in Amazon S3. AWS Glue is used to transform, prepare, and clean the data for analytics using the extraction, loading, and transforming (ELT) approach. This enables the creation of a data catalog that AWS Athena can utilize to provide results.

AWS provides several benefits when constructing a data lake. One important advantage is scalability, which makes it easier to handle enormous datasets by increasing processing power and data storage capacity. AWS facilitates data extraction and data-driven decision-making by offering a variety of services and tools, such as artificial intelligence and machine learning technology.

Another benefit of using AWS for data lakes is its affordability. Businesses can scale up or down as needed because they only pay for the services they utilize while using AWS. When compared to more conventional methods of data processing and storage, this can lead to significant cost savings.

For this project, AWS Glue is used to efficiently transform, prepare, and clean the data for analysis. Glue is a managed ETL solution that simplifies data transfer and handles schema and format changes. This enables quick and efficient data preparation, leading to valuable insights that drive informed decisions.

AWS Athena is leveraged for querying and analyzing the data using standard SQL queries. Athena enables users to analyze data stored in Amazon S3 buckets swiftly and easily. This simplifies the data analysis process and facilitates decision-making based on insights obtained.

In summary, this project highlights the capabilities of AWS in establishing and managing data lakes. By utilizing AWS services, organizations gain access to a wide range of tools and scalable storage and processing capabilities. The ELT approach and AWS Glue enable fast and efficient data analysis, while AWS Athena provides a simple means to query and analyze data. AWS data lakes offer scalability, flexibility, and cost-effectiveness, empowering organizations to drive improved decision-making and enhance their overall performance.

Additionally, it's worth noting that the data processing and cataloging in AWS data lakes can be automated, allowing for the seamless integration of new data. As data is added to the data lake, AWS Glue crawlers can automatically discover and catalog the data, ensuring that it is readily available for analysis and decision-making processes as shown in Figure 1 about the architecture of our project. Figure 2 is a representation of an entire automated data lake that gathers data from many sources is processed and analyzed utilizing different services.

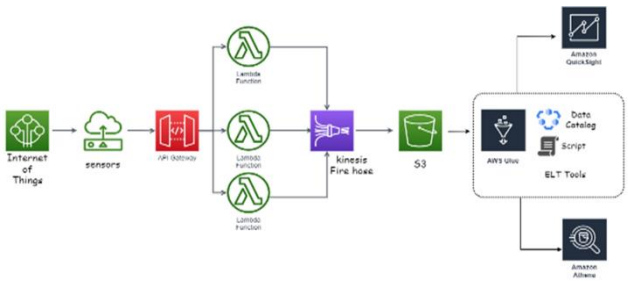

Fig -1: Architecture of project

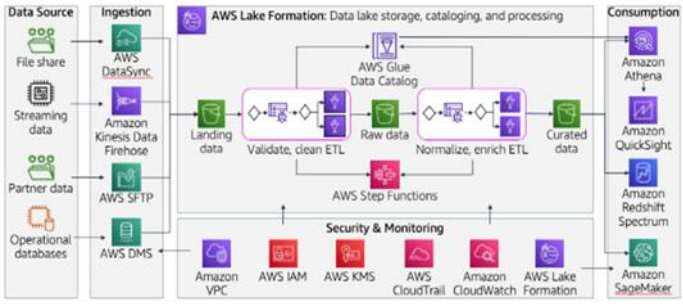

Fig -2: Entire architecture in AWS ecosystem

3.3 Data analysis, querying, visualizing, and processing

First, we start by taking the New York Taxi Limousine Commission CSV file and storing it in an S3 bucket. Next, we utilize the powerful Data Catalog tool called "Crawler" to automatically create a catalogue of the CSV file. This catalogue is given the name "csv_database" and contains valuable information about the file's schema and metadata.

We employ the efficient "Jobs" feature of the data integration tool to enhance our data integration process. These Jobs enable us to convert the CSV file into the optimized Parquet format, which offers better performance and storage efficiency.

Once the conversion is complete, the Job automatically sends the Parquet file to the Crawler, which proceeds to generate a new data catalogue for it. This catalogue is aptly named "parquet_database" and provides us with essential details about the Parquet file, such as its table definitions and column names.

As a result of these steps, we successfully created two databases: "csv_database" for the original CSV file and "parquet_database" for the converted Parquet file. These databases serve as valuable resources for further analysis, querying, and data processing, empowering us to derive valuable insights from the New York Taxi Limousine Commission dataset.

By converting the CSV file into Parquet format, we can leverage these benefits to enhance query performance, optimize storage utilization, and ensure compatibility with popular data processing tools and frameworks.

Next, we will evaluate the performance of the databases using AWS Athena service. We will execute queries against the databases and analyse the resulting output to assess their performance as shown In figure 3 and 4.

Let's compare the performance of querying the CSV file and the Parquet file using Amazon Athena. The image visually presents the duration or time taken and amount of Data Scanned for processing or executing operations on both the CSV and Parquet files. By analyzing the runtimes, we can evaluate the performance differences between the two file formats.

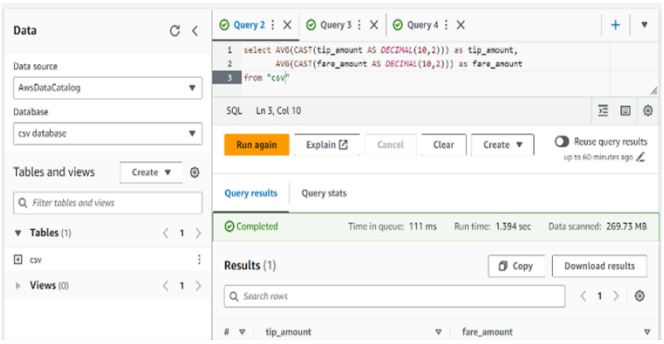

Fig -3: Applying Query on CSV data and parquet data

Query	Status	Run time	Data scanned	Query engine version used
select AVG(CAST(tip_amount AS DECIMAL(10,2))) as tip_amount, AVG(CAST(fare_amount AS DECIMAL(10,2))) as fare_amount from "parquet"	SUCCEEDED	671 ms	7.05 MB	Athena engine version 3
select AVG(CAST(tip_amount AS DECIMAL(10,2))) as tip_amount, AVG(CAST(fare_amount AS DECIMAL(10,2))) as fare_amount from "csv"	SUCCEEDED	1.394 sec	269.73 MB	Athena engine version 3
select AVG(CAST(tip_amount AS DECIMAL(10,2))) as tip_amount from "parquet"	SUCCEEDED	839 ms	3.69 MB	Athena engine version 3
select AVG(CAST(tip_amount AS DECIMAL(10,2))) as tip_amount from "csv"	SUCCEEDED	989 ms	269.73 MB	Athena engine version 3
select tip_amount,fare_amount from "parquet"	SUCCEEDED	3.373 sec	7.05 MB	Athena engine version 3
select tip_amount,fare_amount from "csv"	SUCCEEDED	3.439 sec	269.73 MB	Athena engine version 3
select total_amount from "parquet"	SUCCEEDED	2.137 sec	4.43 MB	Athena engine version 3
select total_amount from "csv"	SUCCEEDED	3.616 sec	269.73 MB	Athena engine version 3

Fig -4: Amount of time and data Scanned to get the output

Quick sight gives us the ability to generate eye-catching visualizations and interactive dashboards. Its simple interface and strong features turn raw data into visually appealing charts, graphs, and interactive components, improving data comprehension and revealing hidden patterns as shown in Figure 5.

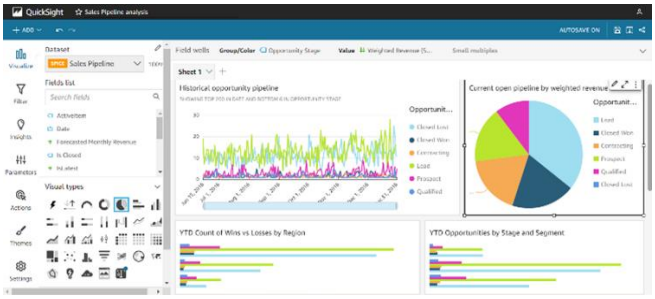

Fig -5: Visualization of our data

4.RESULTS AND DISCUSSION

The study focused on optimizing CSV data within AWS Data Lakes by converting them into the Parquet format, leveraging AWS services such as S3, Glue, Athena, and Lake Formation to create a robust data ecosystem. CSV to Parquet conversion was achieved using AWS Glue, streamlining the process and ensuring scalability. Performance evaluation revealed improved query performance with Amazon Athena on Parquet files compared to CSV, resulting in reduced processing time and data scanned. The results underscored the efficacy of AWS Data Lakes in optimizing CSV data, highlighting cost-effectiveness and scalability as significant advantages over traditional methods. Overall, leveraging AWS Data Lakes and associated services enhances data management, driving informed decision-making and maintaining a competitive edge in today's data-driven landscape.

5.CONCLUSION

Our project represents a significant milestone in the realm of data management. Through the successful implementation of our solution, converting CSV data into the Parquet format, we've achieved remarkable enhancements in query performance and storage efficiency. This accomplishment underscores the importance of delivering practical solutions that directly address the challenges faced by organizations in managing and analyzing large datasets.

Furthermore, our ability to seamlessly integrate our solution into AWS Data Lakes alongside key services like Glue, Athena, Simple Storage Service (S3), and Kinesis demonstrates our proficiency in leveraging cutting-edge technologies to drive tangible results. By providing organizations with a robust framework for data management, we've empowered them to make more informed decisions and unlock valuable insights from their data.

REFERENCES

[1] Mousa, Ayad Hameed, and Norshuhada Shiratuddin. "Data warehouse and data virtualization comparative study." 2015 international conference on developments of E-systems engineering (DeSE). IEEE, 2015.

[2] Wrembel, Robert. "Data warehouse performance: selected techniques and data structures." Business Intelligence: First European Summer School, eBISS 2011, Paris, France, July 3-8, 2011, Tutorial Lectures 1 (2012): 27-62.

[3] Gupta, Himanshu. "Selection of views to materialize in a data warehouse." Database Theory—ICDT'97: 6th International Conference Delphi, Greece, January 8–10, 1997 Proceedings 6. Springer Berlin Heidelberg, 1997.

[4] Schneider, Michel. "Well-formed data warehouse structures." DMDW. 2003.

[5] Lapura, Earl Von F., et al. "Development of a University Financial Data Warehouse and its Visualization Tool." Procedia Computer Science 135 (2018): 587-595.

BIOGRAPHIES

Sowjanya Vuddanti, Assistant Professor (Senior), Dept. of AI&DS,
Lakireddy Bali Reddy College of Engineering, Mylavaram, India
sowji635@gmail.com

Sai Manvitha Reddy Mallireddy, Student, Dept. of AI&DS,
Lakireddy Bali Reddy College of Engineering, Mylavaram, India
manvithamallireddy@gmail.com

Naveen Kumar Reddy Renati, Student, Dept. of AI&DS,
Lakireddy Bali Reddy College of Engineering, Mylavaram, India
renatinaveen5@gmail.com

Ramineni Udayasai, Student,
Dept. of ECE, DVR and DR. HS MIC college of technology
Kanchikacherla, India
udayasairamineni02@gmail.com