**AWS-Based Superstore Data Analysis: A Research-Oriented Approach**

**Marcelo Vicente & Harry Depina**

**Project Purpose and Introduction**

The purpose of this research-oriented project was to explore the end-to-end use of Amazon Web Services (AWS) tools for a real-world business intelligence (BI) and knowledge management (KM) scenario. We chose to use a popular "Superstore" dataset to simulate a real-life small business process. The project involved data storage in Amazon S3, cataloging in AWS Glue, querying in Amazon Athena, and result visualization with Amazon QuickSight. It aimed to create an end-to-end cloud-based, scalable, and cost-effective BI solution for small and medium businesses, emphasizing efficiency, knowledge integration, and real-time analysis.

**Solution Architecture**

Our architecture consisted of several key AWS components seamlessly integrated to form a cohesive data analytics pipeline:

A screenshot of a computer

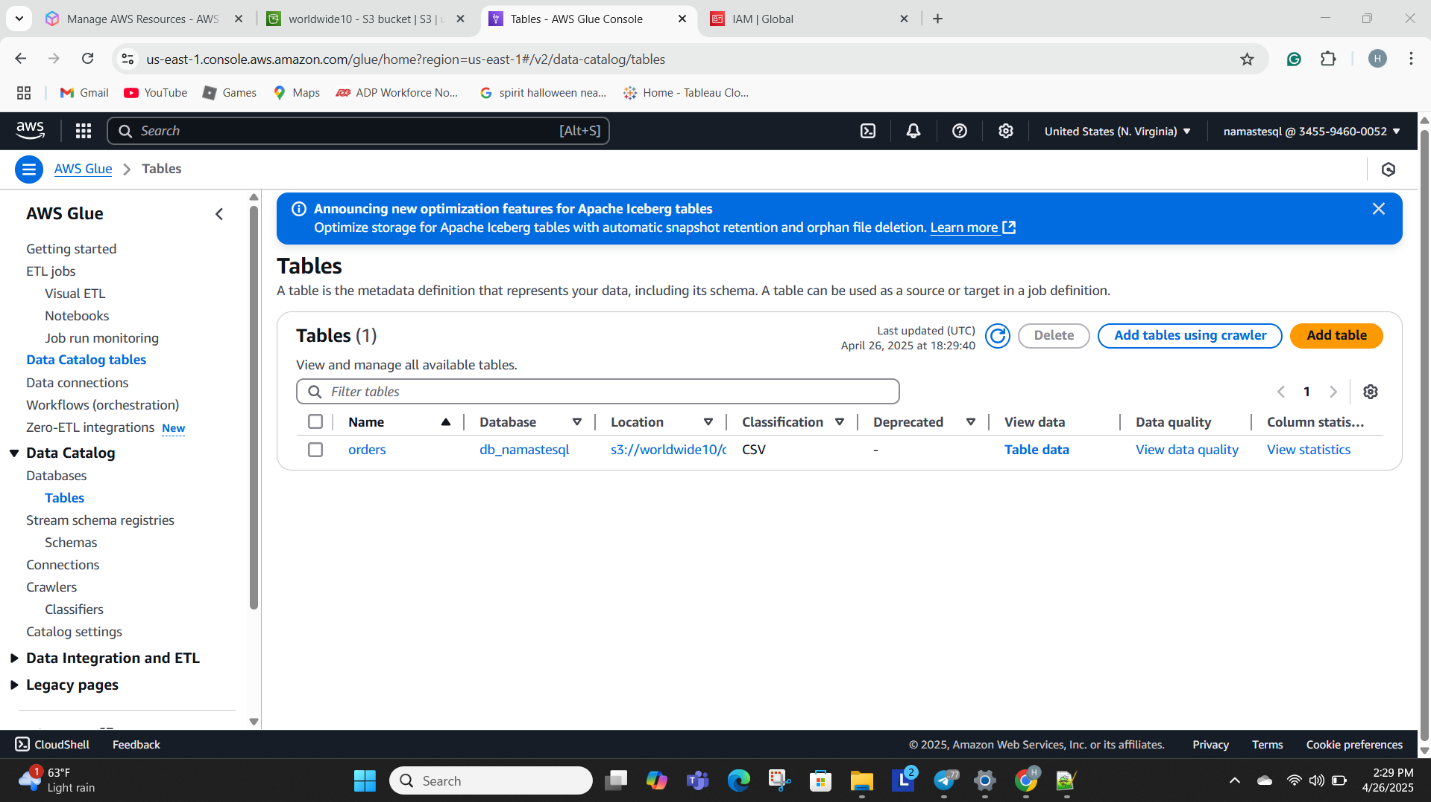
AI-generated content may be incorrect.

* **Amazon S3** was used as the primary cloud storage solution for raw and incremental datasets.

A computer screen with a message

AI-generated content may be incorrect.

* **AWS Glue** created the data catalog by crawling the S3 buckets and generating metadata schemas, enabling structured queries.



* **Amazon Athena** provided serverless SQL querying capabilities directly over the S3 datasets without needing a traditional database.

A computer screen shot of a computer screen

AI-generated content may be incorrect.

* **Amazon QuickSight** served as the visualization layer, connecting to Athena to generate business dashboards and analytical charts.

A screenshot of a computer

AI-generated content may be incorrect.

This architecture emphasized serverless, pay-per-use computing to minimize costs while maintaining high flexibility. Partitioning strategies within S3 helped reduce query sizes, making operations faster and more affordable.

**Key Services in the Solution**

1. **Amazon S3**:
   * Provided secure, scalable storage.
   * Bucket policies and folder structures ("orders/", "snapshot\_date=YYYYMMDD/") ensured organized, partitioned data.
2. **AWS Glue**:
   * Crawlers automatically detected schema changes.
   * The Glue Data Catalog facilitated a seamless connection to Athena.
   * Created a logical structure to make S3 data appear like relational tables.
3. **Amazon Athena**:
   * Enabled serverless SQL querying without provisioning infrastructure.
   * Benefited from data partitioning to reduce scan size, cost, and latency.
4. **Amazon QuickSight**:
   * Provided rapid dashboard creation.
   * Direct integration with Athena allowed near real-time reporting.
   * Enabled the creation of multiple visualization types (bar charts, pie charts) for sales and profit analysis.

**Demonstration and Discussion**

Throughout the project, we simulated the full data lifecycle:

* Uploaded partitioned order datasets to S3 to simulate incremental daily sales.
* Created a Glue Crawler to scan the bucket and update the Data Catalog.
* Queried the data in Athena using simple and advanced SQL queries, including aggregation functions (e.g., SUM(Sales) by Category).
* Built interactive dashboards in QuickSight to visualize sales by category, region, and snapshot day.

One key decision was to use partitioning based on the snapshot day (e.g., snapshot\_date=20170101/). This allowed Athena queries to target only necessary partitions, significantly reducing query costs and runtime. Partitioning was especially important to stay within AWS Free Tier limits and mimic best practices for cloud data management.

**Implementation Challenges and Solutions**

* **Codespace and Docker Issues**: Initially, recovery mode errors occurred due to a misconfigured dev container setup. The team had to review creation logs, update the Dockerfile, and rebuild the container to fix this.
* **IAM Role Confusion**: Assigning correct permissions to Glue, Athena, and QuickSight required multiple iterations. We resolved this by carefully applying least-privilege access and verifying role-trust relationships.

A screenshot of a computer

AI-generated content may be incorrect.

* **Crawler Errors**: During initial crawls, partitions were not detected correctly because of folder misnaming. Renaming folders in the correct key=value format solved the problem.
* **Athena Query Failures**: Forgetting to configure the query result output bucket initially prevented SQL execution. Setting up a designated "athena-results" folder in S3 resolved this.
* **QuickSight Permissions**: QuickSight needed explicit permission to access the S3 buckets containing Athena results. Granting access under "Manage QuickSight" fixed this.

**Example Companies Using Similar Solutions**

1. Walmart: Walmart uses AWS to support big-box store analytics at a massive scale, leveraging cloud storage, big data processing (such as our S3, Glue, and Athena setup), and real-time dashboards to optimize logistics and sales strategy.
2. Expedia: Expedia migrated its on-premise data warehouses to AWS, leveraging services like S3, Athena, and QuickSight to analyze user behaviors, optimize bookings, and improve customer experiences through more targeted insights from massive datasets.

Both use cases reflect the power of serverless analytics to provide scalable, rapid decision-making without the expense of infrastructure management.

**Project Challenges and Lessons Learned**

* **Importance of Clean Architecture**: Having a clear end-to-end plan prevents confusion when connecting services.
* **Early Testing Matters**: Setting up small, incremental datasets during early testing helped us catch errors without blowing through Free Tier limits.
* **Partitioning Strategy Was Critical**: Organizing data by snapshot date greatly reduced query scan size in Athena, allowing fast, inexpensive queries. Without it, Athena would have scanned entire datasets, causing performance and cost issues.
* **Documentation**: Maintaining a log of configuration steps made troubleshooting much easier, especially when permissions or services didn't behave as expected.
* **Career Implications**: This project emphasized the growing importance of cloud-native data solutions in business analytics. As future business analysts, mastering serverless, scalable cloud services like S3, Glue, Athena, and QuickSight will be essential for delivering agile, data-driven insights to organizations.

**References**

* Amazon Web Services. (2024). *What is Amazon Athena?* Retrieved from: https://docs.aws.amazon.com/athena/latest/ug/what-is.html
* Amazon Web Services. (2024). *Getting Started with AWS Glue.* Retrieved from: https://docs.aws.amazon.com/glue/latest/dg/getting-started.html
* Amazon Web Services. (2024). *Amazon S3 Documentation.* Retrieved from: https://docs.aws.amazon.com/s3/index.html
* Amazon Web Services. (2024). *What is Amazon QuickSight?* Retrieved from: https://docs.aws.amazon.com/quicksight/latest/user/welcome.html
* Airbnb Engineering. (2017). *Scaling Data Analysis at Airbnb.* Retrieved from: https://medium.com/airbnb-engineering/scaling-data-analytics-airbnb-222fca079f2f
* FINRA Case Study on AWS. (2024). Retrieved from: https://aws.amazon.com/solutions/case-studies/finra/