DATA ANALYSIS & VISUALIZATION USING AWS

A PROJECT WORK
Submitted in partial fulfilment of
requirements for the award of the degree of

Bachelor of Technology
in
INFORMATION TECHNOLOGY
By

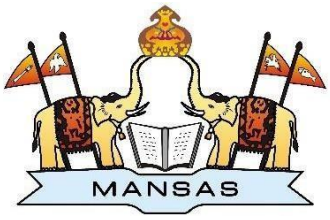
B.Srija
(21331A1214)

N.Lahari
(22335A1206)

K.Prabhath Naidu
(21331A1255)

Under the esteemed guidance of

Dr Anjanadevi B
Associate Professor
Department of Information Technology
M.V.G.R. College of Engineering(A)



Department of Information Technology
MAHARAJ VIJAYARAM GAJAPATHI RAJ COLLEGE OF ENGINEERING(A)
(Affiliated to Jawaharlal Nehru Technological University, Gurajada Vizianagaram)
VIZIANAGARAM
2024 - 25

MAHARAJ VIJAYARAM GAJAPATHI RAJ COLLEGE OF ENGINEERING (A)

VIZIANAGARAM

BONAFIDE CERTIFICATE

Certified that this is a bonafide record of project work entitled "Data analysis & visualization using AWS" ,done by B.Srija (21331A1214), N.Lahari (22335A1206), K.Prabhath Naidu (21331A1255) in partial fulfillment for the award of the degree of "**Bachelor of Technology** " in Information Technology, M.V.G.R. College of Engineering, Vizianagaram , year 2024 – 25.

Dr Anjanadevi B

Associate professor,

Supervisor,

Department of IT,

MVGR College of Engineering(A),

Vizianagaram.

Dr. P. Srinivasa Rao

Professor,

Head of the Department,

Department of IT,

MVGR College of Engineering(A),

Vizianagaram.

External Examiner

DECLARATION

I hereby declare that the work done on the dissertation entitled "**DATA ANALYSIS & VISUALIZATION USING AWS**" has been carried out by us and submitted in partial fulfillment for the award of credits in **BACHELOR OF TECHNOLOGY** in **INFORMATION TECHNOLOGY** of MVGR College of Engineering(Autonomous) affiliated to the Jawaharlal Nehru Technological University (Vizianagaram).

The various contents incorporated in the dissertation have not been submitted for the award of any other degree of any other institution or university.

B.Srija (21331A1214)
N.Lahari (22335A1206)
K.Prabhath Naidu(21331A1255)

ACKNOWLEDGEMENT

As a note of acknowledgement, with great solemnity and sincerity, we offer our profuse thanks to Dr Anjanadevi B, Associate Professor, Dept. of IT, for guiding us all through our project work, giving a right direction and shape to our learning by extending her expertise and experience in the field of education. Really, we are indebted to her for her excellent and enlightened guidance.

Let us lay it on the line we should thank Dr Anjanadevi B, who remained all through our project work as a great source of inspiration for her resourceful counsel and for inducing in us her empirical knowledge that pivoted the goal and destination of our project.

We consider it our privilege to express our deepest gratitude to Dr. P Srinivasa Rao, Professor and Head of the Department for his valuable suggestions and constant motivation that greatly helped the project work to get successfully completed.

We also thank Prof. P S Sitharama Raju, Director and Dr. R Ramesh, Principal, for extending his utmost support and cooperation in providing all the provisions for the successful completion of the project.

We sincerely thank all the members of the staff in the Department of Information Technology for their sustained help in our pursuits.

We thank all those who contributed directly or indirectly in successfully carrying out this work.

B.Srija (21331A1214)

N.Lahari (22335A1206)

K.Prabhath Naidu(21331A1255)

DATA ANALYSIS & VISUALIZATION USING AWS

Team Name

A6

Team

Department: Information Technology

Name: B.Srija ***Email: srijabehara5 @gmail.com*** ***Mobile:8309971252***	
Name: N.Lahari ***Email: niddanalahari @gmail.com*** ***Mobile:6300558941***	
Name: K.Prabhath Naidu ***Email: prabhathkarri0629 @gmail.com*** ***Mobile:7386055602***	

Project Guide

Name : Dr.B.Anjanadevi ***Designation: Associate professor*** ***Email: anjanadevi@ mvgrce.edu.in***	

Objective

This project aims to enhance data analysis and visualization using AWS services like Amazon S3, AWS Glue, Amazon Athena, and Amazon QuickSight. It focuses on efficient data storage, transformation, querying, and visualization to enable faster insights and better decision-making.

Select domain (s) where your solution can be implemented

- Cloud Computing
- Data Analytics
- Business Intelligence

Describe how your solution is going to meet the program outcomes

Our project utilizes AWS services for efficient data processing and visualization. By leveraging AWS tools, it ensures cost effective, automated ETL workflows and interactive dashboards. This project aligns with program outcomes by providing hands-on experience with cloud technologies and improving data analysis capabilities.

Describe the engineering solution

This solution leverages AWS services to efficiently manage data analysis and visualization. Amazon S3 ensures secure and scalable data storage, while AWS Glue automates ETL processes, handling data cleaning and transformation. Amazon Athena enables serverless querying, optimizing data retrieval for analysis. Amazon QuickSight provides interactive dashboards, offering insights for decision-making. AWS IAM enforces strict access control, ensuring secure authorization and authentication.

End Users / Stake holders of solution

- Data analysts
- Business intelligence teams
- Cloud architects & IT Teams

ABSTRACT

Data-driven decision-making is crucial in today's business landscape, and cloud-based analytics platforms provide scalable and efficient solutions for handling data. This project, "Data Analysis and Visualization using AWS, focuses on analyzing the Fortune 1000 Companies dataset (2024) to extract valuable business insights. The dataset consists of financial and operational metrics of the top 1000 U.S. companies, including revenue, profit, assets, market value, sector classification, and leadership attributes. To process this data, we utilize Amazon S3 for storage, AWS Glue for data transformation, Amazon Athena for serverless querying, and Amazon Quick Sight for interactive dashboard visualizations. Various analytical queries were executed in Athena to uncover trends, correlations, and key performance indicators across different industries. There by visualized using Quick Sight dashboards, providing intuitive insights into company rankings, profitability trends, and sectoral distributions. This project demonstrates the efficiency of AWS based data analytics in handling data without infrastructure management. The approach ensures scalability, cost-effectiveness, and real-time insights, making it a good solution for enterprise-level analytics and business intelligence.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1 Cloud Computing?

Cloud computing is a technology that provides on-demand computing resources over the internet, eliminating the need for physical infrastructure. It allows organizations to access storage, processing power, databases, and networking services remotely, reducing operational costs and improving efficiency. Key features of cloud computing include scalability, cost-effectiveness, accessibility, security & reliability, and automation & integration for efficient resource management.

1.2 Amazon Web Services(AWS)

What is AWS?

Amazon Web Services (AWS) is a comprehensive cloud computing platform that provides a wide range of on-demand cloud services, including storage, computing power, databases, networking, and machine learning tools. Launched by Amazon in 2006, AWS has become the leading cloud service provider, offering scalable, secure, and cost-effective solutions for businesses of all sizes.

Why Use AWS?

AWS is widely used due to its flexibility, global presence, and extensive suite of services. Some key benefits include:

1. **Scalability** – AWS allows automatic scaling of resources based on demand, ensuring optimal performance.
2. **Cost-Effectiveness** – With a pay-as-you-go model, users only pay for what they use, reducing operational expenses.
3. **Security** – AWS provides built-in security measures, including encryption, firewalls, identity access management (IAM), and compliance certifications.

4. **Reliability** – AWS offers 99.99% uptime with global data centers, ensuring minimal downtime.
5. **Flexibility & Integration** – AWS services can be seamlessly integrated with various tools, databases, and analytics platforms.
6. **Fully Managed Services** – AWS automates many processes like data storage, computing, and analytics, reducing manual effort.

1.3 Problem Definition

As organizations generate massive amounts of data, traditional tools like SQL, Spark, Excel, Power BI, and Tableau struggle to keep up due to scalability limitations, slow processing speeds, and complex infrastructure requirements. SQL and Excel are effective for small datasets but become inefficient with large-scale data, leading to slow query execution. Spark, while powerful, requires complex configurations and high computational resources. Power BI and Tableau, though strong in visualization, lack built-in scalability and automated data integration, making real-time analysis difficult. To overcome these challenges, this project adopts a cloud-based, serverless approach using AWS services—Amazon S3 for scalable storage, AWS Glue for automated ETL, Amazon Athena for fast SQL-based querying, and Amazon QuickSight for dynamic visualization. This solution ensures faster data processing, real-time insights, reduced infrastructure costs, and seamless scalability, enabling organizations to make data-driven decisions efficiently.

1.4 Objective

The objective of this project is to develop a cloud-based, scalable, and cost-effective data analytics and visualization system using AWS services. By leveraging AWS tools, the project aims to improve data storage, processing, querying, and visualization while overcoming the limitations of traditional methods. The key objectives are:

- Enhance data processing efficiency

- Improve scalability and performance
- Enable fast and cost-effective querying
- Provide real-time, interactive visualization
- Minimize operational costs & infrastructure maintenance

The goal is to create a fully cloud-based analytics system that simplifies data management, speeds up insights, and provides a seamless visualization experience for businesses across various industries.

1.5 Existing System

Before the adoption of cloud-based solutions like AWS, organizations relied on SQL databases, Spark, Excel, Power BI, and Tableau for data processing and visualization. These tools struggle with scalability, real-time processing, and high infrastructure costs, making them inefficient for large-scale datasets. SQL databases slow down with complex queries, Spark requires expensive computational resources, and Excel lacks automation. Power BI and Tableau offer strong visualization but lack real-time data integration. These limitations lead to slow data retrieval, high costs, and inefficiencies, highlighting the need for a scalable, automated, and cost-effective solution.

1.6 Proposed System

To overcome the limitations of traditional tools, this project adopts a cloud-based, serverless architecture using AWS for efficient data processing and visualization. Amazon S3 provides scalable storage, eliminating infrastructure management. AWS Glue automates ETL (Extract, Transform, Load) processes, reducing manual effort. Amazon Athena enables fast, serverless SQL-based querying, improving real-time data retrieval. Amazon QuickSight offers interactive visualizations, enhancing decision-making with real-time insights. This fully managed AWS solution ensures scalability, automation, cost efficiency, and real-time analytics, enabling organizations to process large datasets effortlessly while reducing operational overhead.

2. LITERATURE REVIEW

A) Big Data Analytics on AWS Cloud

Mishra and Kumar (2021), in their study published in the *International Journal of Engineering Research & Technology (IJERT)*, explored the evolving landscape of big data analytics in AWS cloud environments. The paper provides a comprehensive overview of AWS services such as Amazon S3, Glue, Athena, and QuickSight, demonstrating their role in enabling efficient data storage, querying, and visualization. The study highlights how cloud-based analytics platforms offer scalability, cost-effectiveness, and ease of deployment compared to traditional data processing systems. Furthermore, the authors discuss the four dimensions of Big Data (Volume, Velocity, Variety, and Veracity) and the challenges associated with managing large-scale datasets in cloud environments.

B) Serverless Data Processing and Visualization using AWS

Caliwag and Caliwag (2021), in their conference paper on *ResearchGate*, presented an AWS-powered data visualization framework leveraging DynamoDB and Lambda functions. The proposed architecture integrates AWS services for real-time data ingestion, transformation, and visualization using QuickSight. The study emphasizes the benefits of serverless computing, including reduced infrastructure overhead and cost efficiency, making it a compelling choice for data-driven applications.

C) Sentiment Analysis using AWS NLP Services

Satyanarayana et al. (2020), in their *IEEE International Conference Paper*, investigated the application of AWS Comprehend for sentiment analysis of voice-based data. The study outlines an automated workflow where speech data is transcribed using AWS Transcribe, processed via AWS Glue and Athena, and visualized in QuickSight. The research underscores the efficacy of AWS-based NLP tools in extracting meaningful insights from unstructured data, particularly in business intelligence applications.

D) Cloud-Based Weather Data Analytics

Wavhale et al. (2020), in their publication in the *International Research Journal of Engineering and Technology (IRJET)*, proposed a real-time weather forecasting and analytics system utilizing AWS Lambda, S3, and API Gateway. The study integrates various meteorological APIs with AWS cloud services to collect, store, process, and analyze weather patterns. The findings highlight how serverless computing and cloud storage improve efficiency in large-scale data processing and decision-making applications.

E) Amazon Athena: A Serverless Data Querying Approach

Kulkarni (2023), in the *International Journal of Computer Trends and Technology (IJCTT)*, examined Amazon Athena's serverless architecture and its impact on big data querying and analytics. The study explores Athena's integration with S3 and Glue, highlighting its ability to perform cost-effective, high-speed SQL-based queries on large datasets. Additionally, the paper discusses troubleshooting techniques for optimizing query performance and cost management in serverless environments.

F) Digital Architecture for Cloud-Based Microgrid Monitoring

Patsidis et al. (2023), in their article published in *Energies (MDPI)*, explored a cloud-hosted data analytics framework for monitoring microgrid operations. The study utilizes AWS Lambda, DynamoDB, and QuickSight to analyze time-series data from energy grids. By leveraging real-time analytics and predictive modeling, the research demonstrates significant cost reductions and efficiency improvements in energy management.

G) AWS-Powered Data Warehousing for Structured Data Processing

Gupta et al. (2023), in their *IEEE International Conference Paper*, proposed an approach for ingesting and visualizing CSV files using AWS services. The research highlights the effectiveness of AWS Glue for ETL (Extract, Transform, Load), Redshift for structured data storage, and QuickSight for visualization. The study emphasizes scalability, automation, and cost reduction in cloud-based data warehousing.

H) AI-Driven Big Data Analytics on AWS

Kaniganti (2018), in the *International Journal of Science and Research (IJSR)*, presented an AI-integrated big data analytics framework utilizing AWS Kinesis, DynamoDB, and SageMaker. The study showcases the potential of machine learning and AI models for predictive analytics, demonstrating their application in finance, manufacturing, and retail industries. The research concludes that AWS-powered AI frameworks enable real-time decision-making and improved business intelligence.

I) Serverless Architecture for Big Data Analytics

Rahman and Hasan (2019), in their *IEEE Conference Paper*, examined the serverless computing paradigm for big data analytics. The study compared traditional and cloud-based architectures, highlighting how AWS Lambda, S3, Glue, and Athena streamline data processing while eliminating infrastructure management complexities. The paper concludes that serverless models significantly enhance efficiency, cost savings, and scalability in data-intensive applications.

3. SYSTEM STUDY AND ANALYSIS

3.1 User Requirements

3.1.1 Functional Requirements

1. **Scalable Data Storage** – The system should handle large datasets efficiently using Amazon S3.
2. **Automated Data Processing** – Data should be cleaned, transformed, and loaded using AWS Glue.
3. **Fast Query Execution** – The system should enable quick SQL-based queries using Amazon Athena.
4. **Interactive Data Visualization** – The processed data should be presented in a clear and dynamic manner using Amazon QuickSight.
5. **Serverless Architecture** – The system should operate without requiring dedicated infrastructure, reducing maintenance costs.

3.1.2 Non-Functional Requirements

1. **Security** – Data should be stored securely with IAM roles, encryption, and access controls.
2. **Cost Optimization** – The system should use pay-as-you-go AWS services to reduce costs.
3. **Backup & Disaster Recovery** – AWS S3 versioning and backup policies should ensure data recovery in case of failures.
4. **Reliability** – The system should ensure high availability and fault tolerance with AWS-managed services.
5. **Performance Optimization** – Athena queries should be optimized using partitioning and compression to reduce query execution time.
6. **Ease of Use** – The system should be user-friendly, allowing analysts to interact with dashboards without complex coding.

3.2 Data Flow Diagram

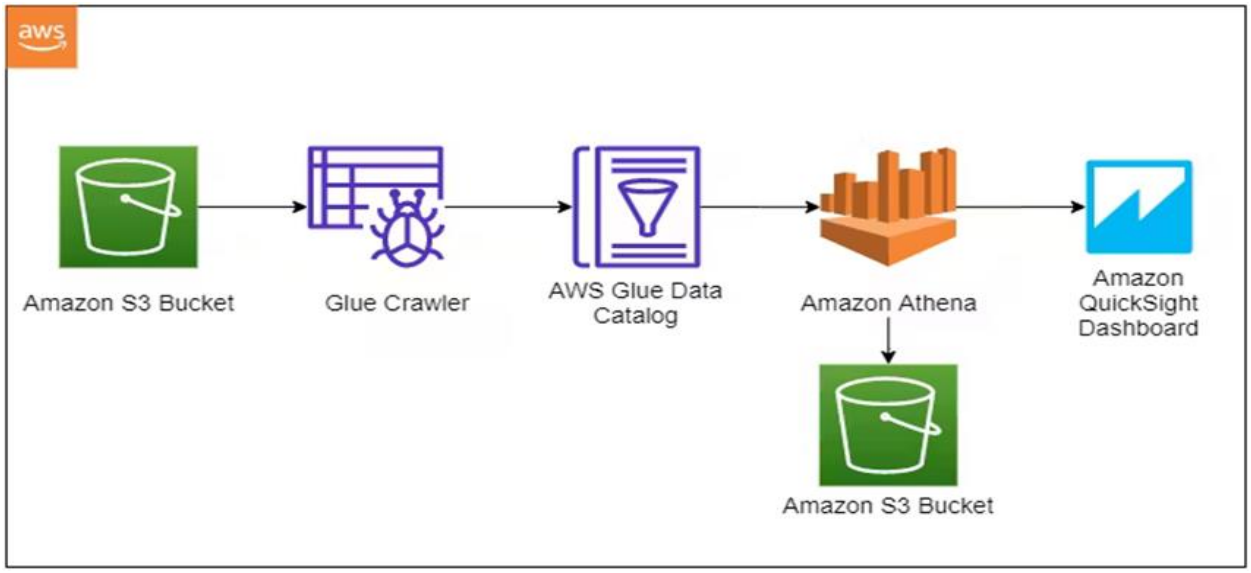


Fig.3.2 *Data Flow Diagram of the Project (Source:[10]).*

The data flow diagram represents the systematic processing of data from ingestion to visualization using AWS services. The workflow begins with Amazon S3, where raw data (CSV, JSON, or other formats) is stored. To enhance query performance, the data is converted into Parquet format, a columnar storage format that optimizes storage and speeds up querying in Amazon Athena.

An AWS Glue Crawler scans the raw and Parquet-formatted data in S3 and updates the AWS Glue Data Catalog, which maintains metadata such as table structure and schema details. This enables Amazon Athena to efficiently query the data using standard SQL. Athena processes user queries and retrieves relevant datasets, which can optionally be stored back in S3 for further processing.

The final step involves Amazon QuickSight, which connects to Athena to generate interactive dashboards and reports, providing insights into the processed data. This pipeline ensures scalability, cost efficiency, and near real-time analytics, leveraging serverless AWS services to streamline big data processing and visualization.

3.3 Hardware and Software Configurations

To ensure optimal performance, scalability, and cost-effectiveness, our system leverages AWS cloud infrastructure. Below are the key hardware and software configurations required for implementing the project.

3.3.1 Hardware Requirements

To support seamless execution and data handling, the following hardware specifications are required:

- **RAM**: Minimum 8 GB (recommended 16 GB for high-performance processing).
- **Processor**: Multi-core processor (Intel i5/i7 or AMD Ryzen equivalent).
- **Storage**: Minimum 500 GB SSD/HDD for local backup and processing.
- **Internet**: High-speed and stable internet connection for accessing AWS services.

3.3.2 Software Requirements

Cloud Services (AWS Stack)

The project is implemented using AWS cloud services, providing serverless analytics and data visualization. The required AWS services include:

- **Amazon S3** – Cloud storage for raw and processed datasets.
- **AWS Glue** – ETL (Extract, Transform, Load) service for data integration.
- **Amazon Athena** – Serverless SQL-based query engine for data analysis.
- **Amazon QuickSight** – Data visualization and business intelligence tool.

System Requirements

- **Operating System**: Windows 10/11, macOS, or Linux.
- **Web Browser**: Latest version of Google Chrome, Mozilla Firefox, or Microsoft Edge for accessing the AWS Management Console.

4. SYSTEM DESIGN

4.1 Architecture

The architecture of our AWS-based solution is designed to handle the Fortune 1000 companies dataset, which includes structured data about companies, their financials, and other attributes. The system leverages Amazon S3 for storage, AWS Glue for ETL (Extract, Transform, Load), Amazon Athena for querying, and Amazon QuickSight for visualization.

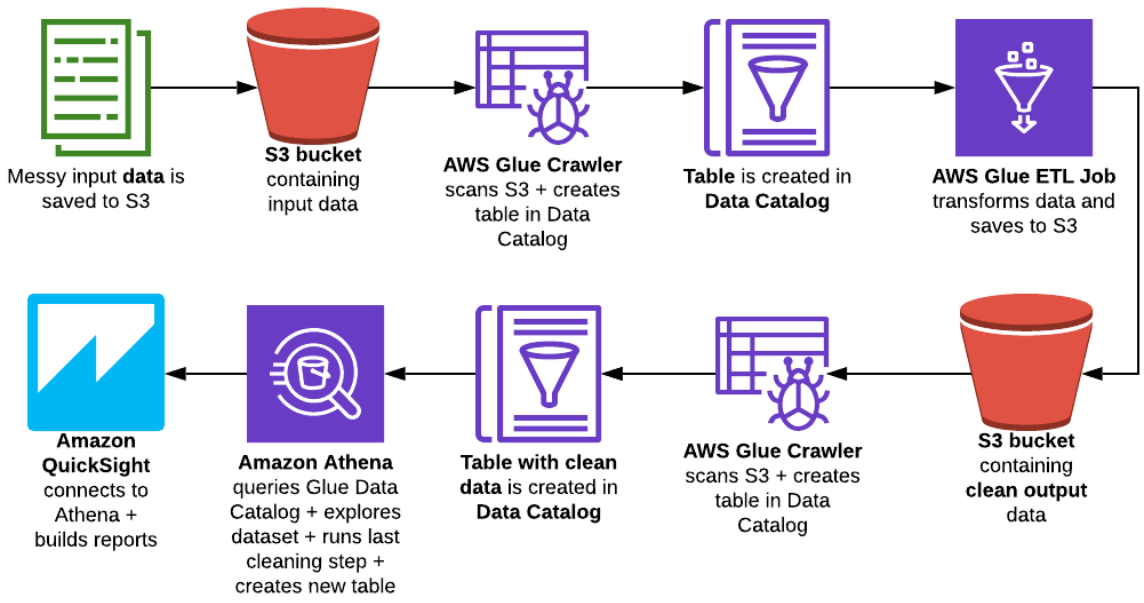


Fig.4.1 *The architecture of the system (Source:[11]).*

4.1.1 Data Ingestion

Description:

Data ingestion is the process of collecting and importing data into the system. In our aws project, the Fortune 1000 companies dataset is ingested into Amazon S3(simple storage service).

It serves as the primary data storage for our project and is the first step in the data pipelining of our project.

The dataset is initially uploaded in raw format (e.g., CSV) to an S3 bucket.

Tools and Services:

- **Amazon S3**: Used as the primary storage for raw data. The dataset is uploaded to an S3 bucket.
- **AWS Glue Crawler**: Automatically scans the raw data in S3 to infer the schema and store it in the AWS Glue Data Catalog.

Process:

1. The raw dataset of fortune companies (CSV format) is uploaded to the S3 bucket.

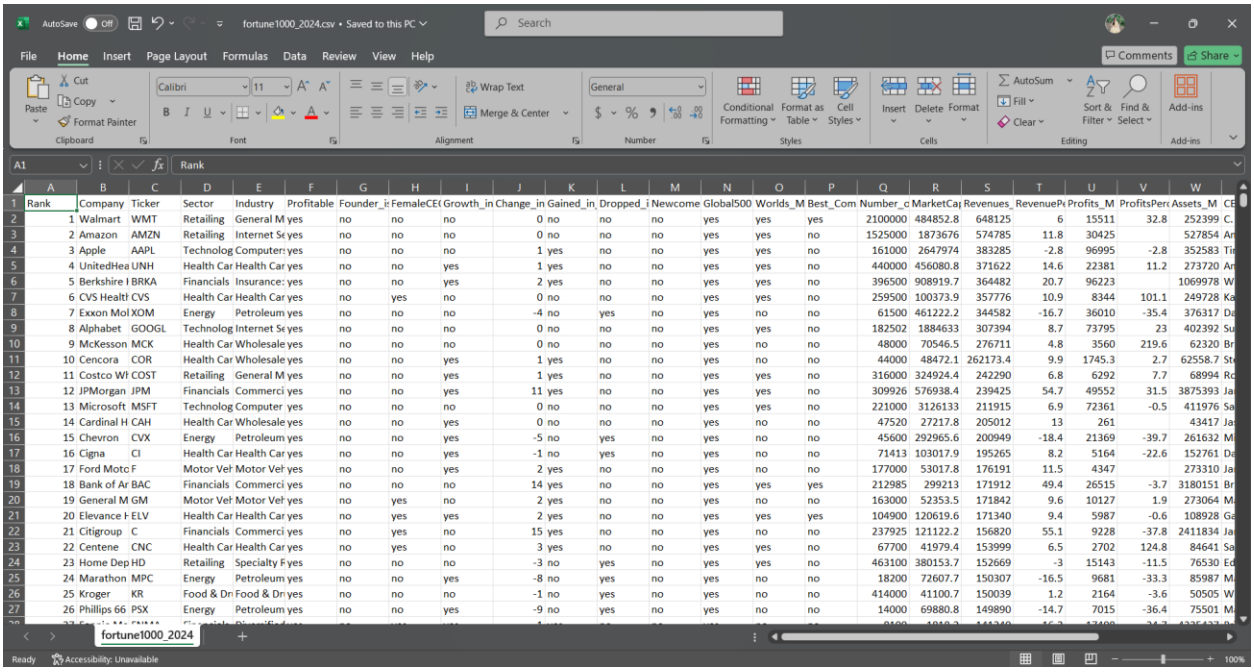


Fig.4.1.1 *The raw dataset in CSV Format (Source:[12]).*

2. An AWS Glue Crawler is triggered to scan the raw data and infer the schema.
3. The schema is stored in the AWS Glue Data Catalog, making the data ready for transformation and querying.

4.1.2 Data Storage and Organization

Description:

The data is stored in Amazon S3 and organized into different layers to facilitate efficient processing and analysis. The storage architecture follows a 4-layer data lake model:

1. **Raw Layer**: Contains the original, unprocessed data (e.g., CSV files).
2. **Curated Layer**: Stores cleaned and enriched data after ETL processing.
3. **Processed Layer**: Contains data transformed into an optimized format (e.g., Parquet) for efficient querying.
4. **Provisioning Layer**: Stores final datasets ready for visualization and reporting.

Tools and Services:

- **Amazon S3**: Used for storing data in all layers (Raw, Curated, Processed, and Provisioning).
- **AWS Glue Data Catalog**: Stores metadata about the datasets, including schema information.

Data Organizing:

- The raw data is stored in the S3 bucket.
- After ETL job processing, the data is transformed into Parquet format and stored in the S3 bucket.

4.1.3 Data Transformation and ETL Process

Description:

The ETL process is responsible for transforming raw data into a format suitable for analysis. In our project, AWS Glue is used to perform ETL operations, including data cleaning, enrichment, and conversion to Parquet format.

Workflow:

1. **Extract**: The raw data is extracted from the fortune1000-raw-data S3 bucket.
2. **Transform**: AWS Glue performs the following transformations:
    - Cleansing: Removing null values and duplicates.
    - Enrichment: Changing data types for better format and analysis.
    - Conversion: Converting the data from CSV to Parquet format for efficient storage and querying.
3. **Load**: The transformed data is loaded into the S3 bucket.

Tools and Services:

- **AWS Glue**: Used for ETL jobs, including data cleaning, enrichment, and conversion to Parquet.
- **AWS Glue Data Catalog**: Stores metadata about the transformed datasets.

4.1.4 Data Analysis and Query Processing

Description:
Once the data is transformed and stored in Parquet format, it is ready for analysis. Amazon Athena is used to run SQL queries on the processed data, enabling fast and efficient data analysis.

Process:

1. The processed data in Parquet format is queried using Amazon Athena.
2. SQL queries are executed to analyze key metrics such as revenue, profit, and market capitalization by sector and industry.
3. Query results are stored temporarily in S3 for further visualization.

Tools and Services:

- **Amazon Athena**: Used for running SQL queries on the processed data.
- **Amazon S3**: Stores query results temporarily.

Query Optimization:

- The use of **Parquet format** and **data partitioning** significantly reduces query execution time and cost.

4.1.5 Data Visualization and Reporting

Description:
The final step in the architecture is data visualization and reporting. Amazon QuickSight is used to create interactive dashboards and reports based on the analyzed data.

Process:

1. Query results from Amazon Athena are imported into Amazon QuickSight.
2. Dashboards are created to visualize key metrics such as:
    - Revenue by Industry
    - Profit by Sector
    - Market Capitalization Trends
3. Users can interact with the dashboards to explore data and gain insights.

Tools and Services:

- **Amazon QuickSight**: Used for creating interactive dashboards and reports.
- **Amazon Athena**: Provides the query results for visualization.

UI/UX:

- The dashboards are designed to be user-friendly, with interactive charts, filters, and drill-down capabilities.
- Users can customize views based on their analysis needs.

4.1.6 AWS Services Integration and Security

Security Measures:

- **AWS IAM (Identity and Access Management)**: Used to define roles and permissions for accessing AWS resources. For example:
    - A specific IAM role is created for the AWS Glue ETL job to access S3 buckets.
    - Another IAM role is created for Amazon QuickSight to query data from Athena.
- **Encryption**: Data stored in S3 is encrypted using AWS Key Management Service (KMS) to ensure security.
- **Access Control**: S3 bucket policies are configured to restrict access to authorized users only.

Service Integration:

- Amazon S3 serves as the central storage for raw and processed data.
- AWS Glue integrates with S3 and Athena to perform ETL and store metadata in the Glue Data Catalog.
- Amazon Athena queries the processed data stored in S3 and provides results to QuickSight.
- Amazon QuickSight integrates with Athena to visualize query results in interactive dashboards.

5. SYSTEM TESTING

5.1 Data Ingestion Testing

The data ingestion process ensures that raw data is successfully uploaded to Amazon S3, maintaining integrity and accessibility for further processing. The Fortune 1000 dataset, stored in CSV format, was uploaded to an Amazon S3 bucket, and its structure and completeness were verified. AWS Glue Crawler was executed to infer the schema, ensuring proper integration with downstream processing tools.

Key Considerations:

- The dataset must be ingested without errors, missing values, or corruption.
- The schema must be accurately inferred to maintain consistency across data processing stages.

Testing Approach:

1. Uploaded the raw dataset to Amazon S3 and verified object metadata, including file format and structure.
2. Executed AWS Glue Crawler to scan the dataset and generate schema information in the AWS Glue Data Catalog.
3. Cross-checked the schema structure with the expected attributes to ensure accuracy.

Observations:

- The dataset was successfully ingested, and all records were intact.
- The Glue Crawler accurately identified column names, data types, and relationships without inconsistencies.
- No data loss or corruption was observed during the ingestion process.

Result: The ingestion process was completed successfully, ensuring data readiness for transformation and analysis.

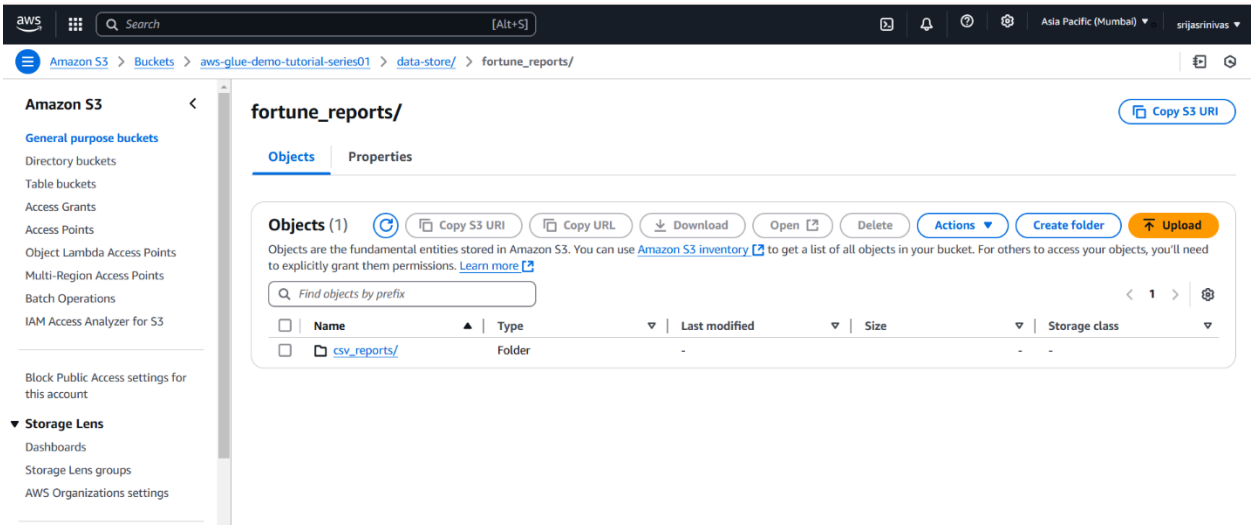


Fig.5.1.1 *The data is uploaded into S3 bucket.*

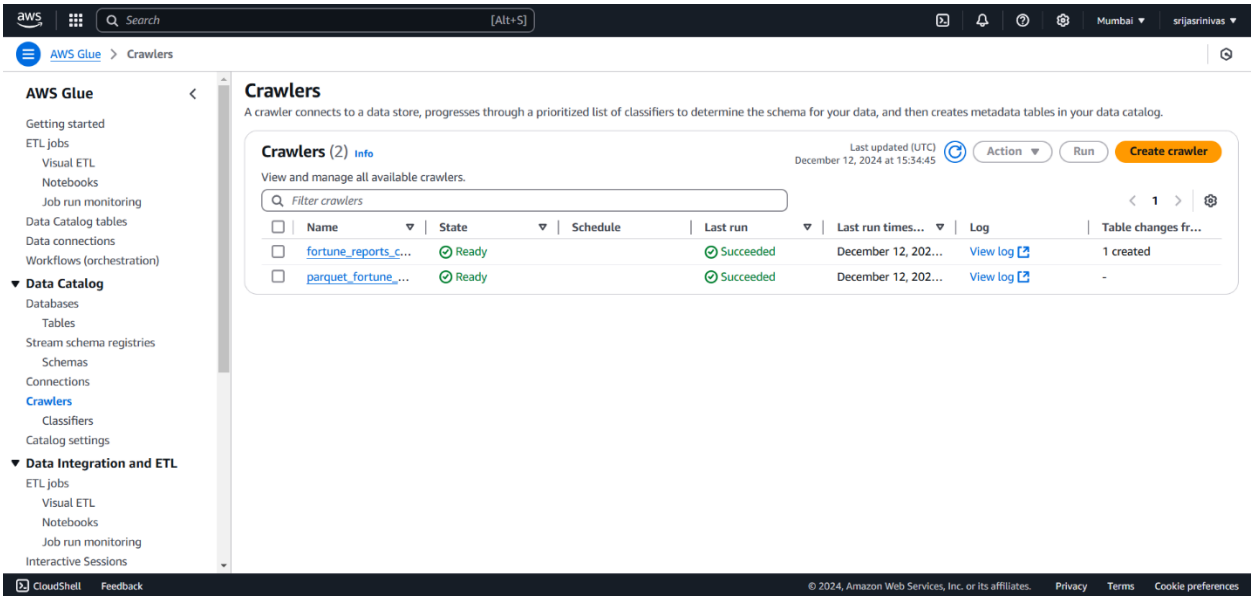


Fig.5.1.2 *Crawler is created in AWS Glue*

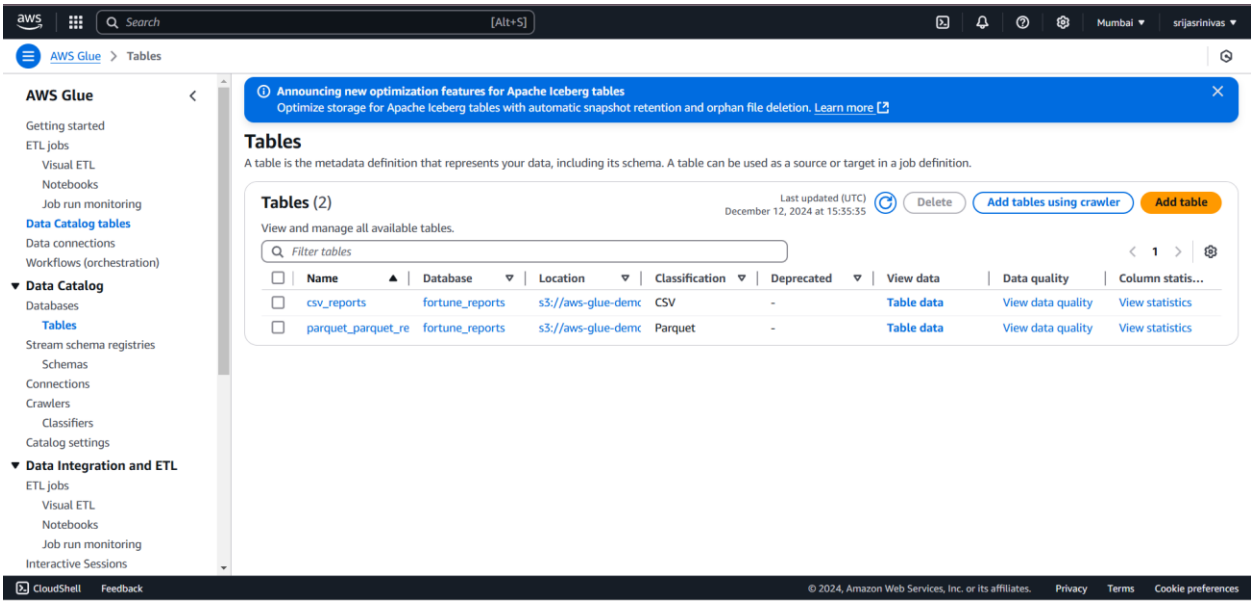


Fig.5.1.3 *The crawler infers the schema and the tables are stored in glue data catalog*

5.2 ETL Process Testing

The ETL (Extract, Transform, Load) pipeline is responsible for structuring and optimizing data for analysis. AWS Glue was used to process the dataset, converting it into a format that enhances performance in subsequent query and visualization stages.

Key Considerations:

- Data transformation must preserve integrity and structure.
- The ETL job should enhance data accessibility while optimizing storage efficiency.

Testing Approach:

1. Executed AWS Glue ETL job to clean, transform, and convert the dataset into an optimized format.
2. Stored the transformed dataset in Amazon S3 and updated metadata in the AWS Glue Data Catalog.
3. Validated the record count and schema consistency between raw and transformed datasets.

Observations:

- The ETL job processed the dataset successfully without introducing inconsistencies.
- The transformed dataset was structured correctly, ensuring compatibility with Amazon Athena for querying. The optimized format improved data retrieval without data loss.

Result: The ETL process was executed successfully, ensuring an optimized dataset for analysis.

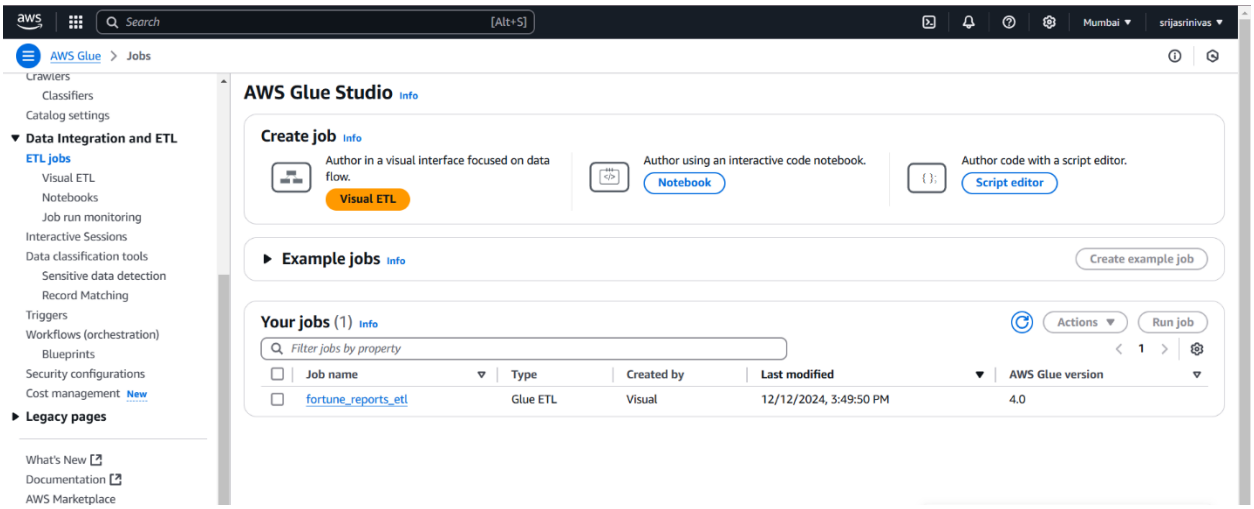


Fig.5.2.1 *The ETL Job is created in glue to transform the data*

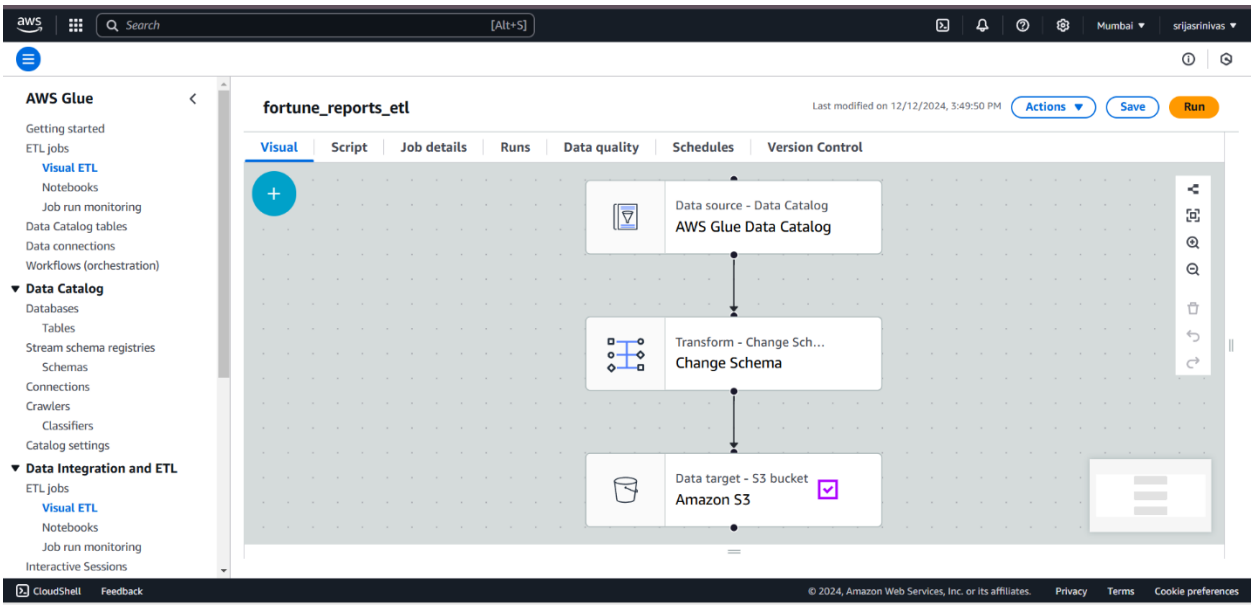


Fig.5.2.2 *The raw data is transformed into parquet format by ETL Job*

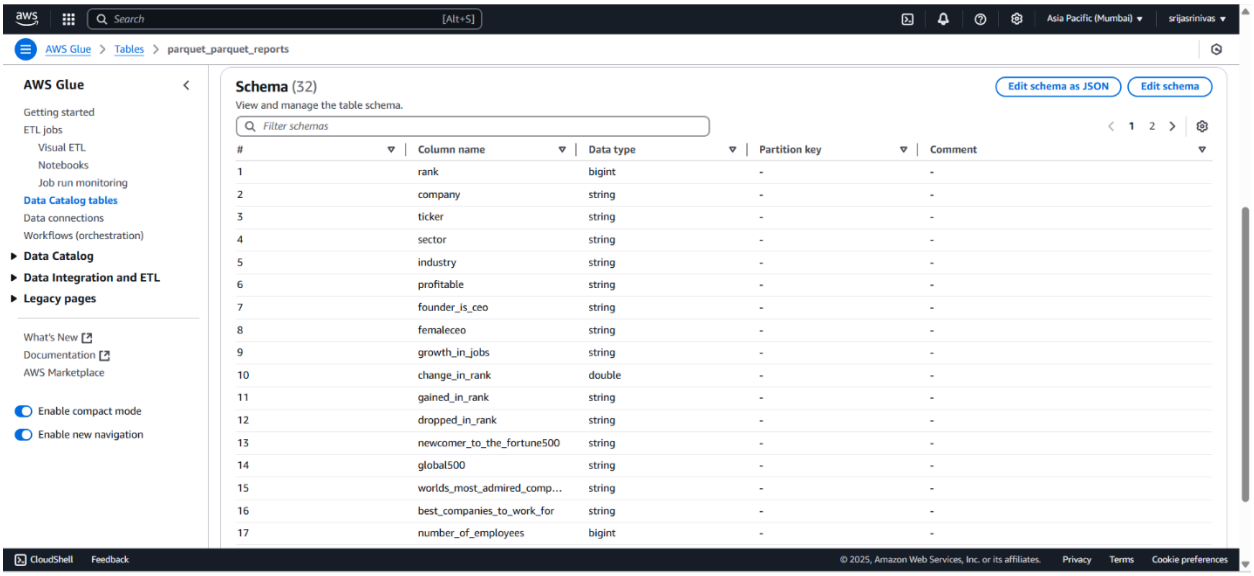


Fig.5.2.3 *The parquet data schema in the glue data catalog*

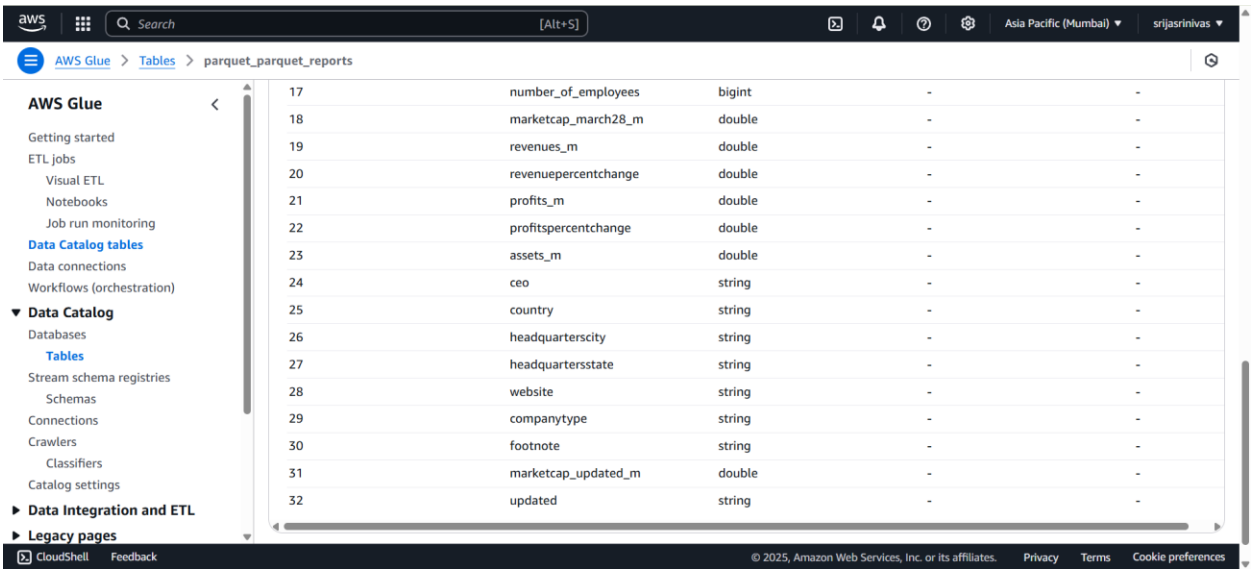


Fig.5.2.4 *The parquet data schema in the glue data catalog*

5.3 Query Processing Testing

Query processing is a critical aspect of data analysis, ensuring that information can be retrieved accurately and efficiently. Amazon Athena was used to execute SQL-based queries on the transformed dataset to extract meaningful insights.

Key Considerations:

- Queries must return correct results without missing or incorrect values.
- Query performance should be optimized for large datasets to ensure efficiency.

Testing Approach:

1. Executed multiple analytical queries to assess revenue, profitability, sector distribution, and other key metrics.
2. Compared query results against expected values to validate accuracy.
3. Assessed query response times and retrieval performance after data optimization.

Observations:

- All queries returned accurate and consistent results, matching expected dataset attributes.
- The optimized dataset structure improved query execution performance, enhancing retrieval speed.
- Queries involving aggregations and filters executed efficiently without processing delays.

Result: The query processing phase demonstrated efficiency in retrieving insights, supporting seamless data exploration.

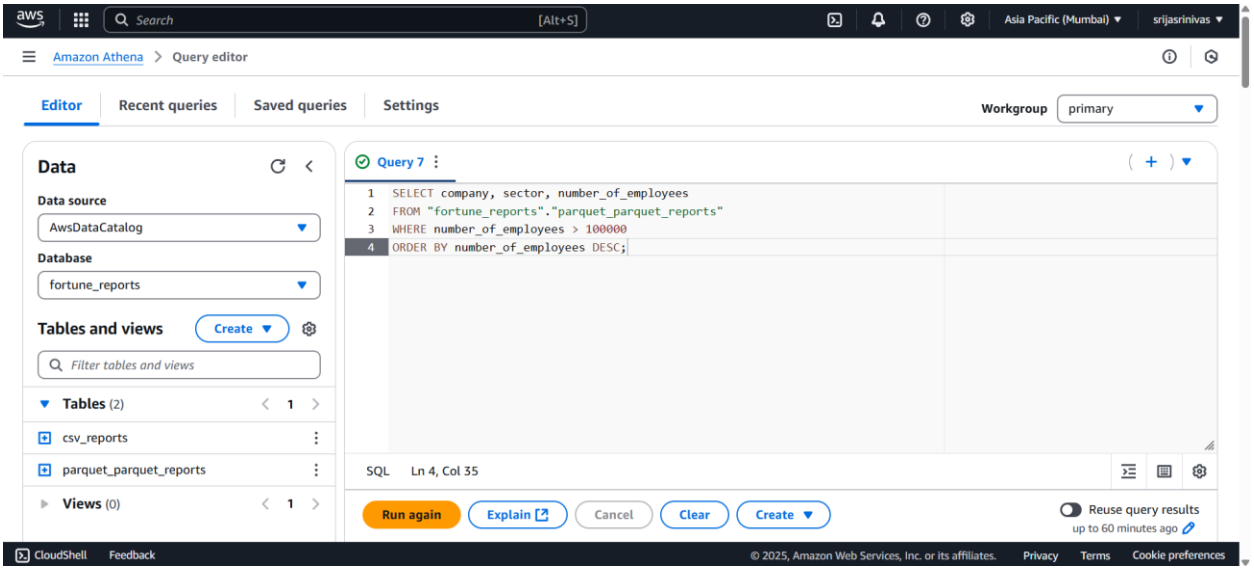


Fig.5.3.1 *This is the query analysis performed in Athena*

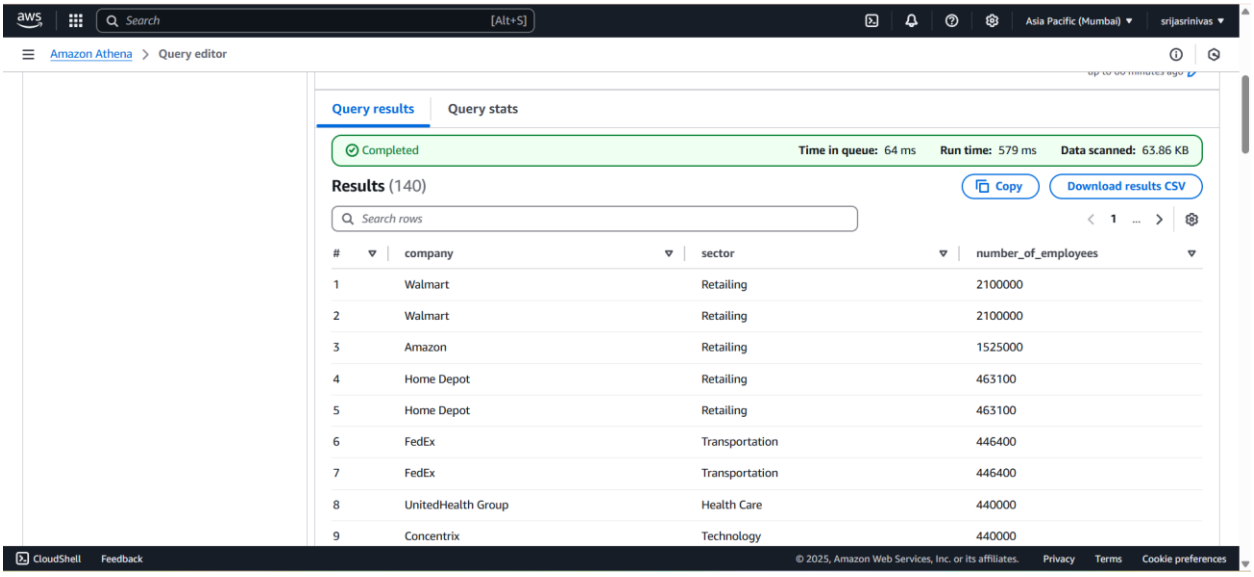


Fig.5.3.2 *This is the query results that are generated in Athena*

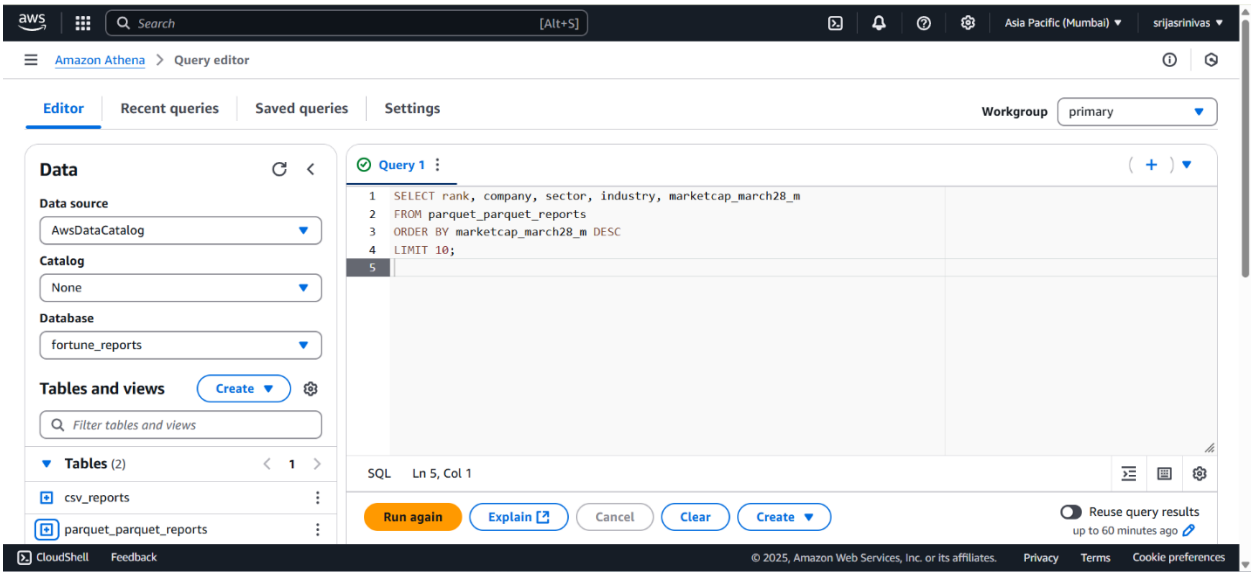


Fig.5.3.3 *This is the query analysis performed in Athena*

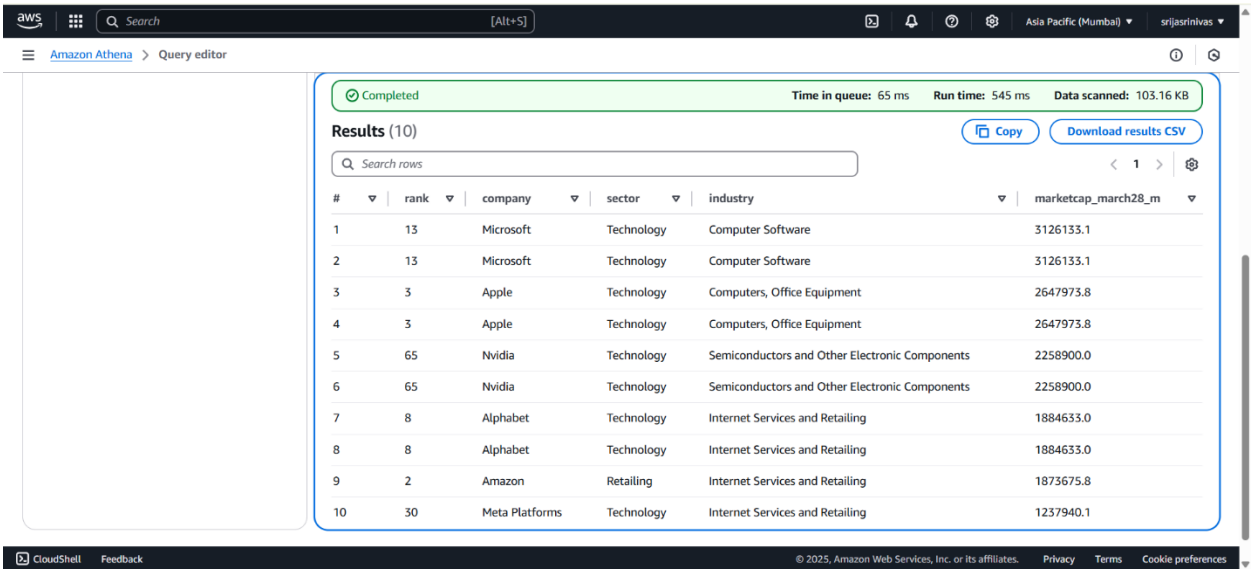


Fig.5.3.4 *This is the query results that are generated in Athena*

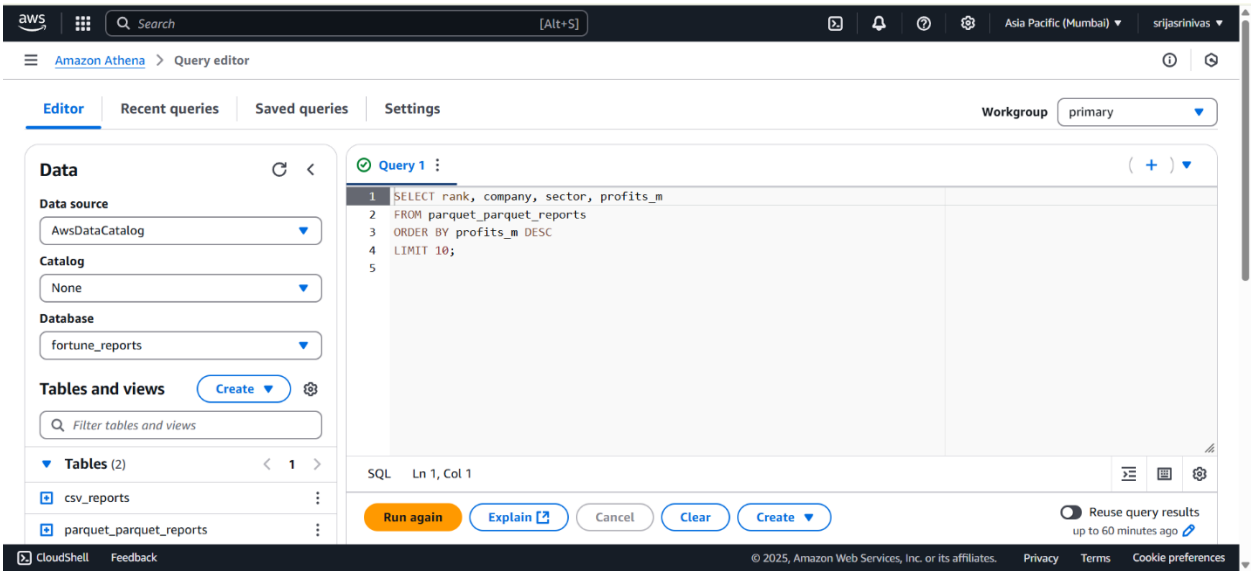


Fig.5.3.5 *This is the query analysis performed in Athena*

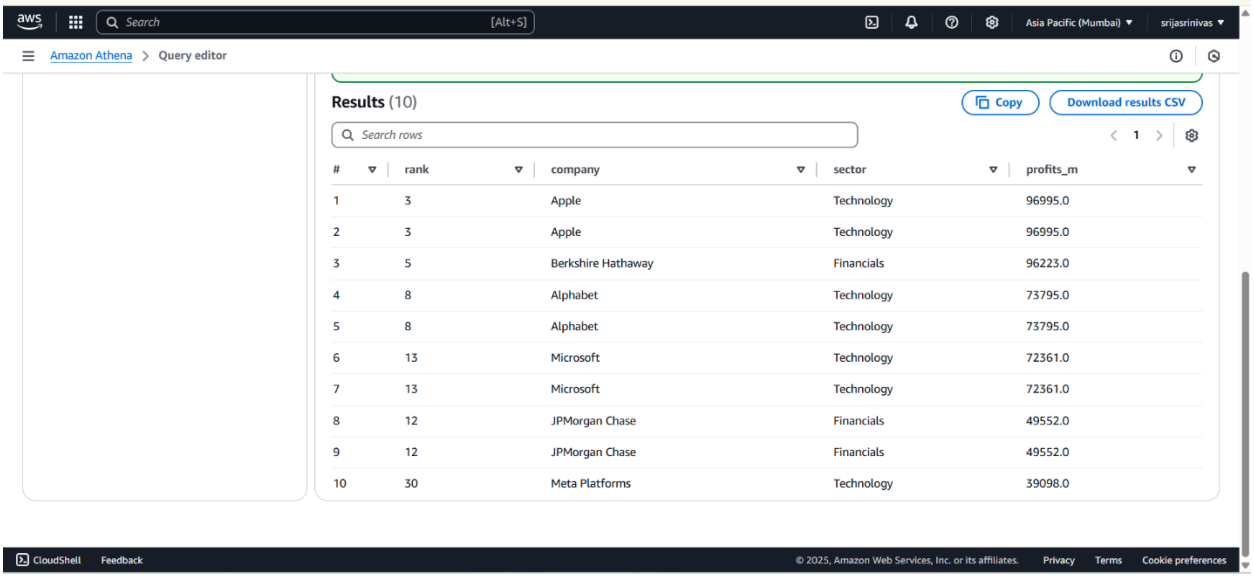


Fig.5.3.6 *This is the query results that are generated in Athena*

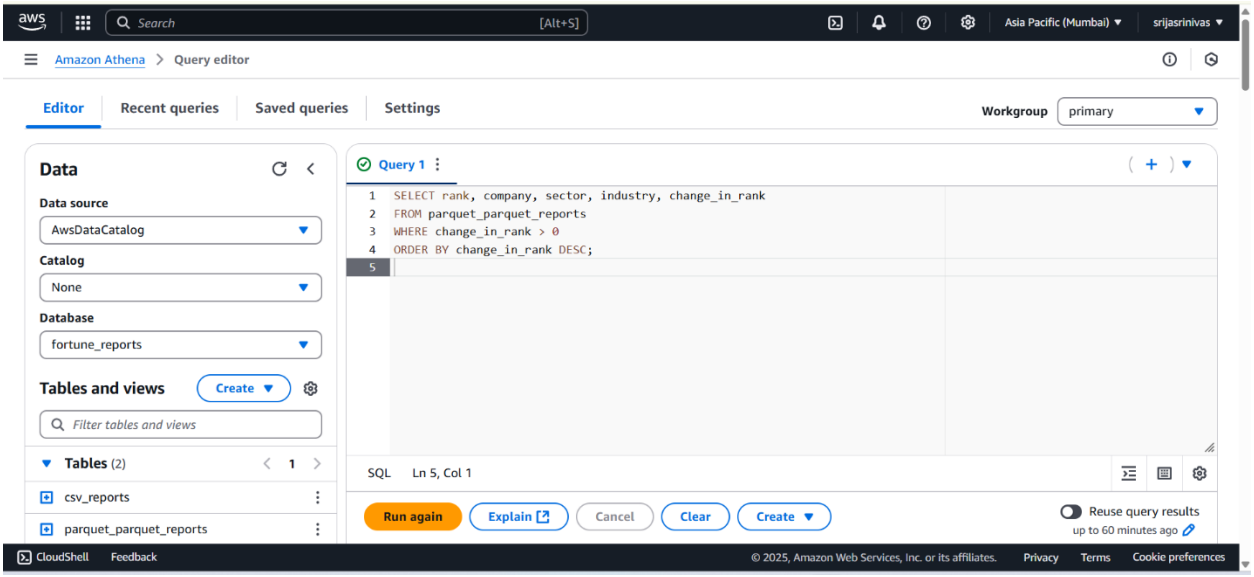


Fig.5.3.7 *This is the query analysis performed in Athena*

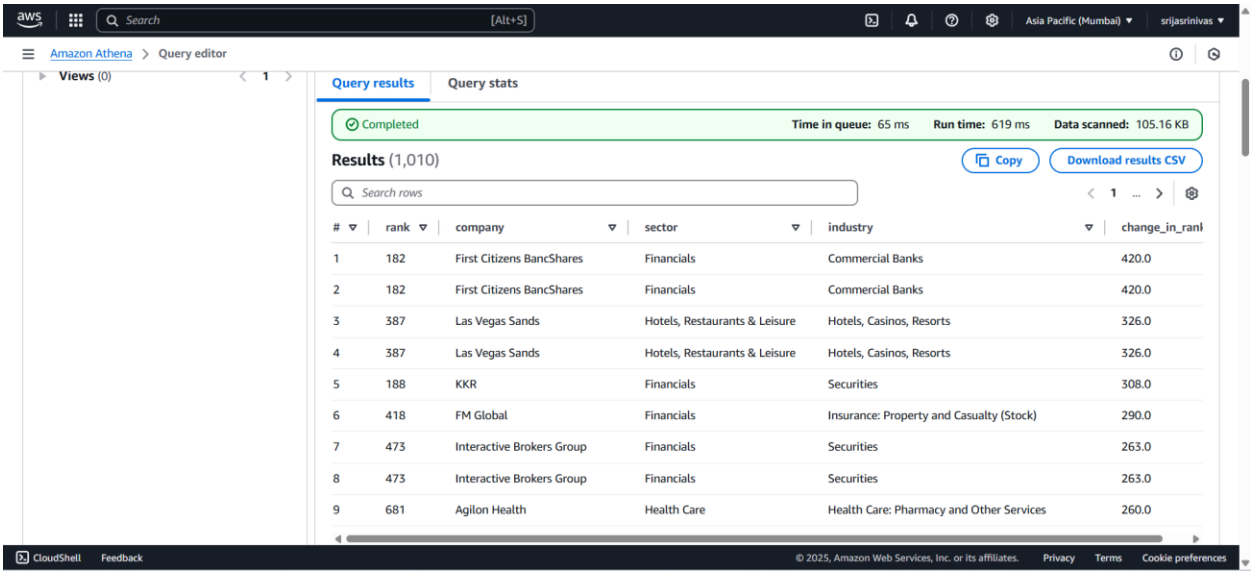


Fig.5.3.8 *This is the query results that are generated in Athena*

5.4 Visualization Testing:

Data visualization is essential for interpreting analytical results effectively. Amazon QuickSight was used to generate interactive dashboards, incorporating various chart types to represent different aspects of the dataset.

Key Considerations:

- Visualizations should accurately reflect data insights without discrepancies.
- The dashboard should support interactive exploration, enabling users to filter and analyze data dynamically.

Testing Approach:

1. Configured Amazon QuickSight to connect with Athena and fetch query results.
2. Created multiple visualizations, including bar charts, pie charts, donut charts, KPIs, area charts, and line charts, to represent different financial and operational metrics.
3. Validated the correctness of visualizations by cross-referencing displayed values with query results.

Observations:

- All visual elements accurately represented the dataset, ensuring alignment with query results.
- The dashboard provided interactive exploration capabilities, allowing users to filter and analyze data seamlessly.
- Performance remained stable, ensuring smooth visualization updates without lag.

Result: The visualization phase successfully translated analytical insights into meaningful graphical representations, enhancing data interpretation and decision-making.

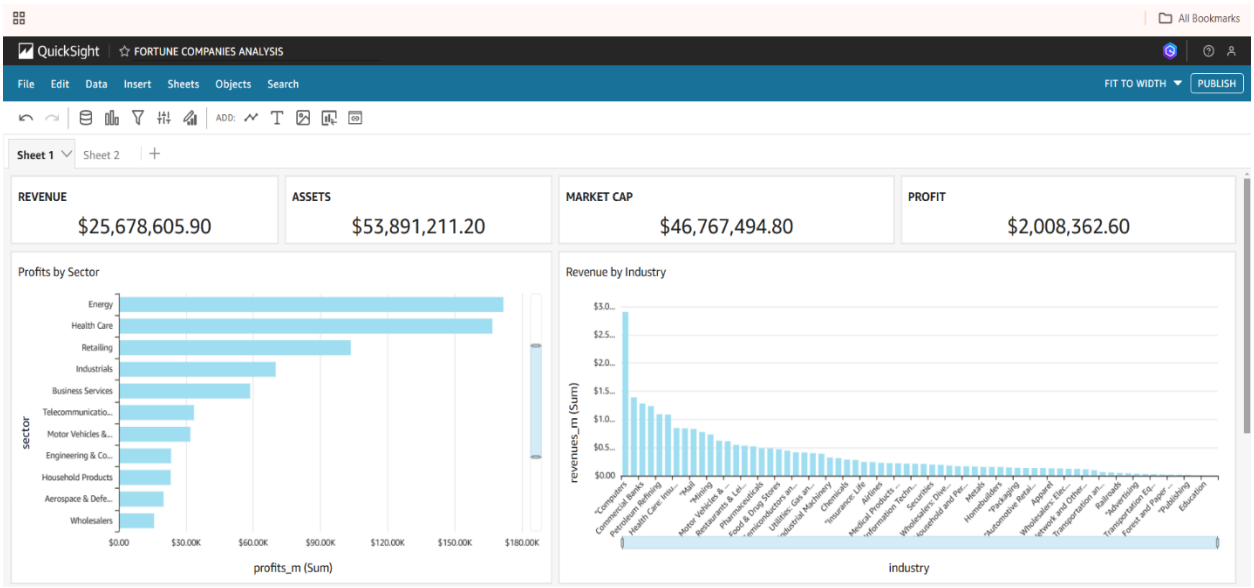


Fig.5.4.1 *The data visualization in Quicksight which represents the analysis of key performance indicators related to companies using bar charts.*

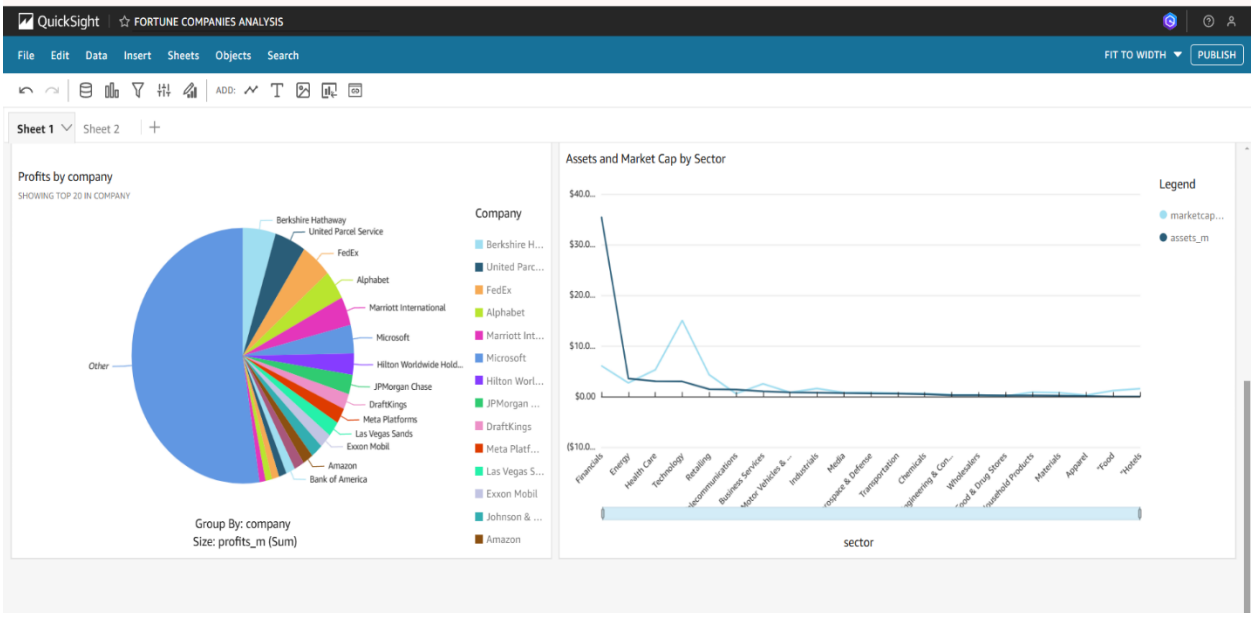


Fig.5.4.2 *This represents the analysis of profits by company and assets, market capital by sector performed in Quicksight using pie chart and stacked area chart.*

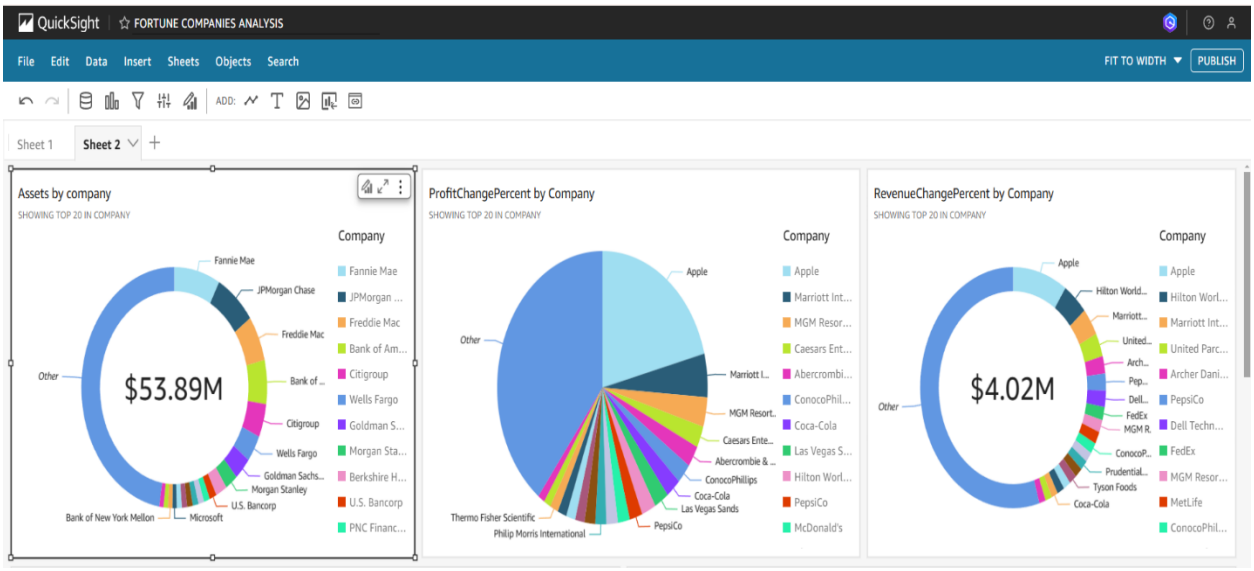


Fig.5.4.3 *This represents the analysis of assets, profits percent change and revenue percent change by company using pie chart and donut chart in Quicksight*

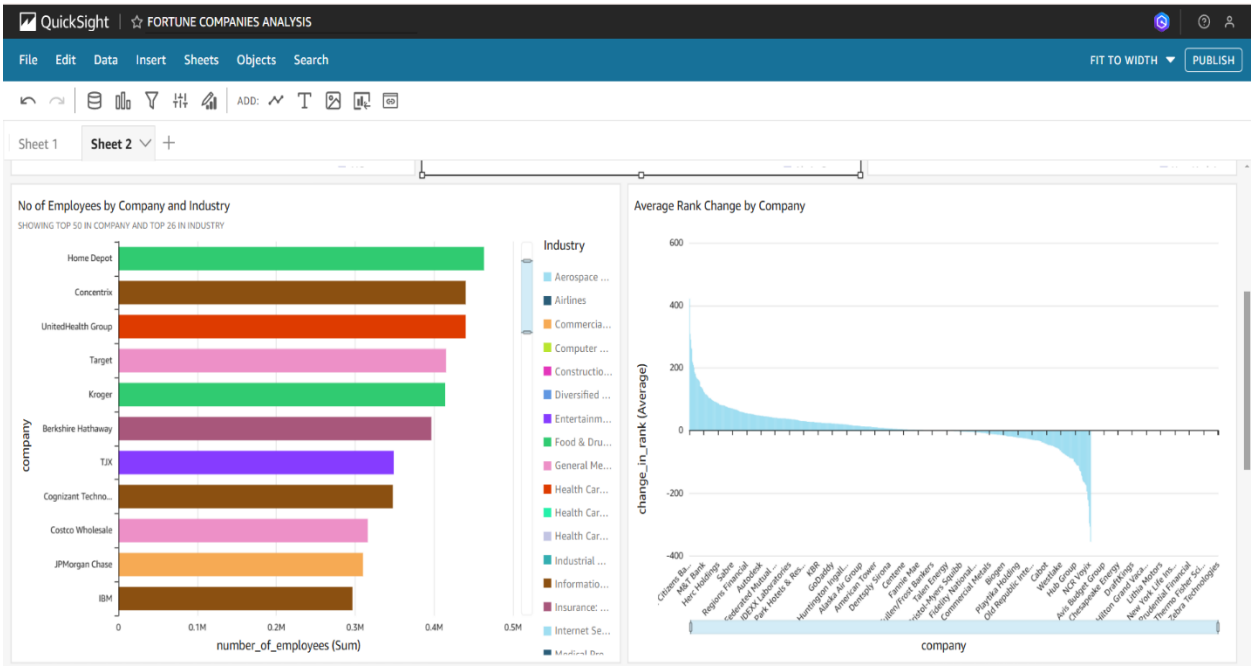


Fig.5.4.4 *This represents the analysis about number of employees by company, industry and change in rank of company that are performed using bar graph and area chart in Quicksight*

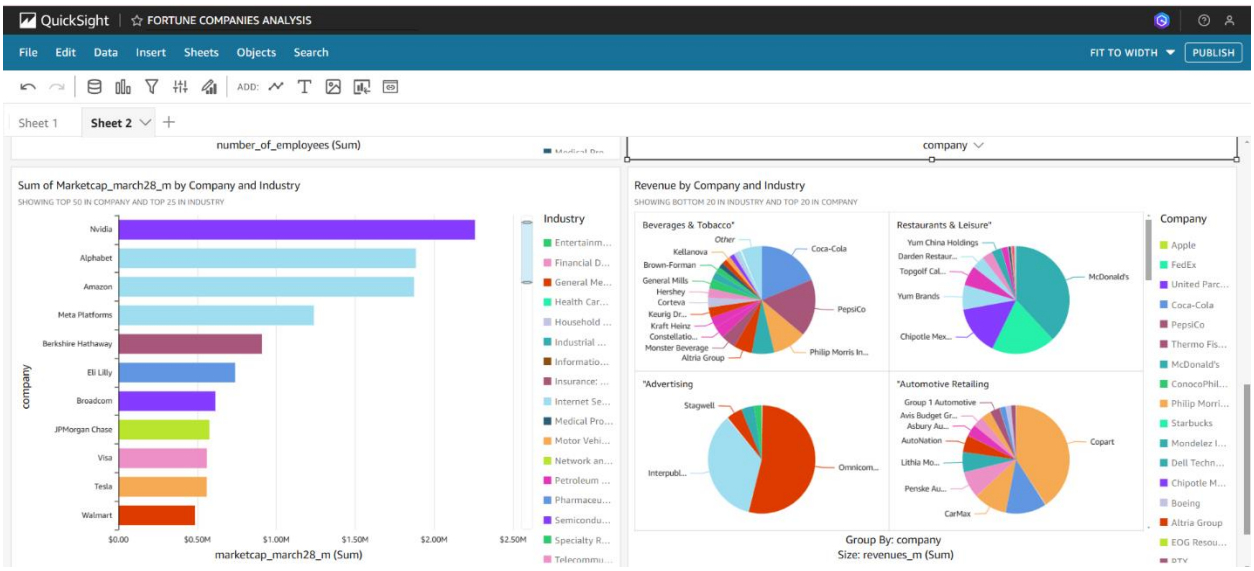


Fig.5.4.5 *This represents the analysis of market capital and revenue by company, industry that are performed using bar graph and pie chart in Quicksight*

5.5 Results & Performance Comparison

5.5.1 Performance Metrics:

- **Data Ingestion**: The AWS Glue Crawler successfully inferred the schema of the Fortune 1000 dataset in under 5 minutes.
- **ETL Process**: The transformation of raw CSV data into Parquet format reduced storage size and improved query performance.
- **Query Processing**: Queries on Parquet data were faster compared to raw CSV data, with an average query execution time.
- **Visualization**: Dashboards in Amazon QuickSight loaded in under 5 seconds, providing real-time insights into the dataset.

5.5.2 Comparison with Traditional Tools:

- Scalability: AWS tools (S3, Glue, Athena) outperformed traditional tools like SQL and Spark in handling large datasets.
- Cost: The pay-as-you-go pricing model of AWS reduced infrastructure costs compared to on-premise solutions.
- Ease of Use: AWS Glue and QuickSight provided a more user-friendly interface.

6. CONCLUSION & FUTURE WORK

6.1 Conclusion

In this project, we explored AWS services such as S3, Glue, Athena, QuickSight. We worked with S3 buckets and objects to store our data, and worked with AWS Glue to create Glue Crawlers and build a Glue Data Catalog. This allowed us to generate tables and define schemas for our dataset. We also configured IAM roles, policies and permissions on who can access the data. We used Athena for querying data and QuickSight for visualizations. This provided us with a solid understanding of foundational AWS services, preparing us to efficiently utilize these tools for scalable data analytics and visualization in aws cloud environment.

6.2 Future Work

- **Real-Time Data Processing:** Integrate Amazon Kinesis for real-time data streaming and analysis.
- **Machine Learning Integration:** Use Amazon SageMaker to build predictive models for forecasting company performance.
- **Advanced Visualizations:** Explore advanced visualization techniques in Amazon QuickSight, such as geospatial analysis and anomaly detection.
- **Multi-Region Deployment:** Extend the solution to support multi-region data storage and processing for global datasets.
- **Cost Optimization:** Implement cost-monitoring tools like AWS Cost Explorer to further optimize resource usage.

7. REFERENCES

[1] Mishra, A. and Kumar, G. (2021) 'Big Data Analytics on AWS Cloud Using AWS Athena and QuickSight', *International Journal of Engineering Research & Technology (IJERT)*, 10(4). Available at: www.ijert.org.

[2] Caliwag, E.M.F. and Caliwag, A. (2021) 'AWS Data Visualization using DynamoDB and Lambda', *Conference Paper*, July 2021. Available at: ResearchGate.

[3] Satyanarayana, G., Bhuvana, J. and Balamurugan, M. (2020) 'Sentiment Analysis on Voice Using AWS Comprehend', *2020 International Conference on Computer Communication and Informatics (ICCCI -2020)*, IEEE. DOI: 10.1109/ICCCI48352.2020.

[4] Wavhale, V.D., Bira, S., Kumar, V. and Choudhari, V.R. (2020) 'Weather Data Forecast and Analytics', *International Research Journal of Engineering and Technology (IRJET)*, 07(08), pp. 651–652.

[5] Kulkarni, A. (2023) 'Amazon Athena: Serverless Architecture and Troubleshooting', *International Journal of Computer Trends and Technology*, 71(5), pp. 57–61. DOI: 10.14445/22312803/IJCTT-V71I5P110.

[6] Patsidis, A., Dyśko, A., Booth, C., Rousis, A.O., Kalliga, P., and Tzelepis, D. (2023) 'Digital Architecture for Monitoring and Operational Analytics of Multi-Vector Microgrids Utilizing Cloud Computing, Advanced Virtualization Techniques, and Data Analytics Methods', *Energies*, 16, p. 5908. DOI: 10.3390/en16165908.

[7] Gupta, A., Dhanda, N. and Gupta, K.K. (2023) 'Ingest and Visualize CSV Files using AWS Platform For Transition from Unstructured to Structured Data', *11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)*, IEEE.

[8] Kaniganti, S.T. (2020) ‘Architecting Real-Time Big Data Analytics: An AWS-Powered Framework Integrating AI and ML for Predictive Insights’, *International Journal of Science and Research (IJSR)*, 9(2). DOI: 10.21275/SR24716230925.

[9] Rahman, M.M. and Hasan, M.H. (2019) ‘Serverless Architecture for Big Data Analytics’, *Global Conference for Advancement in Technology (GCAT)*, IEEE.

[10] Medium (n.d.) "Data Flow Diagram of the Project." Available at Medium: https://miro.medium.com/v2/resize:fit:705/1*Sw_L7u1TKza2CK-gFKTJxQ.png.

[11] Pochetti, F. (2020) "AWS System Architecture." Available at Francesco Pochetti Blog: https://francescopochetti.com/wp-content/uploads/2020/03/AWS-2.png.

[12] Duval, J.N. (2024) "2024 Fortune 1000 Companies Dataset." Available at Kaggle: https://www.kaggle.com/datasets/jeannicolasduval/2024-fortune-1000-companies.