Superstore Data Analysis I End to End AWS Data Engineering I Project

PROBLEM STATEMENT:

- In today's data-driven world, organizations often struggle to handle and analyze the ever-increasing volumes of data generated from various sources.
- Traditional data analytics solutions require significant infrastructure setup, management, and maintenance efforts, which can be time-consuming, resource-intensive, and costly.
- Furthermore, organizations may face challenges in streamlining the processes of data discovery, querying, and visualization, leading to inefficiencies and delayed insights.
- To address these challenges, there is a need for a scalable, cost-effective, and serverless data analytics solution that enables organizations to store and analyze vast amounts of data without the overhead of managing underlying infrastructure.
- The solution should leverage AWS cloud-native services to automate data ingestion, cataloging, querying, and visualization, thereby providing a seamless and streamlined experience for data analysts and business users.

SOLUTION:

1. **Serverless Architecture**: Implement a serverless architecture that eliminates the need for provisioning and managing servers or databases, allowing organizations to focus on data analytics rather than infrastructure management.
2. **Scalability and Cost-Efficiency**: Provide a highly scalable and cost-efficient solution that can handle varying data volumes and workloads without upfront investment in hardware or resources.
3. **Data Discovery and Cataloging**: Enable efficient data discovery and cataloging by automatically extracting metadata from various data sources, creating a centralized repository for easy access and understanding of data assets.
4. **Querying and Analysis**: Offer a user-friendly interface for querying and analyzing data using standard SQL, enabling data analysts and business users to extract insights without the need for specialized skills or tools.
5. **Data Visualization**: Provide powerful data visualization capabilities, allowing users to create interactive dashboards, reports, and visualizations to effectively communicate insights and facilitate data-driven decision-making.
6. **Integration with Existing Data Sources**: Seamlessly integrate with various data sources, such as object storage, relational databases, and data lakes, ensuring compatibility with existing data infrastructure.
7. **Security and Compliance**: Incorporate robust security measures and compliance features to protect sensitive data and ensure adherence to industry standards and regulations.

By implementing this serverless data analytics solution, organizations can unlock the full potential of their data, gain valuable insights, and make informed decisions while minimizing infrastructure overhead and operational costs.

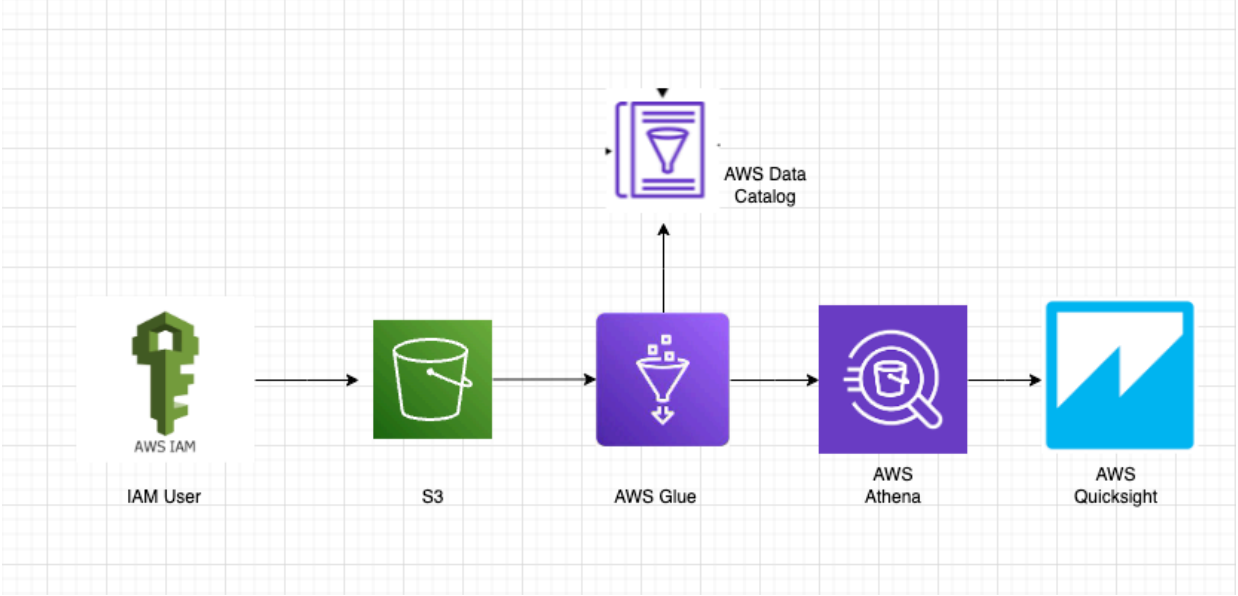
Data set used: Kaggle dataset - https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

  - This data set is about a Superstore giant who is trying to understand what works best for them. They would like to understand which products, regions, categories and customer segments they should target or avoid.
- Metadata:
  - Row ID => Unique ID for each row.
  - Order ID => Unique Order ID for each Customer.

    - Order Date => Order Date of the product.
    - Ship Date => Shipping Date of the Product.
    - Ship Mode=> Shipping Mode specified by the Customer.
    - Customer ID => Unique ID to identify each Customer.
    - Customer Name => Name of the Customer.
    - Segment => The segment where the Customer belongs.
    - Country => Country of residence of the Customer.
    - City => City of residence of of the Customer.
    - State => State of residence of the Customer.
    - Postal Code => Postal Code of every Customer.
    - Region => Region where the Customer belong.
    - Product ID => Unique ID of the Product.
    - Category => Category of the product ordered.
    - Sub-Category => Sub-Category of the product ordered.
    - Product Name => Name of the Product
    - Sales => Sales of the Product.
    - Quantity => Quantity of the Product.
    - Discount => Discount provided.
    - Profit => Profit/Loss incurred.

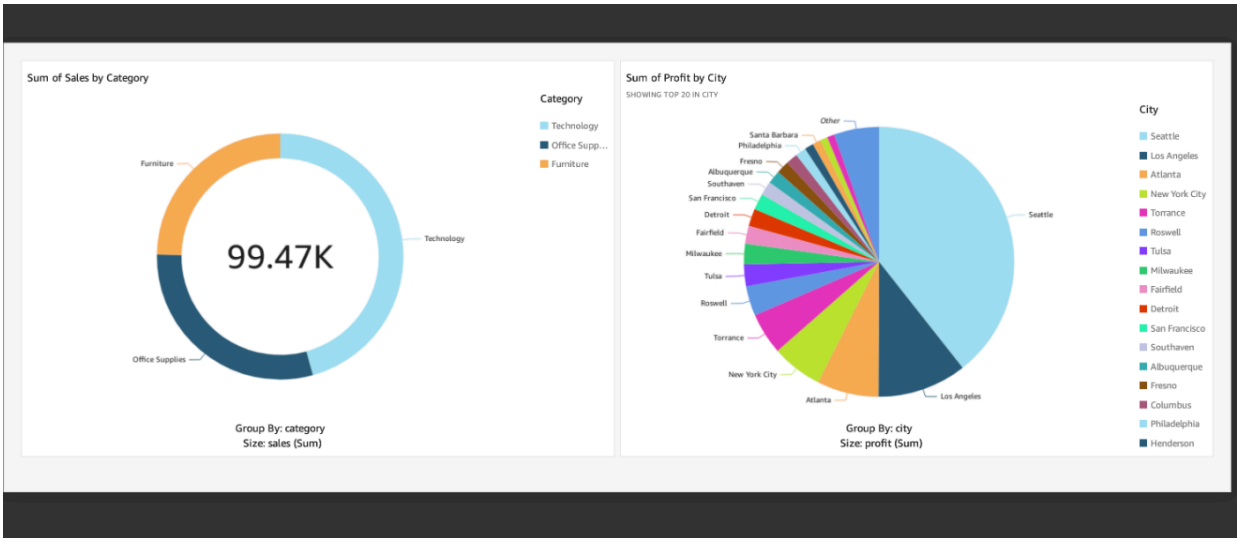
- AWS Services:
    - IAM
    - S3
    - AWS Glue
    - AWS Athena
    - AWS Quicksight

Project Architecture:



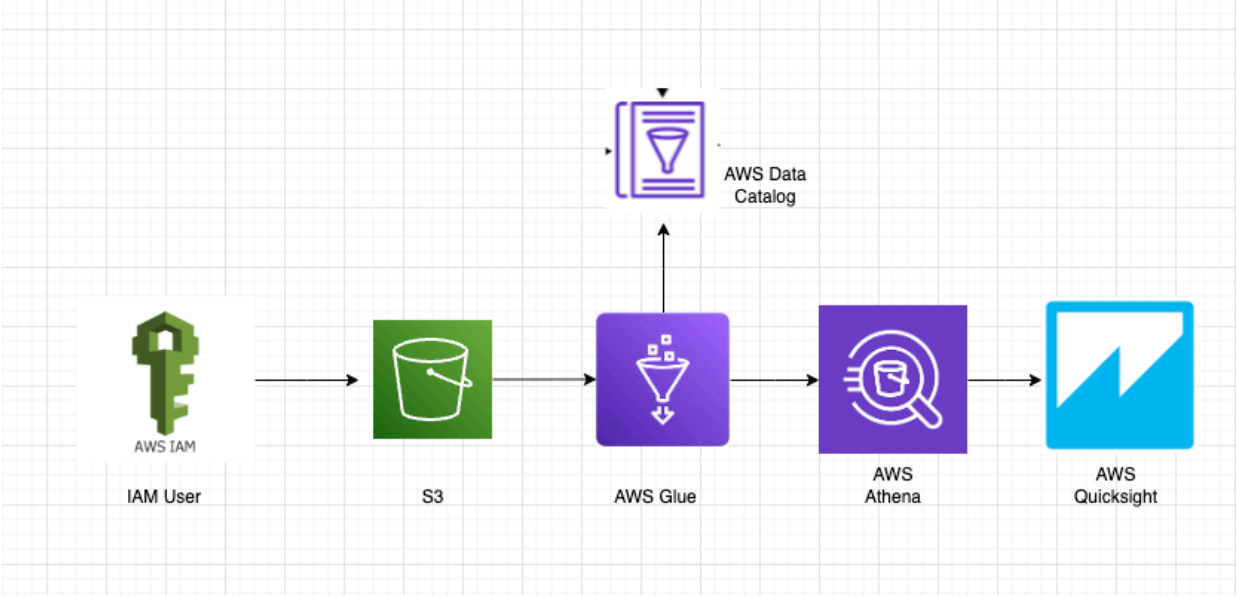
Quicksight DB

Sheet_1_2024-05-20T01_26_18.pdf



1. **Sales By Product Category**
2. **Profit by City**

SUMMARY:



This architecture diagram illustrates end to end basic data analytics workflow using AWS services. Here's an explanation of the components involved:

1. **AWS IAM (Identity and Access Management)**: IAM User represents an authenticated user or application that initiates the data analytics process.
2. **Amazon S3 (Simple Storage Service)**: S3 is an object storage service where the raw data or files are stored.
3. **AWS Glue**: Glue is a fully managed extract, transform, and load (ETL) service. It crawls the data stored in S3, infers the schema, and creates metadata tables in the AWS Data Catalog.
4. **AWS Data Catalog**: The Data Catalog is a centralized repository that stores metadata about the data sources, data formats, and data schemas. It maintains a unified view of data across different data stores, making it easier to discover and access data for analytics.
5. **AWS Athena**: Athena is an interactive query service that allows you to analyze data directly from S3 using standard SQL queries. It uses the metadata stored in the Data Catalog to understand the structure of the data and perform queries.
6. **AWS QuickSight**: QuickSight is a cloud-native business intelligence (BI) service that allows you to create visualizations, dashboards, and reports based on the data queried from Athena or other data sources.

Workflow:

- The workflow can be summarized as follows:

1. The IAM User (or an application) uploads raw data or files to Amazon S3.
2. AWS Glue crawls the data stored in S3 and infers the schema, creating metadata tables in the AWS Data Catalog.
3. The AWS Data Catalog stores and maintains the metadata about the data sources, formats, and schemas.
4. AWS Athena uses the metadata from the Data Catalog to query and analyze the data stored in S3 using SQL.
5. The queried data from Athena can be visualized and analyzed further using AWS QuickSight, which creates dashboards and reports based on the data.

- This architecture allows for serverless data analytics, where you can store and analyze vast amounts of data without having to manage any underlying infrastructure.

- It leverages AWS services like Glue, Data Catalog, Athena, and QuickSight to streamline the process of data discovery, querying, and visualization.

Appendix:

- IAM User - Limited access to AWS Services for this project
- S3 Bucket - Will store the Orders Incremental Data In Folders/Sub-Folders
- AWS Glue - Used to run the crawler and create Data Catalog
- AWS Athena - To perform Analysis on S3 data, run SQL queries
- AWS Quicksight - To create Dashboards / Visualization.
- AWS Glue is an ETL Tool. You can create and run Crawler to create a Data Catalog.
- Create Database in AWS Glue, this is not a DB but instead is like a folder in Glue.
- While creating a Crawler, used Crawl All new Sub folders only, in order to get incremental data.
- Created an IAM Role for Crawler.
- To query the data from an S3, you need to know its metadata. Crawler will help create that.
- Crawler crawls over the s3 file and will create a Table (metadata only). It does not create Data. Data will remain in S3. It identifies the schema.
- With the sub-folders created in s3, the Crawler created partitions. These can be used to run query in Athena.
- You can also schedule your crawler to run based on frequency.
- Data catalog holds only data about data (metadata) in the form of tables.
- While using Athena, create an S3 to store the results of the query that runs.
- With Partitions your query will run faster, scans will be based on partition, less cost.
- In Quicksight you can create charts. Use SPICE for creating visualizations for extracted data.
- Use Directly query your data for live data.