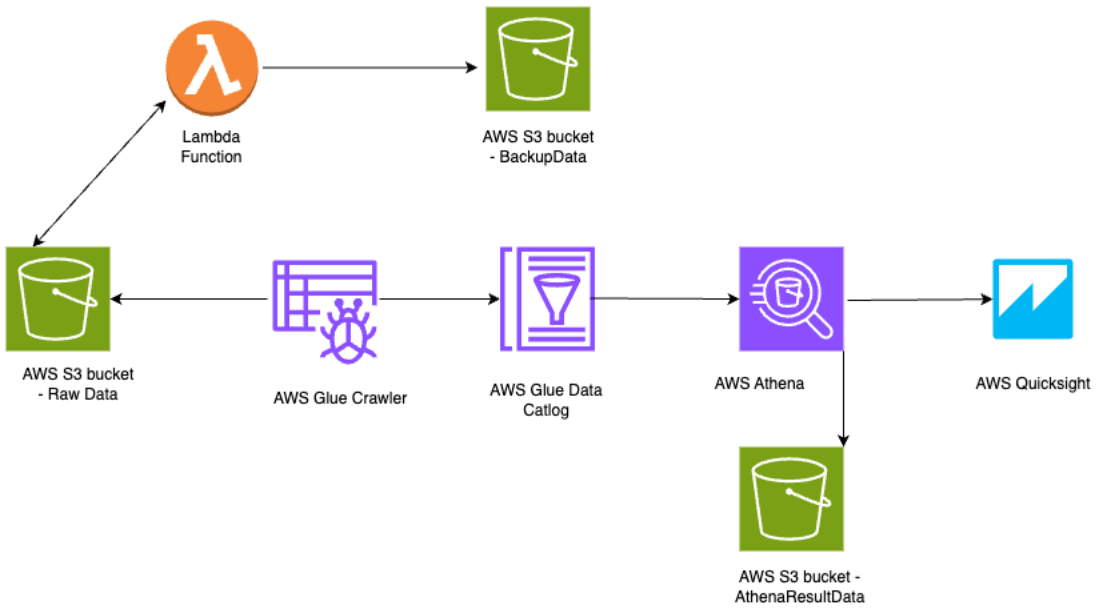
This setup provides a streamlined approach to managing and analyzing raw data while ensuring data redundancy and ease of access for reporting and insights generation. By leveraging AWS services, the project establishes a robust data pipeline that supports scalability and efficient data analysis.

This project involves setting up a data processing pipeline using AWS services, including S3, Athena, Glue, Lambda, and QuickSight. The goal is to efficiently manage raw data, perform queries, and generate insights.

Below diagram shows the data pipeline architecture of this project:



Step-by-Step Implementation

1. **User Creation**
    - Created an IAM user with necessary permissions to access S3, Athena, Glue, Lambda, and QuickSight. Assigned policies for data read/write operations and service access.
2. **S3 Buckets Setup**
    - **Raw Data Bucket**: Created an S3 bucket designated for storing raw data uploads.
    - **Athena Query Results Bucket**: Created a separate S3 bucket to store the results generated by Athena queries.
    - **Backup Data Bucket**: Established an S3 bucket to serve as a backup location for raw data.
3. **AWS Glue and Athena Configuration**

- Opened Amazon Athena and created a Glue Crawler to scan the raw data bucket, which automatically catalogs the data.
- Configured the Glue Data Catalog, ensuring it properly defines the schema based on the raw data.
- Created Athena tables based on the Glue catalog, allowing for efficient querying of the raw data.
- Assigned the query results bucket for storing Athena query outputs.

4. **Data Processing with AWS Lambda**
   - Developed an AWS Lambda function that triggers on data uploads to the raw data S3 bucket.
   - Implemented logic within the function to automatically copy uploaded objects to the backup data bucket for redundancy.
   - Create a new IAM role and set necessary IAM permissions for the Lambda function to read from the raw data bucket and write to the backup bucket.
5. **Data Analysis with Amazon QuickSight**
   - Opened Amazon QuickSight and connected it to the Athena dataset, selecting the relevant tables created from the Glue catalog.
   - Created visual insights and dashboards based on the queried data, customizing visualizations to highlight key metrics and trends.
   - Configured permissions in QuickSight to share insights with stakeholders or team members as needed.
6. **Monitoring and Maintenance**
   - Set up CloudWatch to monitor the Lambda function and receive alerts for any failures or performance issues.
   - Implemented S3 bucket lifecycle policies for automatic data archiving or deletion based on data retention requirements.
   - Regularly reviewed IAM roles and permissions to ensure security best practices are maintained.