# AWS Project

Superstore Sales Data Analysis

Executive Summary:

This project demonstrates the use of Amazon Web Services (AWS) to analyze superstore sales data. The analysis was performed using AWS S3 for data storage, AWS Glue for data cataloging, AWS Athena for SQL-based serverless data querying, and AWS Quick Sight for visualization. The project's objective was to provide actionable insights into sales trends, customer behavior, and product performance using a scalable and cost-effective cloud-based solution.

Project Structure:
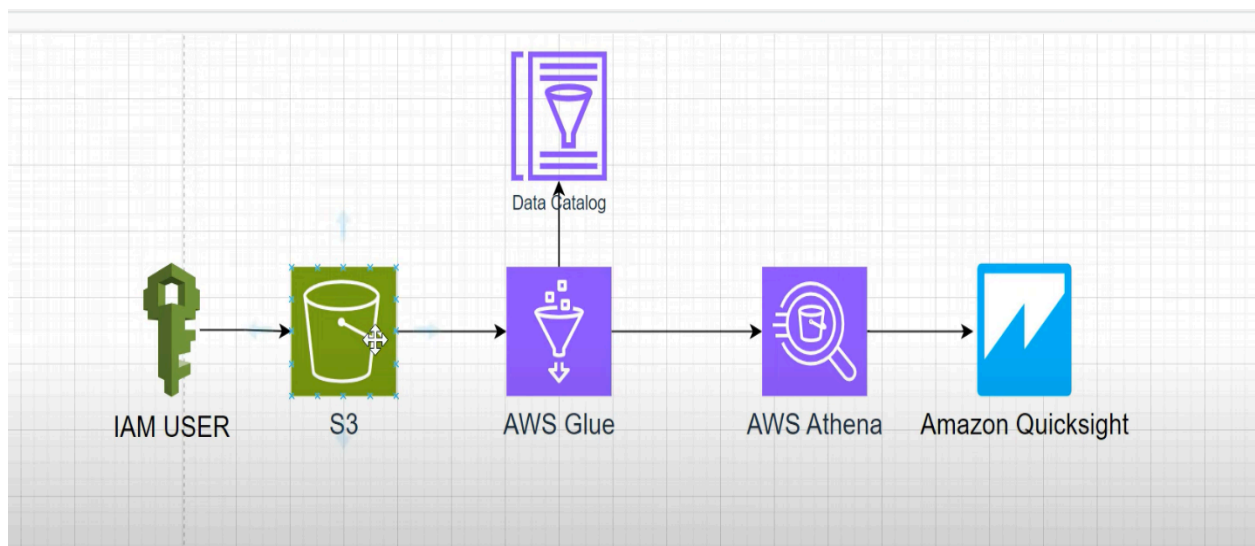
Project Objectives

1. Create a scalable and secure data pipeline for analyzing superstore sales data.

2. Leverage AWS Glue to extract metadata for efficient querying.

3. Use AWS Athena to perform SQL-based analysis without the need to load data into a database.

4. Design an interactive dashboard in AWS Quick Sight to visualize key findings.

Tools and Services Used:

- **AWS S3**: For storing raw sales data sourced from Kaggle.

- **AWS Glue**: For building a crawler to fetch metadata and catalog data stored in S3.

- **AWS Athena**: For querying data directly from S3 using SQL.

- **AWS QuickSight**: For creating an interactive dashboard to present the analysis.

Project Workflow:

1. Data Ingestion

- A dataset of superstore sales was downloaded from Kaggle.

- An Amazon S3 bucket was created to securely store the dataset. The data was uploaded in CSV format for ease of access.

2. Metadata Extraction with AWS Glue

- An AWS Glue Crawler was configured and executed to scan the S3 bucket.

- The crawler fetched metadata and created a data catalog in AWS Glue.

- This catalog allowed seamless querying of the data stored in S3 without manual schema definitions.

- Used partitions to access relevant data for faster processing

3. Data Analysis with AWS Athena

- AWS Athena was connected to the Glue Data Catalog.

- SQL queries were used to analyze sales data directly from the S3 bucket.

- Key metrics analyzed included:

  - Total Sales by Region.

  - Sales Analysis by Product Category.

  - Customer Interests in Purchase Frequency.

  - Discounts by Products Category.

4. Visualization with AWS QuickSight

- AWS QuickSight was integrated with Athena to access the data analyzed.

- An interactive dashboard was built to display:

  o City wise Sales using pie chart

  o Sales by Product type using horizontal bar chart.

  o Product category profitability using bar chart.

  o Discount % by category insights using donut chart.
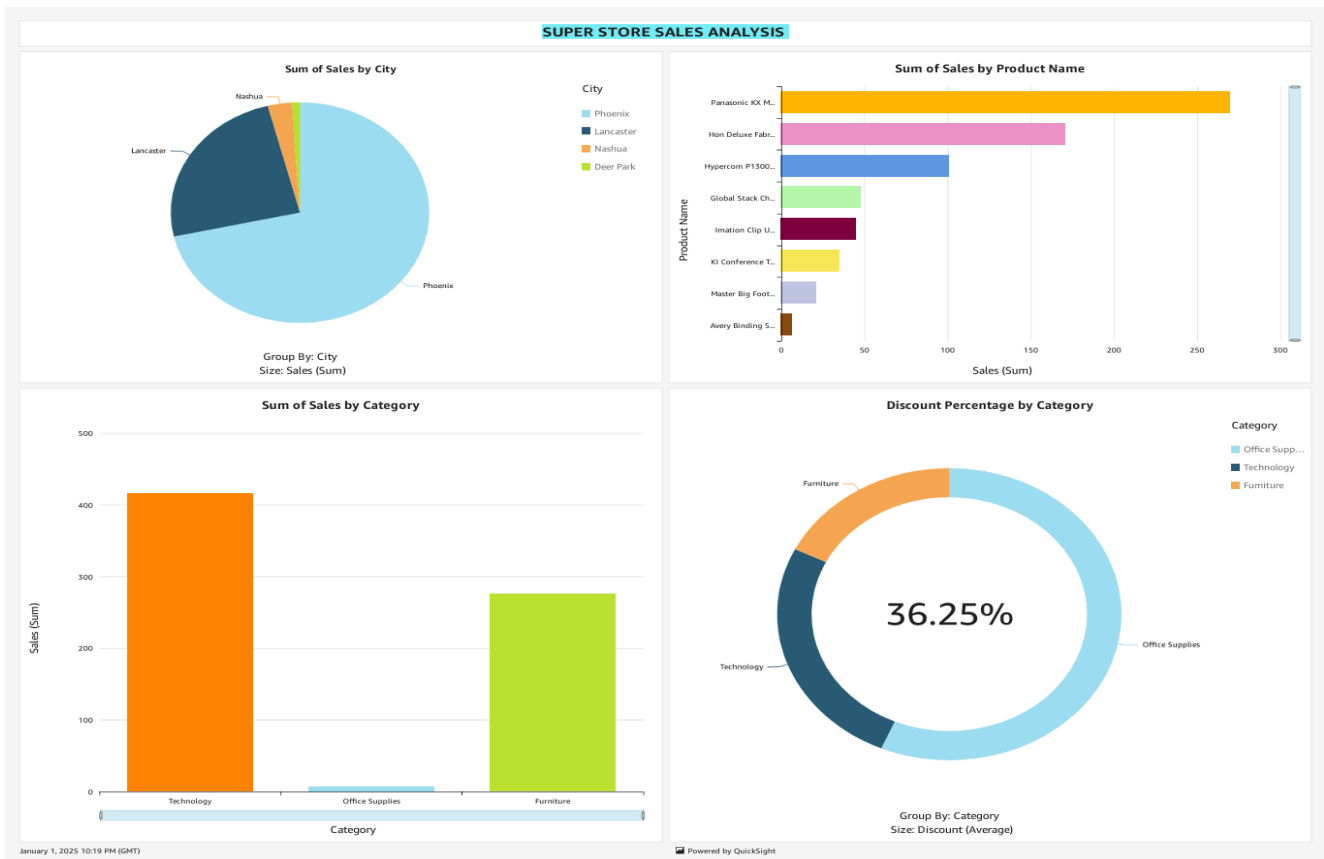
Results and Findings

1. **Regional Sales Insights**:

   o The western region had the highest sales volume, while the southern region showed potential for growth.

2. **Profitability Analysis**:

   o Office supplies had consistent profits, whereas technology products had high sales but lower margins.

3. **Customer Behavior**:

   o A significant portion of revenue came from repeat customers.

4. **Seasonal Trends**:

   o Sales peaked during Q4, indicating seasonal shopping behavior.

Challenges and Solutions

1. **Data Format Compatibility**:

   o Ensured proper CSV formatting to avoid crawler errors.

2. **Query Optimization**:

   o Used partitioning and filters in Athena to reduce query runtime and cost.

3. **Dashboard Design**:

   o Incorporated user-friendly visuals and filters to make the QuickSight dashboard intuitive.

Conclusion

The project successfully demonstrated the capabilities of AWS services to create a robust, cloud-based data pipeline for superstore sales data analysis. The integration of S3, Glue, Athena, and QuickSight provided a seamless workflow for serverless ingesting, processing, querying, and visualizing data. The insights derived can help businesses make informed decisions to improve performance and profitability.

Appendix

- **Dataset Source**: Kaggle Superstore Sales Data.

- **AWS Services Used**: S3, Glue, Athena, QuickSight.