

NEXUS DATALENS

Version 2.0

An End-to-End Automated Data Analytics Pipeline on AWS

Automated Serverless Data Analytics Pipeline on AWS

Prepared by
Ruksana Shaikh

Date: September 9, 2025

Abstract

Nexus DataLens 2.0 is a cloud-native, end-to-end automated data analytics pipeline built on Amazon Web Services (AWS). It addresses the challenge of transforming raw CSV data into meaningful insights without manual processing or technical expertise. Users upload CSV files through a web interface, after which the system autonomously performs ingestion into Amazon S3, cleaning with AWS Glue, schema cataloging through AWS Glue Crawlers, and visualization using Amazon QuickSight or Power BI. The result is a scalable, secure, and cost-effective solution that accelerates data-to-decision workflows with minimal effort, showcasing the power of serverless, event-driven architectures in modern data engineering. This solution is applicable to SMEs, research institutions, and enterprises looking for automated analytics

Table of Contents

- 1. Introduction**
 - 1.1 Background
 - 1.2 Motivation
 - 1.3 Scope of the Project
 - 1.4 Technologies Used
- 2. Problem Statement**
- 3. Solution Overview**
 - 3.1 Objectives
- 4. System Architecture**
 - 4.1 Workflow Diagram
 - 4.2 AWS Services Used
- 5. Frontend Application**
 - 5.1 Upload Interface
 - 5.2 S3 Bucket Structure
 - 5.3 Frontend File Upload and Hosting
 - 5.4 Secure File Storage in Amazon S3
- 6. Automation Using AWS Lambda**
 - 6.1 Trigger Mechanism
 - 6.2 Lambda Function Logic
- 7. Data Processing Pipeline**
 - 7.1 AWS Glue Crawler
 - 7.2 AWS Glue Job for Data Cleaning
 - 7.3 AWS Glue DataBrew (Optional)
 - 7.4 Second Lambda Invocation
- 8. Data Storage & Querying**
 - 8.1 Cleaned Data Output in S3
 - 8.2 Querying via Athena
- 9. Visualization and Reporting**
 - 9.1 Amazon QuickSight Dashboard
 - 9.2 Power BI Integration (Optional)
- 10. Challenges and Resolutions**
- 11. Key Features & Benefits**
- 12. Use Cases of Nexus DataLens**
- 13. Results and Outcomes**

14. Conclusion and Future Enhancements

14.1 Recommendations

14.2 Future Scope

15. References

16. Appendix

16.1 Source Code Links (GitHub)

16.2 Sample Input/Output

16.3 Additional Screenshots

1. Introduction

1.1 Background

In today's data-driven environment, organizations collect large volumes of raw data from various sources. However, transforming this data into actionable insights typically requires technical expertise, manual cleaning, and lengthy processes. There is a growing need for scalable, automated solutions that can streamline data ingestion, processing, and visualization particularly for non-technical users. Nexus DataLens addresses this gap by providing a fully automated pipeline built on reliable AWS services.

1.2 Motivation

This project is motivated by the need to simplify and democratize data analytics. Manual workflows often introduce delays, errors, and scalability issues. With increasing data volumes, it becomes essential to implement a system that automates every step—from file ingestion to dashboard generation—thereby saving time, reducing effort, and enhancing accessibility for all users, regardless of technical background.

1.3 Scope of the Project

This project focuses on building an end-to-end data analytics solution capable of:

- Ingest CSV files via a web-based frontend
- Process and clean data automatically using AWS Glue and DataBrew
- Create and update Athena tables for SQL querying
- Visualize data dynamically in Amazon QuickSight and optionally Power BI
- Enables seamless, near real-time insights without human intervention

The scope is limited to CSV-based structured data and primarily uses AWS services under the free tier wherever possible.

1.4 Technologies Used

The Nexus DataLens project integrates a diverse range of AWS services and technologies to deliver a seamless, automated data analytics workflow. The technologies used are grouped below into **Frontend**, **Backend**, and **Support Services** for clarity:

Frontend Technologies

- **HTML, CSS, JavaScript** – Developed a clean, responsive user interface for CSV upload
- **Amazon S3 (Static Website Hosting)** – Hosts the frontend as a static web application
- **AWS Cognito** – Provides secure user authentication

- **Upload Path** – Files are uploaded directly to:
s3://nexus-web-frontend1/data/filename.csv

Backend & Data Processing Technologies

- **Amazon S3** – Stores both raw uploaded files and cleaned outputs
- **AWS Lambda** – Automates Glue Crawlers and Glue Jobs upon file upload
- **AWS Glue Crawler** – Detects schema from new data files and updates Glue Catalog
- **AWS Glue (Spark-based Jobs)** – Cleans and transforms CSV data efficiently
- **AWS Glue DataBrew** – Provides no-code UI-based transformation for advanced cleaning
- **Amazon Athena** – Performs SQL queries on cleaned datasets stored in S3
- **Amazon QuickSight** – Creates dynamic, interactive dashboards from Athena datasets
- **Power BI (*optional*)** – Enterprise-level reporting via Athena ODBC integration
- **Python** – Used in AWS Glue Jobs and Lambda scripts for logic implementation

Security, Monitoring & Integration Services

- **AWS IAM** – Role-based access control for services and users
- **AWS KMS** – Key management for data encryption at rest
- **AWS CloudWatch** – Real-time logging and monitoring of Lambda and Glue operations
- **AWS API Gateway** – Exposes RESTful endpoints if integrating external tools
- **Amazon CloudFront** – Content delivery network (CDN) for global access to the frontend
- **AWS Cognito** – Adds user pool-based authentication to restrict uploads

2. Problem Statement

In many organizations, data analysts and business users often spend significant time manually processing and cleaning uploaded data for reporting purposes. This traditional approach is not only time-consuming but also prone to human error, lacks scalability, and delays decision-making. There is a growing need for an automated mechanism that can reliably process raw user-uploaded CSV files and generate meaningful insights with minimal technical effort.

3. Solution Overview

Nexus DataLens addresses this challenge by implementing a fully automated, event-driven data analytics pipeline built on AWS. The architecture eliminates manual intervention and delivers real-time insights using the following key services:

- **Amazon S3** – for secure and scalable file storage
- **AWS Lambda** – for automating workflow execution upon file upload
- **AWS Glue (Crawler + Job)** – to infer schema and perform data cleaning and transformation
- **Amazon Athena** – to query structured, cleaned datasets using standard SQL
- **Amazon QuickSight (and optionally Microsoft Power BI)** – to build rich, interactive dashboards for end users

The entire pipeline is cloud-native, scalable, and designed for seamless integration, ensuring data moves from ingestion to visualization without any manual processing.

3.1. Objectives

The primary objectives of the Nexus DataLens project are as follows:

- To build a robust and scalable platform using AWS services for automated CSV data ingestion
- To clean and transform datasets using AWS Glue with no manual effort
- To enable seamless data visualization through **Amazon QuickSight** and **Power BI**
- To create a frontend user interface that allows non-technical users to upload datasets and receive insights effortlessly

Note: This project supports visualization through both Amazon QuickSight and Microsoft Power BI. Both platforms were successfully integrated and tested during implementation.

4. System Architecture

The architecture of **Nexus DataLens** is designed for **end-to-end automation**, **real-time data processing**, and **scalable analytics** — all built on a **serverless AWS stack**.

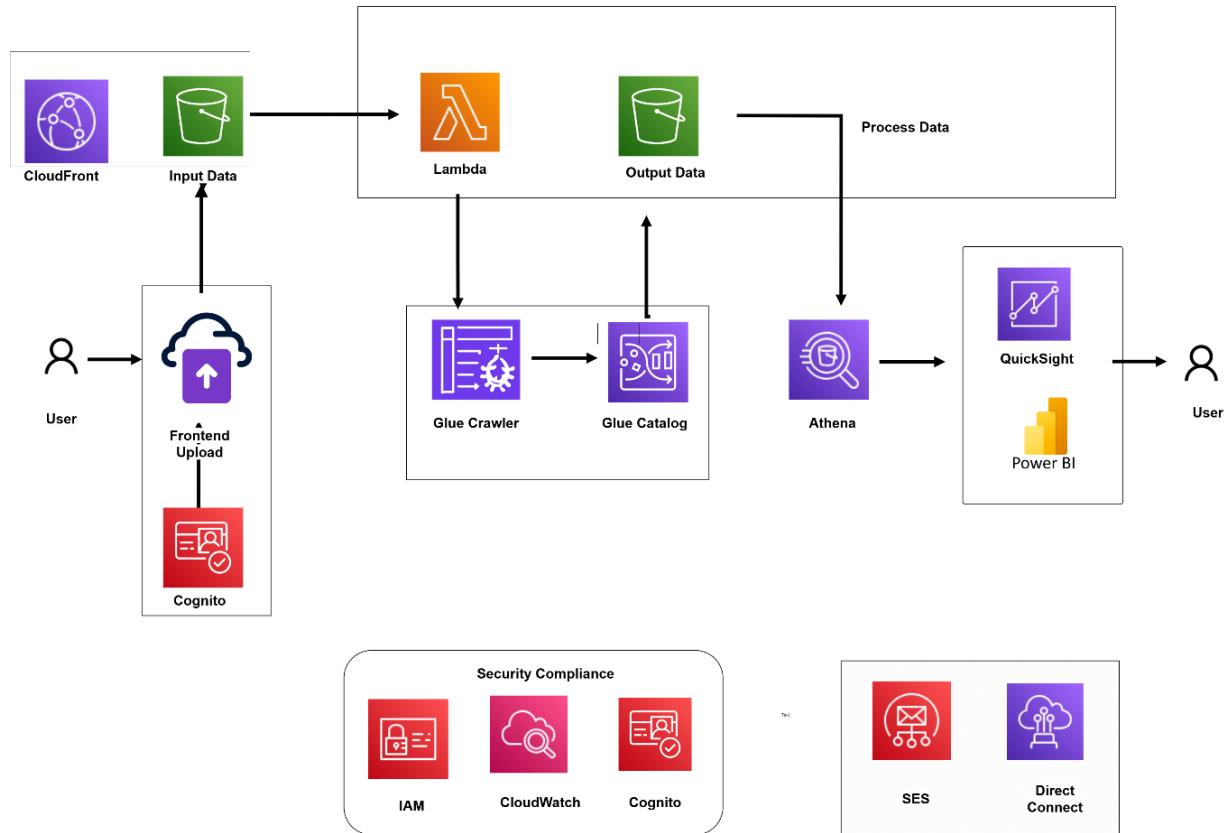


Figure 1: Nexus DataLens – End-to-End Serverless Architecture

Data Flow Explanation

This architecture enables seamless data ingestion, processing, and visualization using AWS services. Here's how the components interact:

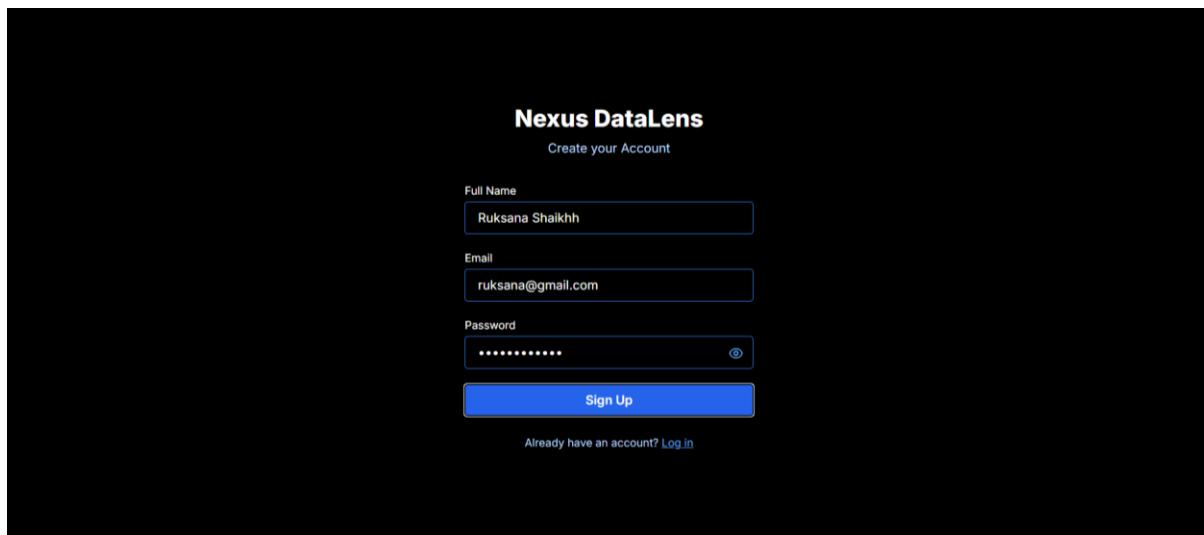
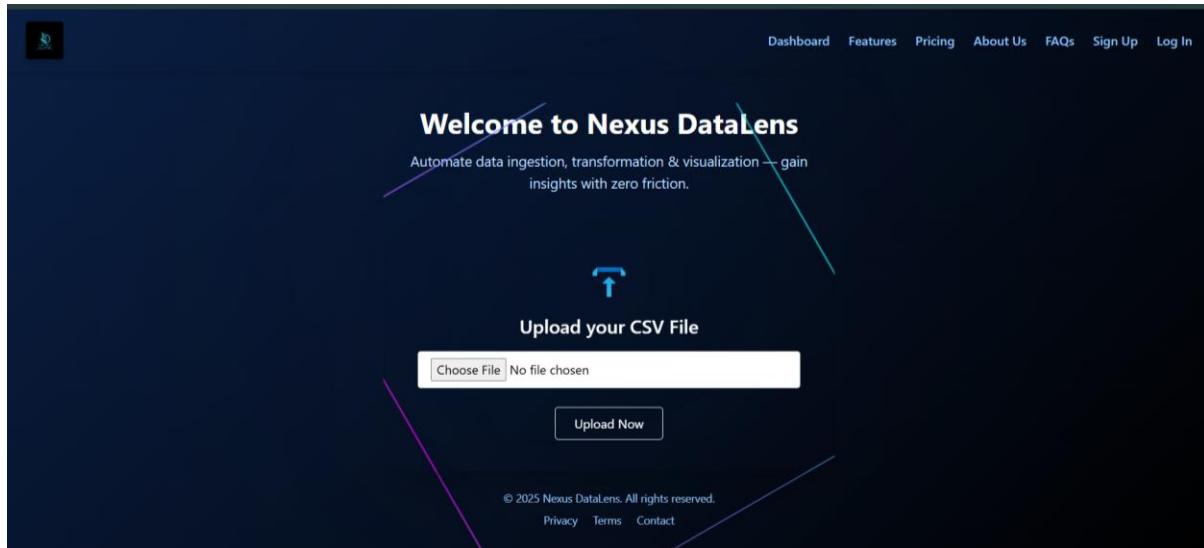
1. **Frontend Upload & Authentication**
 - Users upload CSV files through a web-based frontend.
 - **Amazon Cognito** secures access with optional user authentication.
 - Files are stored in the S3 input folder and served via **CloudFront**.
2. **AWS Lambda Orchestration**
 - The S3 upload triggers **AWS Lambda**, which initiates:
 - A **Glue Crawler** to scan and update the **Glue Catalog**
 - A **Glue ETL Job** to clean and format the data

- Cleaned data is written to the cleaned-output/ folder in S3.
- 3. **Athena & Data Querying**
 - Cleaned data is queried using **Amazon Athena**.
 - Athena integrates with both **QuickSight** and **Power BI** for dashboarding.
- 4. **Visualization & Reporting**
 - Users can visualize the data using **Amazon QuickSight** or **Microsoft Power BI** dashboards.
- 5. **Security & Compliance**
 - **IAM** ensures access control.
 - **CloudWatch** handles monitoring and alerts.
 - **Cognito** adds secure authentication for both the frontend and backend services.
- 6. **Optional Enterprise Features**
 - **Amazon SES** can send automated email reports/alerts.
 - **AWS Direct Connect** can be used in enterprise setups for faster, private network access.

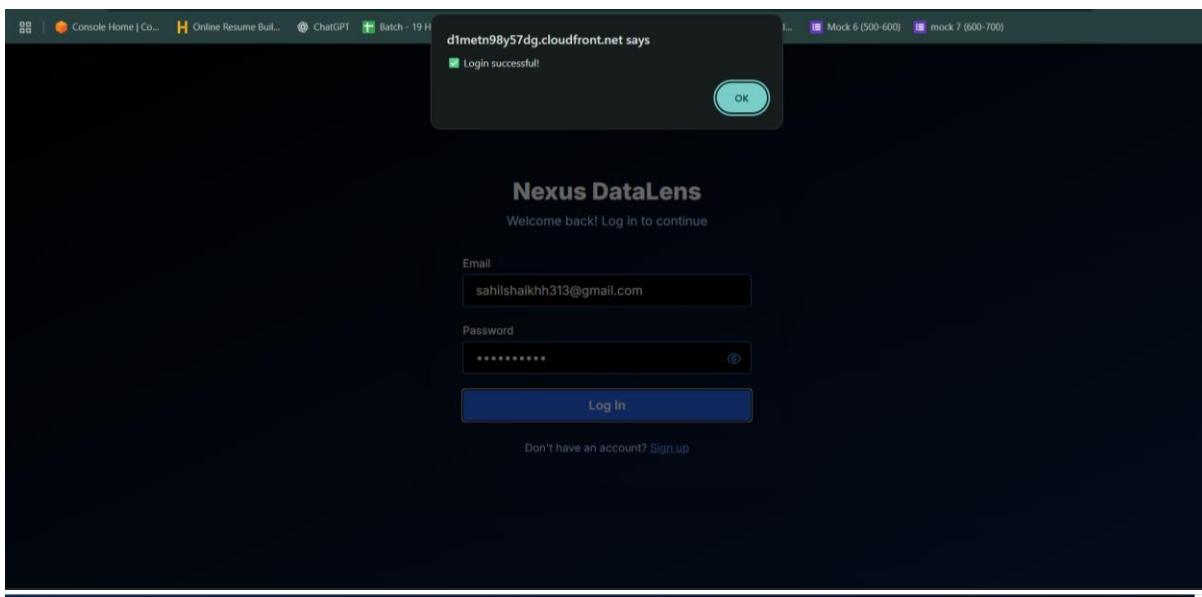
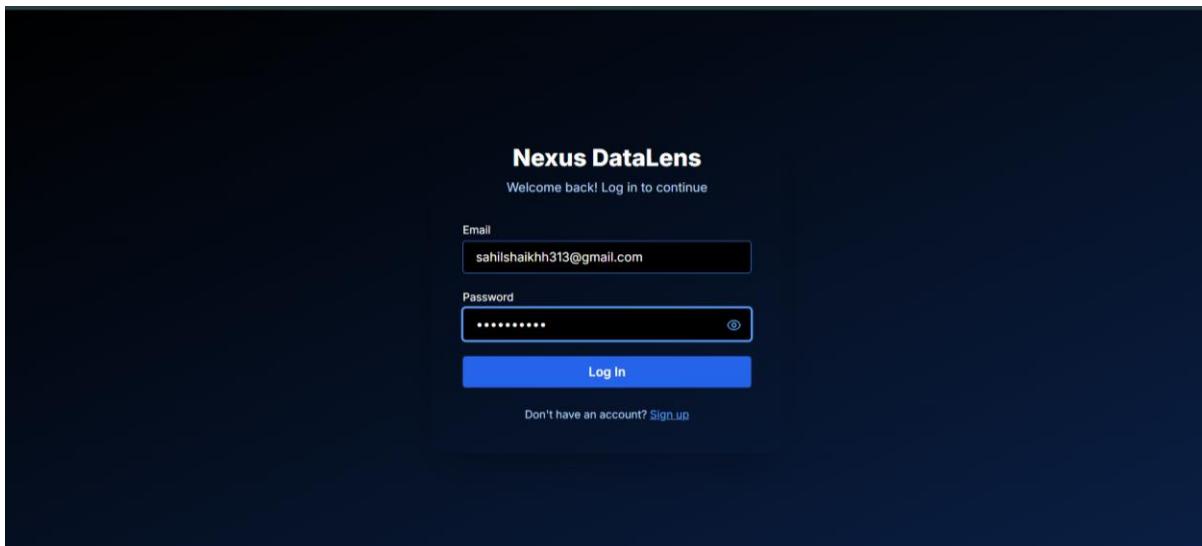
5. Frontend Application

5.1 Upload Interface

This section covers the CSV file upload mechanism built using a simple HTML/CSS/JavaScript interface hosted on Amazon S3.

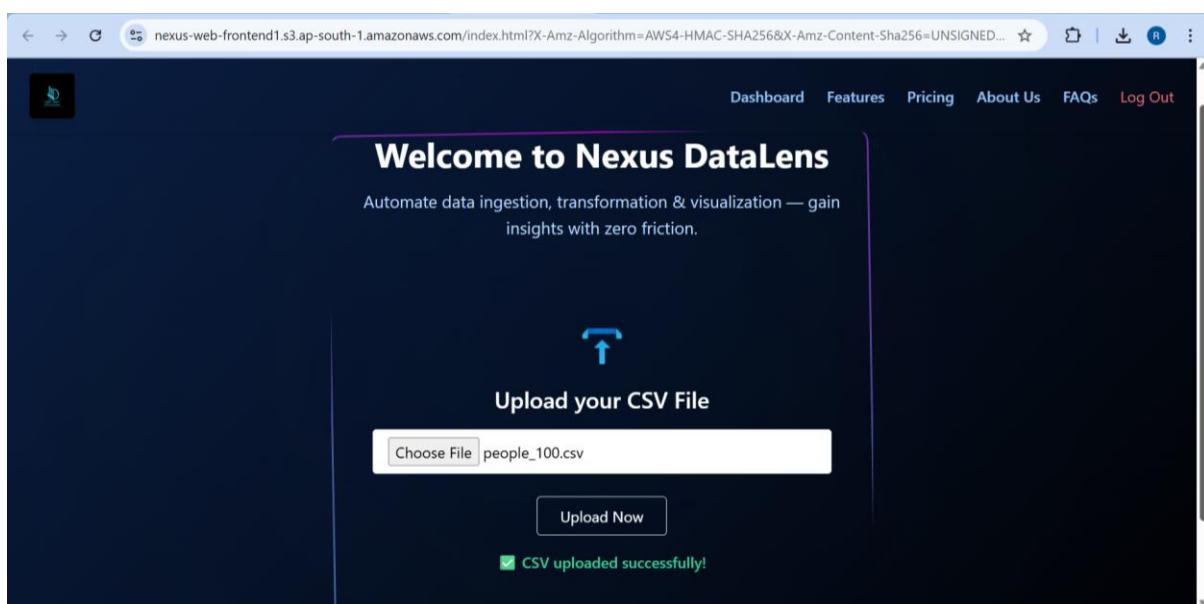
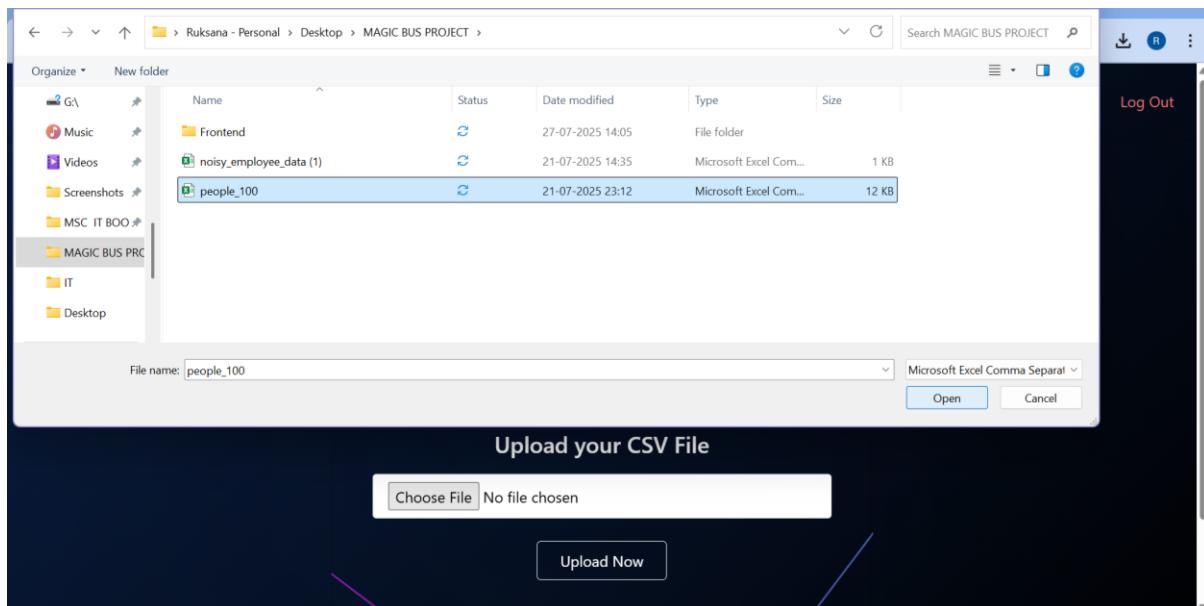


Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

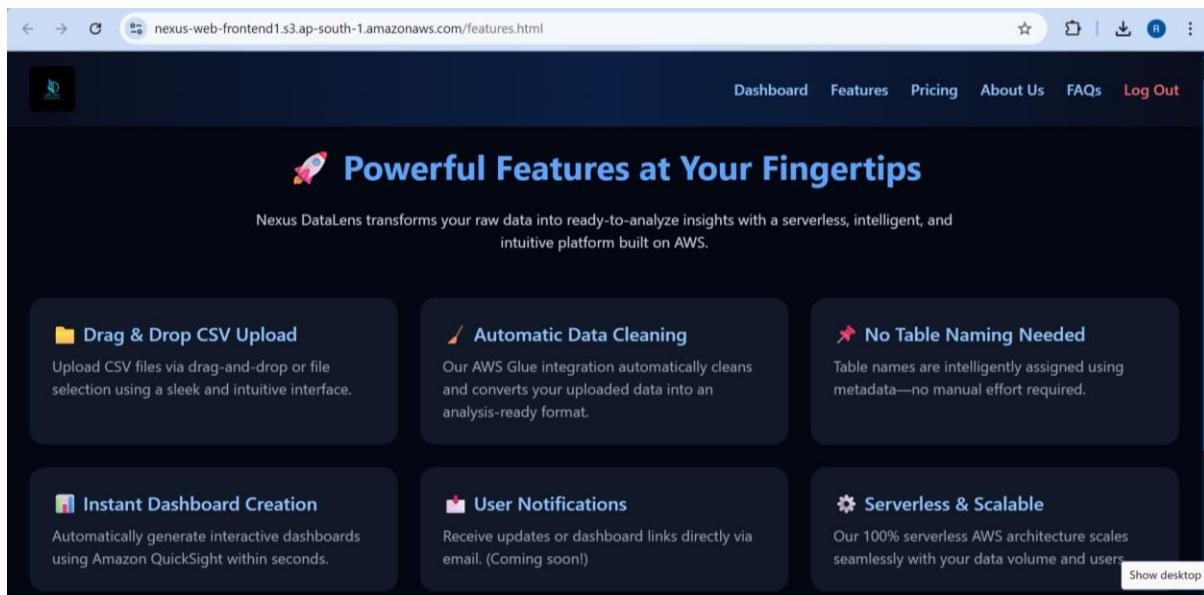


A screenshot of the Nexus DataLens dashboard. The top navigation bar includes links for "Dashboard", "Features", "Pricing", "About Us", "FAQs", and "Log Out". The main content area has a dark background with white text. It features a heading "Welcome to Nexus DataLens" and a subtext "Automate data ingestion, transformation & visualization — gain insights with zero friction." Below this is a large blue upward-pointing arrow icon. The text "Upload your CSV File" is displayed above a file input field containing "Choose File No file chosen". A blue "Upload Now" button is located below the input field. In the bottom right corner of the dashboard, there is a small "Show desktop" button.

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



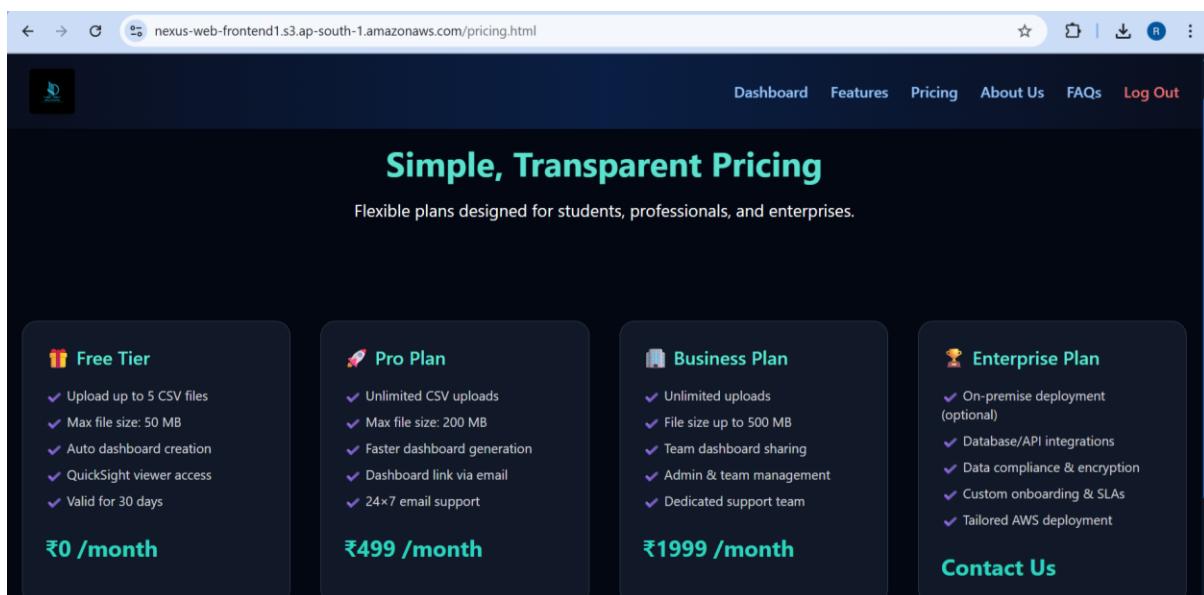
Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



The screenshot shows the 'Features' page of the Nexus DataLens 2.0 web application. At the top, there's a navigation bar with links for Dashboard, Features, Pricing, About Us, FAQs, and Log Out. Below the navigation, a main heading reads 'Powerful Features at Your Fingertips' with a rocket icon. A subtext below it says 'Nexus DataLens transforms your raw data into ready-to-analyze insights with a serverless, intelligent, and intuitive platform built on AWS.' The page is divided into six feature cards:

- Drag & Drop CSV Upload**: Upload CSV files via drag-and-drop or file selection using a sleek and intuitive interface.
- Automatic Data Cleaning**: Our AWS Glue integration automatically cleans and converts your uploaded data into an analysis-ready format.
- No Table Naming Needed**: Table names are intelligently assigned using metadata—no manual effort required.
- Instant Dashboard Creation**: Automatically generate interactive dashboards using Amazon QuickSight within seconds.
- User Notifications**: Receive updates or dashboard links directly via email. (Coming soon!)
- Serverless & Scalable**: Our 100% serverless AWS architecture scales seamlessly with your data volume and users.

A 'Show desktop' button is located in the bottom right corner of the card area.



The screenshot shows the 'Pricing' page of the Nexus DataLens 2.0 web application. At the top, there's a navigation bar with links for Dashboard, Features, Pricing, About Us, FAQs, and Log Out. Below the navigation, a main heading reads 'Simple, Transparent Pricing' with a rocket icon. A subtext below it says 'Flexible plans designed for students, professionals, and enterprises.' The page is divided into four pricing plan cards:

- Free Tier**: Includes upload up to 5 CSV files, max file size 50 MB, auto dashboard creation, QuickSight viewer access, and valid for 30 days. Priced at ₹0 /month.
- Pro Plan**: Includes unlimited CSV uploads, max file size 200 MB, faster dashboard generation, dashboard link via email, and 24x7 email support. Priced at ₹499 /month.
- Business Plan**: Includes unlimited uploads, file size up to 500 MB, team dashboard sharing, admin & team management, and dedicated support team. Priced at ₹1999 /month.
- Enterprise Plan**: Includes on-premise deployment (optional), database/API integrations, data compliance & encryption, custom onboarding & SLAs, and tailored AWS deployment. Priced at ₹TBA /month.

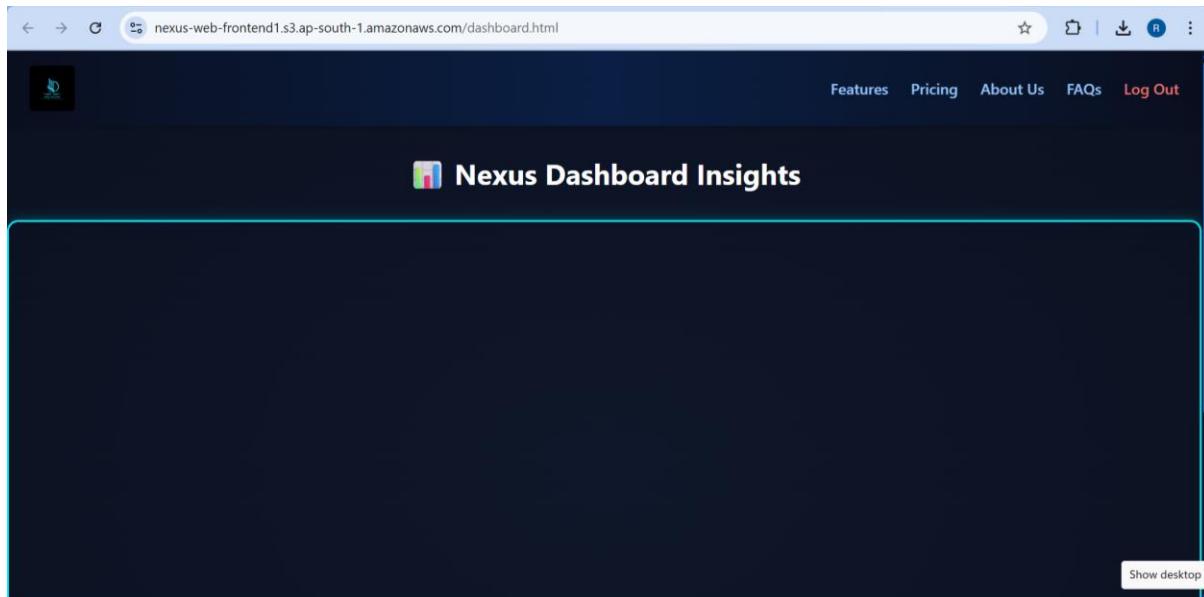
A 'Contact Us' button is located in the bottom right corner of the card area.

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the 'About' page of the Nexus DataLens website. At the top, there's a navigation bar with links to Dashboard, Features, Pricing, About Us, FAQs, and Log Out. Below the header, a main title 'Welcome to Nexus DataLens' is displayed, followed by a brief introduction: 'Nexus DataLens is your intelligent data companion designed to eliminate the complexities of Data Analysis. Whether you're a student, analyst, or founder, we've built this platform with you in mind.' Two main sections are visible: 'Why We Exist' and 'What Makes Us Unique'. Under 'Why We Exist', it says: 'Working with data shouldn't be intimidating. Traditional tools require technical skills, time, and a steep learning curve. Nexus DataLens changes that—our goal is to make data analytics as easy as uploading a file.' Under 'What Makes Us Unique', it says: 'We offer an end-to-end automated experience, from file upload to dashboard generation, with zero code involved. Our intelligent backend transforms raw CSVs into clean, structured data ready for insights—all through a smooth web interface.' At the bottom of the page, there's a section titled 'Our Technology Stack' which lists: AWS S3 for secure file storage, AWS Glue for schema detection and data cleaning, AWS Lambda for backend automation, and Amazon QuickSight for generating professional dashboards.

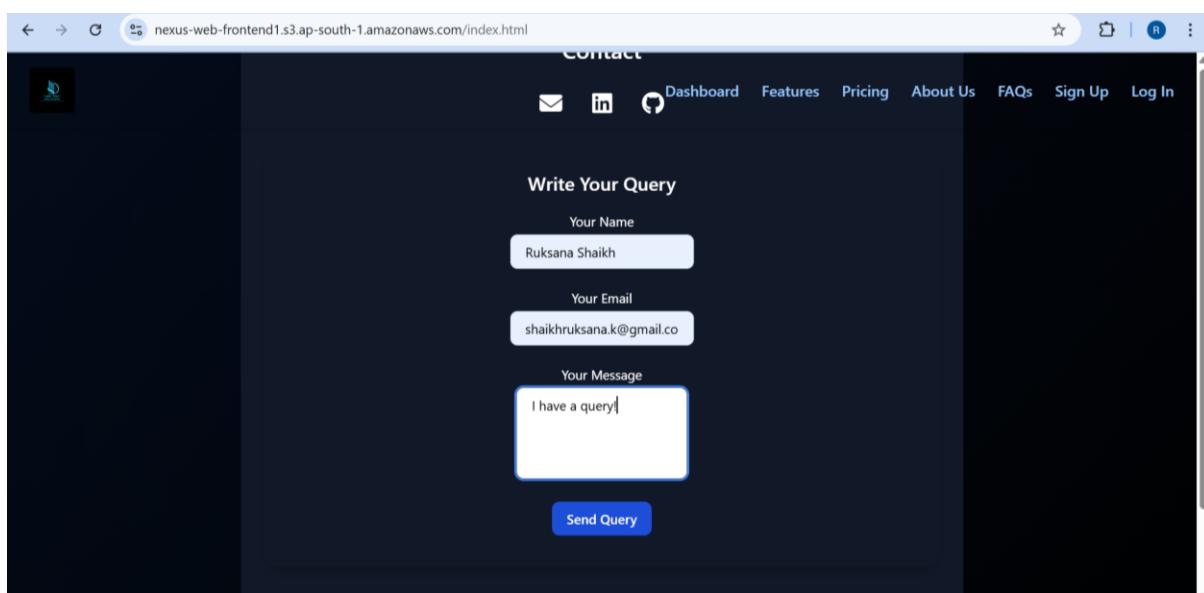
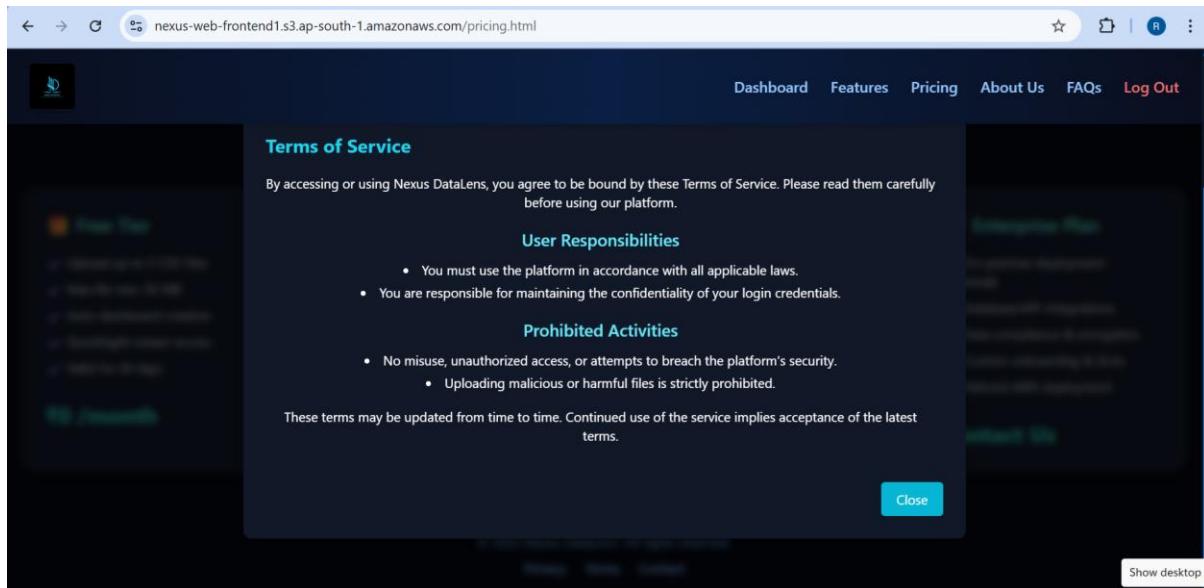
The screenshot shows the 'FAQs' page of the Nexus DataLens website. At the top, there's a navigation bar with links to Dashboard, Features, Pricing, About Us, FAQs, and Log Out. Below the header, a main title 'Frequently Asked Questions' is displayed. There are several questions listed with their corresponding answers: **What is Nexus DataLens?** (Nexus DataLens is a cloud-based platform that allows users to upload datasets (CSV, Excel, etc.), clean the data automatically, and generate insightful visualizations using AWS-powered services.) **Is login required to upload files?** (Yes. For security reasons and to provide personalized insights, users must sign up or log in before uploading files.) **Are my files and data secure?** (Absolutely. All files are stored in encrypted S3 buckets and accessed securely via API Gateway. Your data is never shared with third parties.) **Which file formats are supported?** (Currently, we support CSV, XLSX (Excel), and JSON file uploads. Support for databases and APIs is coming soon.) **How is my data cleaned?** (Our backend uses AWS Lambda and Glue Job scripts to remove null values, handle duplicates, normalize column formats, and detect common anomalies.) **Can I download cleaned data and reports?** (Yes. After the data is cleaned and processed, you can download the cleaned version and export insights as PDFs, images, or spreadsheets.)

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



A screenshot of a web browser displaying the 'Pricing' page. The URL in the address bar is 'nexus-web-frontend1.s3.ap-south-1.amazonaws.com/pricing.html'. The page features a dark blue header with 'Dashboard', 'Features', 'Pricing', 'About Us', 'FAQs', and 'Log Out' links. A central modal window is open, titled 'Privacy Policy'. It contains text about privacy, a 'Free Tier' section listing file upload details, and an 'Information We Collect' section with a bulleted list. To the right of the modal, a 'Enterprise Plan' section is visible with a list of features: 'On-premise deployment (optional)', 'Database/API integrations', 'Data compliance & encryption', 'Custom onboarding & SLAs', and 'Tailored AWS deployment'. At the bottom of the modal, there is a 'Close' button and copyright information: '© 2025 Nexus DataLens. All rights reserved.' followed by links to 'Privacy', 'Terms', and 'Contact'.

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



5.2 S3 Bucket Structure

Overview of the S3 bucket layout including folders for raw and cleaned data.

Name	Type	Last modified	Size	Storage class
about.html	html	July 22, 2025, 12:59:18 (UTC+05:30)	16.0 KB	Standard
athena-output/	Folder	-	-	-
cleaned-output/	Folder	-	-	-
contact.js	js	July 22, 2025, 12:59:18 (UTC+05:30)	1.0 KB	Standard
dashboard.html	html	July 26, 2025, 13:49:55 (UTC+05:30)	4.8 KB	Standard
Data/	Folder	-	-	-
faq.html	html	July 22, 2025, 12:59:18 (UTC+05:30)	12.1 KB	Standard
features.html	html	July 22, 2025, 12:59:18	6.7 KB	Standard

Name	Type	Last modified	Size	Storage class
(1).csv	CSV	July 21, 2025, 14:31:02 (UTC+05:30)	291.0 B	Standard
noisy_employee_data.csv	CSV	July 22, 2025, 12:56:26 (UTC+05:30)	11.2 KB	Standard
people_100.csv	CSV	July 21, 2025, 23:36:29 (UTC+05:30)	11.2 KB	Standard
people_101.csv	CSV	July 30, 2025, 13:32:54 (UTC+05:30)	11.2 KB	Standard
people-100.csv	CSV	July 22, 2025, 10:52:55 (UTC+05:30)	11.2 KB	Standard
people-101.csv	CSV	July 19, 2025, 02:36:57 (UTC+05:30)	591.0 B	Standard
Projects.csv	CSV	July 21, 2025, 19:36:57 (UTC+05:30)	47.0 B	Standard
s_8 (1).csv	CSV	-	-	-

5.3 Frontend File Upload and Hosting

The frontend is a **React-based web application**, hosted on **Amazon S3 (Static Website Hosting)** where users interact to upload .csv files.

- Uploaded files are securely stored in:
- `s3://nexus-web-frontend1/data/`
- **Amazon CloudFront** is used as a CDN to deliver the frontend globally with low latency.
- **AWS Cognito** enables **secure user authentication**, ensuring that only authorized users can upload CSV files.

❖ Technologies Used:

- Amazon S3 (Website Hosting)

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

- Amazon CloudFront (CDN)
- AWS Cognito (Authentication)

💡 Suggested Screenshots:

- Frontend UI with file upload section
- S3 bucket hosting the static site (console)
- Cognito user pool dashboard

The screenshot shows the AWS CloudFront Distributions page. On the left, there's a sidebar with links like Policies, Functions, Static IPs, VPC origins, and Reports & analytics. The main area has a table titled 'Distributions (1)'. The table has columns for ID, Status, Description, Type, Domain name..., Alternate dom..., Origins, and Last modified. One row is visible, showing 'E2G46IJUV2UH4K' with 'Enabled' status, 'Standard' type, and 'd1metrn98y5...' as the origin. A 'Create distribution' button is at the top right.

This screenshot shows the detailed view of the 'nexus' distribution. The left sidebar is identical to the previous screenshot. The main page has tabs for General, Security, Origins, Behaviors, Error pages, Invalidations, Tags, and Logging. The General tab is selected. It shows the distribution's name ('nexus'), ARN ('arn:aws:cloudfront:242201271328:distribution/E2G46IJUV2UH4K'), and last modified time ('July 21, 2025 at 11:52:40 AM UTC'). There are sections for Settings (Description, Price class, Supported HTTP versions), Continuous deployment (with a 'Create staging distribution' button), and Standard logging (which is off). Buttons for 'Edit' and 'View metrics' are also present.

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the 'Overview' page for a user pool named 'User pool - an7nzm'. The left sidebar contains navigation links for Amazon Cognito, Applications, User management, Authentication, and Security. The main content area displays 'User pool information' including the user pool name, token signing key URL, user pool ID, ARN, estimated number of users (5), and feature plan (Essentials). It also includes 'Recommendations' for setting up an app, applying branding to login pages, detecting risks, and setting up passwordless sign-in.

5.4 Secure File Storage in Amazon S3

- Upon upload, CSV files are stored in the raw data folder:
bash
Copy code
s3://nexus-web-frontend1/data/
 - AWS KMS (Key Management Service)** ensures encryption at rest.
 - Fine-grained access is controlled using **AWS IAM roles and policies**.

❖ Technologies Used:

- Amazon S3 (Data Storage)
- AWS IAM (Access Control)
- AWS KMS (Encryption)

📸 Suggested Screenshots:

- S3 bucket structure with uploaded CSV in data/
- IAM role used for S3 access
- KMS key policy for encryption

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the Amazon S3 console interface. On the left, there's a sidebar with navigation links like 'Amazon S3', 'General purpose buckets', 'Storage Lens', and 'CloudWatch Metrics'. The main area displays a list of objects (33) in the 'nexus-web-frontend1' bucket. The objects are all CSV files, mostly named 'test_1_s_x.csv' or 'people_100.csv'. The table includes columns for Name, Type, Last modified, Size, and Storage class. At the bottom right of the table, there are links for 'Edit', 'Simulate', 'Remove', and 'Add permissions'.

The screenshot shows the AWS IAM console. The left sidebar has sections for 'Identity and Access Management (IAM)', 'Access management', 'Access reports', and 'Access logs'. The main panel shows the 'Summary' of the 'NEXUS_USER_ROLE'. It includes details like 'Creation date' (July 19, 2025, 10:14 (UTC+05:30)), 'ARN' (arn:aws:iam::242201271328:role/NEXUS_USER_ROLE), and 'Last activity' (16 minutes ago). Below the summary, the 'Permissions' tab is selected, showing a list of four managed policies attached to the role: 'AmazonCognitoPowerUser', 'AmazonDynamoDBFullAccess', 'AmazonS3FullAccess', and 'AWSLambdaBasicExecutionRole'. Each policy is listed with its name, type (AWS managed), and the number of entities it's attached to.

The screenshot shows the AWS KMS console. The left sidebar has sections for 'Key Management Service (KMS)', 'Customer managed keys', and 'Custom key stores'. The main panel shows the 'General configuration' of a key with ID '0772dc11-a8b6-48cf-8a6e-b6109c991d43'. It includes fields for 'Alias' (alias/threatxray-key), 'Status' (Enabled), 'ARN' (arn:aws:kms:ap-south-1:242201271328:key/0772dc11-a8b6-48cf-8a6e-b6109c991d43), 'Description' (''), 'Creation date' (Jul 30, 2025 19:59 GMT+5:30), and 'Regionality' (Single Region). Below the general configuration, there are tabs for 'Key policy', 'Cryptographic configuration', 'Tags', 'Key material and rotations', and 'Aliases'. The 'Key policy' tab is selected, showing a JSON representation of the policy:

```

1 {
2   "Version": "2012-10-17",
3   "Id": "Key policy created by CloudTrail",
4   "Statement": [
5     {
6       "Sid": "Enable IAM User Permissions",
7       "Effect": "Allow",
8       "Action": "sts:AssumeRole"
9     }
10  ]
11 }
  
```

Project Workflow – Step-by-Step Breakdown

The **Nexus DataLens** project leverages a robust serverless architecture to automate the entire data lifecycle — from file upload to advanced business visualizations. Below is a comprehensive breakdown of how each AWS service integrates into the workflow:

6. Automation Using AWS Lambda

6.1 Trigger Mechanism

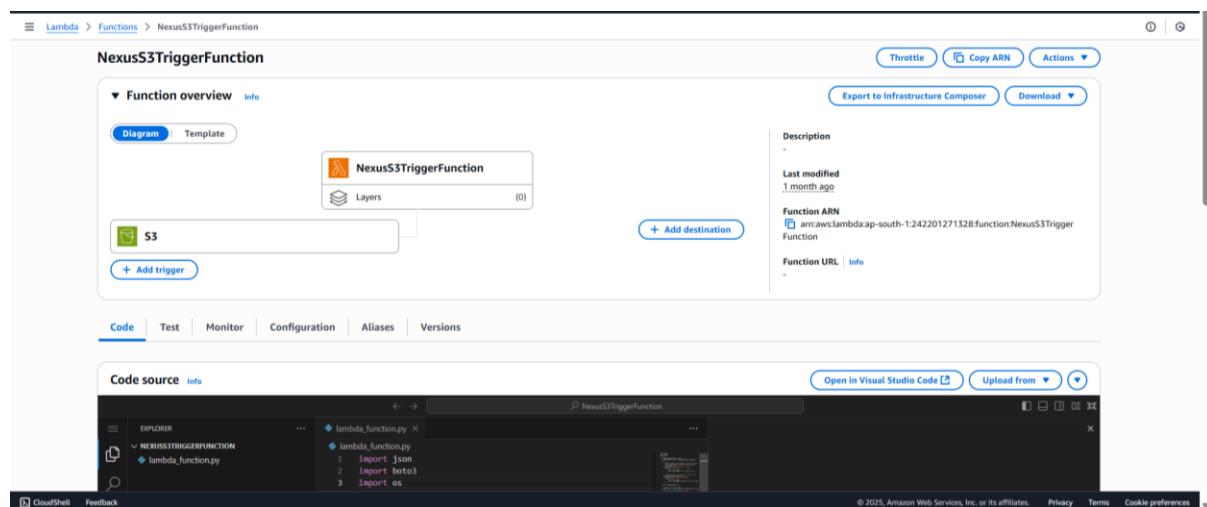
- An S3 PUT event notification automatically triggers the LambdaTriggerQuicksight Lambda function. It performs the following operations:
 - Invokes **Glue Crawler** to scan the raw file
 - Waits for crawler completion via polling logic
 - Runs the **Glue Job** for cleaning

❖ Technologies Used:

- AWS Lambda
- Amazon S3 Event Notifications
- AWS CloudWatch (Monitoring)

⌚ Suggested Screenshots:

- Lambda trigger config (S3 event → Lambda)
- Lambda code snippet (Python)
- CloudWatch logs showing process



Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the AWS Lambda function editor. The top navigation bar includes 'Lambda', 'Functions', 'NexusS3TriggerFunction', 'Code', 'Test', 'Monitor', 'Configuration', 'Aliases', and 'Versions'. The main area is titled 'Code source' with an 'Info' tab. It displays the 'lambda_function.py' file content:

```
lambda_function.py
1 import json
2 import boto3
3 import os
4
5 def lambda_handler(event, context):
6     print("Received event: " + str(event))
7
8     try:
9         bucket = event['Records'][0]['s3']['bucket']['name']
10        key = event['Records'][0]['s3']['object']['key']
11    except Exception as e:
12        print("Error parsing S3 event: " + str(e))
13        return {
14            'statusCode': 500,
15            'body': json.dumps('Error reading S3 event')
16        }
17
18    # Prevent infinite loop: skip files already cleaned
19    if key.startswith("cleaned-output/"):
20        print("Skipping cleaned file: " + key)
21        return {
22            'statusCode': 200,
23            'body': json.dumps("Skipped cleaned file: " + key)
24        }
25
```

The left sidebar shows the 'EXPLORER' view with 'lambda_function.py' selected. Below it are 'TEST EVENTS' and a 'DEPLOY' section with 'Deploy (Ctrl+Shift+U)' and 'Test (Ctrl+Shift+I)' buttons.

6.2 Lambda Trigger Function

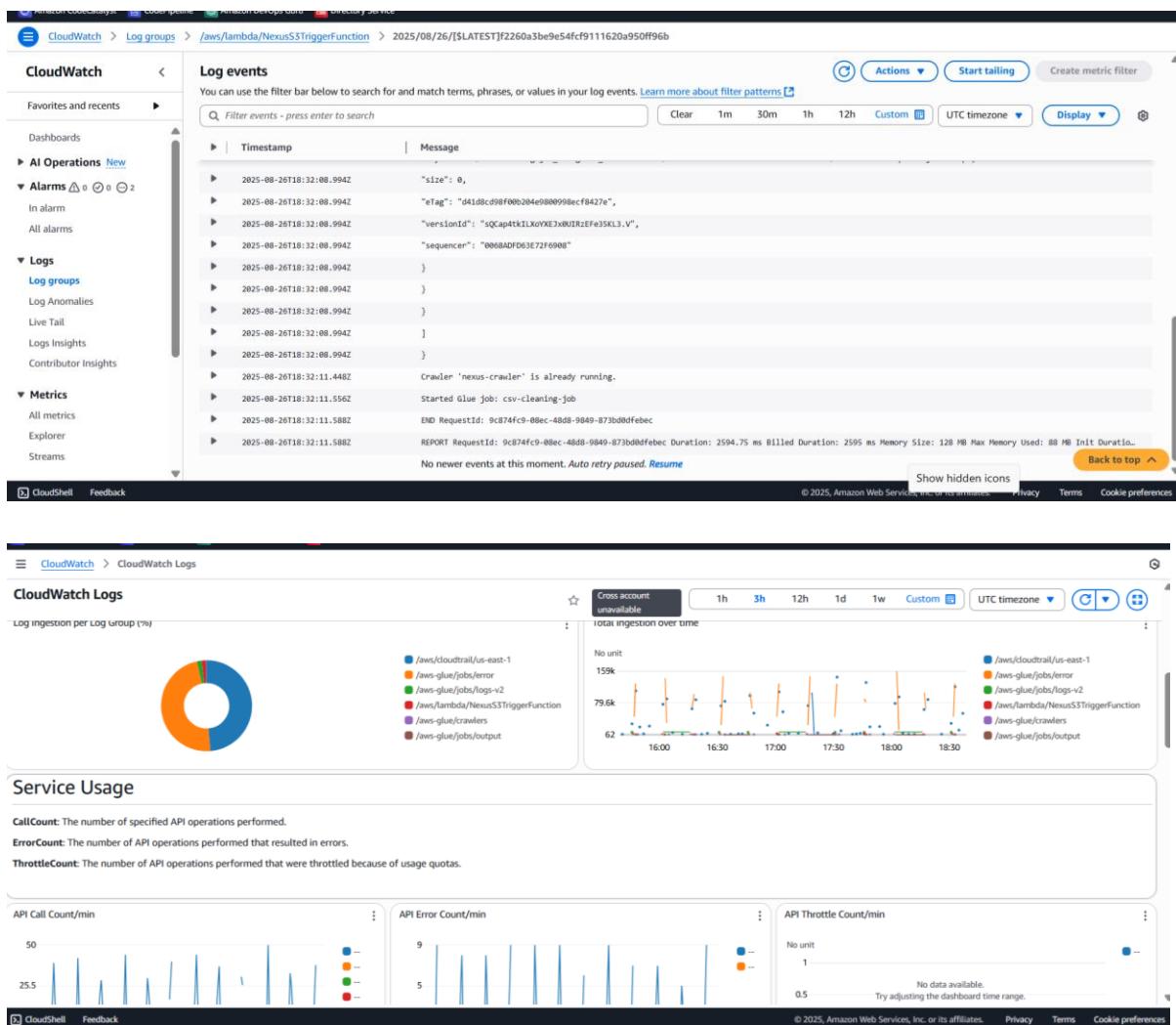
Lambda Trigger Function 2 → Automatically creates Athena-based datasets and QuickSight dashboards after Glue crawler finishes.

It ensures user-specific datasets are generated and visualized in QuickSight dynamically.

The screenshot shows the AWS CloudWatch Log groups interface. The left sidebar includes 'CloudWatch' and sections for 'AI Operations', 'Logs', and 'Metrics'. The main area shows the log group '/aws/lambda/NexusS3TriggerFunction' with 'Log streams (100+)'. A table lists log streams with their last event time:

Last event time
2025-08-26 18:32:11 (UTC)
2025-08-26 18:31:01 (UTC)
2025-08-26 18:16:04 (UTC)
2025-08-26 18:02:21 (UTC)
2025-08-26 18:01:25 (UTC)
2025-08-26 17:46:17 (UTC)
2025-08-26 17:31:27 (UTC)

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



7. Data Processing Pipeline

7.1 Schema Detection with AWS Glue Crawler

- The Glue Crawler scans the uploaded file in the data/ folder, detects the schema, and updates the **AWS Glue Data Catalog**.
- The resulting table is used by **Athena** for SQL querying.

❖ Technologies Used:

- AWS Glue Crawler**
- AWS Glue Catalog**

📸 Screenshot Suggestions:

- Glue Crawler configuration page
- Table in Glue Catalog pointing to data/ S3 path

Name	State	Last run	Last run timestamp	Log	Table changes from last run
cleaned-output-crawler	Ready	Succeeded	July 21, 2025 at 12:46:58 UTC	View log	45 created
nexus-crawler	Ready	Succeeded	August 26, 2025 at 18:05:17 UTC	View log	-

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
August 26, 2025 at 18:46:02	August 26, 2025 at 18:46:58	55 s	Completed	0.064	-
August 26, 2025 at 18:31:03	August 26, 2025 at 18:32:30	01 min 26 s	Completed	0.073	-
August 26, 2025 at 18:16:03	August 26, 2025 at 18:17:24	01 min 20 s	Completed	0.071	-
August 26, 2025 at 18:01:24	August 26, 2025 at 18:05:17	01 min 52 s	Completed	0.045	-
August 26, 2025 at 17:46:17	August 26, 2025 at 17:47:23	01 min 06 s	Completed	0.072	-

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the AWS Glue Tables page. The left sidebar navigation includes 'AWS Glue', 'Data Catalog', and 'Data Integration and ETL'. The main content area displays a table titled 'Tables (1/33)' with columns: Name, Database, Location, Classification, Deprecated, View data, Data quality, and Column stats... The table lists various CSV and JSON files from the 'nexusdb' database, such as 'deliveries_csv', 'heart_csv', 'index_html', etc. A specific row for 'people_100_csv' is selected.

The screenshot shows the AWS Glue Schema details for the 'people_100_csv' table. The left sidebar navigation is identical to the previous screenshot. The main content area shows the schema definition with 9 columns: index (bigint), user id (string), first name (string), last name (string), sex (string), email (string), phone (string), date of birth (string), and job title (string). The 'Schema' tab is selected.

The screenshot shows the AWS Glue Studio Jobs page. The left sidebar navigation includes 'AWS Glue', 'Data Catalog', and 'Data Integration and ETL'. The main content area displays a table titled 'Your jobs (1)' with columns: Job name, Type, Created by, Last modified, and AWS Glue version. One job, 'csv-cleaning-job', is listed, created by 'Glue ETL' and last modified on 8/13/2025 at 11:20:46 AM, using version 5.0. A 'Create example job' button is visible.

7.2 Glue Job – Data Cleaning

The Spark-based job csv-cleaning-job processes the raw data:

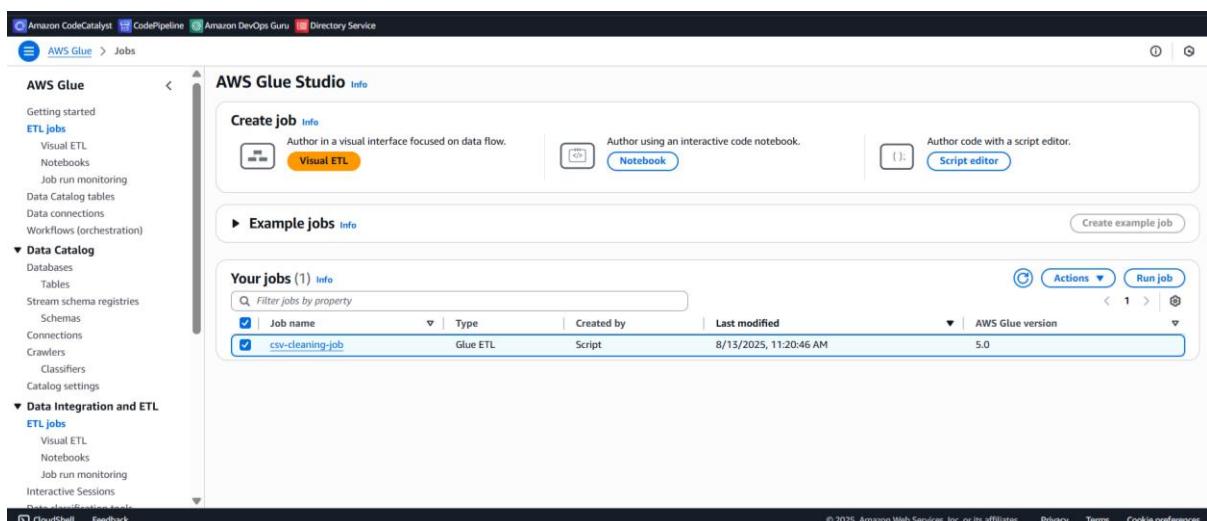
- Drops rows and columns that are entirely null
- Filters out invalid values
- Standardizes column headers
- Repartitions to a single flat file

Cleaned output is written to:

s3://nexus-web-frontend1/cleaned-output/<filename>.csv

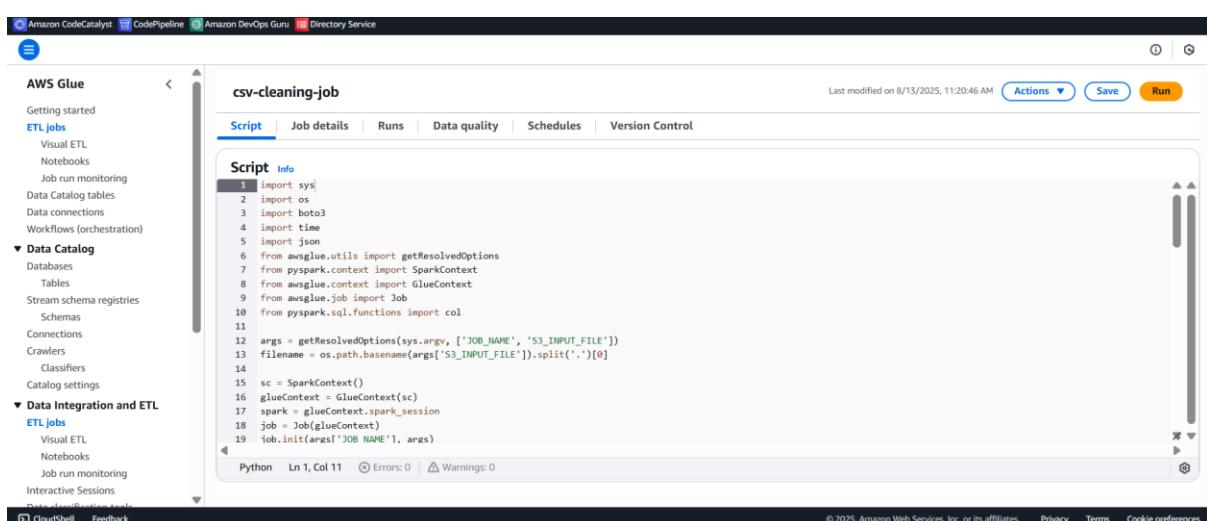
Screenshot Suggestions:

- Glue job script code
- Sample cleaned output file in S3
- Before vs After comparison of the CSV (raw vs cleaned)



The screenshot shows the AWS Glue Studio interface. On the left, there's a sidebar with navigation links for AWS Glue, ETL jobs, Data Catalog, and Data Integration and ETL. The main area is titled "AWS Glue Studio" and shows the "Create job" interface. It has three tabs: "Visual ETL" (selected), "Notebook", and "Script editor". Below this is a section for "Example jobs" with a "Create example job" button. The main content area is titled "Your jobs (1) Info" and lists a single job named "csv-cleaning-job". The job details are as follows:

Job name	Type	Created by	Last modified	AWS Glue version
csv-cleaning-job	Glue ETL	Script	8/13/2025, 11:20:46 AM	5.0



This screenshot shows the details page for the "csv-cleaning-job". The top navigation bar includes "Actions", "Save", and "Run" buttons. The main content area is titled "Script" and contains the Python code for the job:

```

1 import sys
2 import os
3 import boto3
4 import time
5 import json
6 from awsglue.utils import getResolvedOptions
7 from pyspark.context import SparkContext
8 from awsglue.context import GlueContext
9 from awsglue.job import Job
10 from pyspark.sql.functions import col
11
12 args = getResolvedOptions(sys.argv, ['JOB_NAME', 'S3_INPUT_FILE'])
13 filename = os.path.basename(args['S3_INPUT_FILE']).split('.')[0]
14
15 sc = SparkContext()
16 glueContext = GlueContext(sc)
17 spark = glueContext.spark_session
18 job = Job(glueContext)
19 job.init(args['JOB_NAME'], args)

```

The status bar at the bottom indicates "Last modified on 8/13/2025, 11:20:46 AM".

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The image contains two screenshots of the Amazon S3 console. Both screenshots show the 'Objects' tab for the 'nexus-web-frontend1' bucket.

Screenshot 1 (Top): This screenshot shows the 'cleaned-output/' folder expanded. It contains the following objects:

Name	Type	Last modified	Size	Storage class
about.html	html	July 22, 2025, 12:59:18 (UTC+05:30)	16.0 KB	Standard
athena-output/	Folder	-	-	-
cleaned-output/	Folder	-	-	-
contact.js	js	July 22, 2025, 12:59:18 (UTC+05:30)	1.0 KB	Standard
dashboard.html	html	July 26, 2025, 13:49:35 (UTC+05:30)	4.8 KB	Standard
Data/	Folder	-	-	-
faq.html	html	July 22, 2025, 12:59:18 (UTC+05:30)	12.1 KB	Standard
feature.html	html	July 22, 2025, 12:59:18	6.7 KB	Standard

Screenshot 2 (Bottom): This screenshot shows the 'people_100.csv' file selected in the 'cleaned-output/' folder. It contains the following objects:

Name	Type	Last modified	Size	Storage class
people_100.csv	csv	August 6, 2025, 09:54:49 (UTC+05:30)	11.1 KB	Standard
people_101.csv	csv	July 21, 2025, 23:37:41 (UTC+05:30)	11.1 KB	Standard
people-100/	Folder	-	-	-
people-101.csv	csv	July 30, 2025, 13:55:32 (UTC+05:30)	11.1 KB	Standard
Projects/	Folder	-	-	-
quicksight-embedding-js-sdk.csv	csv	July 22, 2025, 10:54:00 (UTC+05:30)	311.0 B	Standard
s_8.csv	csv	July 21, 2025, 19:28:45 (UTC+05:30)	47.0 B	Standard
s_8/	Folder	-	-	-
sahils/	Folder	-	-	-
test_1_e_F.csv	CSV	July 21, 2025, 19:47:04	47.0 B	Standard

7.3 Advanced Cleaning with Glue DataBrew (*Optional*)

- For more complex or no-code transformations, **AWS Glue DataBrew** jobs can be triggered.
- It enables:
 - Data profiling
 - Pattern matching
 - Custom transformations

❖ Technologies Used:

- AWS Glue DataBrew

💡 Suggested Screenshots:

- DataBrew project and recipe
- Preview of cleaned results
- Profile summary (data quality insights)

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the DataBrew interface under the 'Datasets' tab. It displays a list of datasets with one entry: 'nexus-dataset'. The dataset details are as follows:

Dataset name	Data type	Data profile	Source	Location	Create date	Created by	Tags
nexus-dataset	CSV	View data profile	S3	s3://nexus-web-frontend1/Data/noisy_employee_data.csv	2 months ago June 21, 2025, 11:35:47 am	-	-

Below the dataset list, there is a 'Dataset preview' section showing the first few rows of the CSV file:

# Emp_ID	At: Name	# Age	At: Dept	At: Join_Date	# \$
101	Ali	28	HR	2020-01-15	50000
102	Zara	31	IT	2019-05-10	60000
null	Ali	twenty-nine	HR	2020-01-15	50000
104	null	40	Finance	invalid-date	70000
105	Ahmed	null	IT	2021-07-01	NA
106	Sara	25	HR	null	55000
106	Sara	25	HR	2019/05/10	55000
107	@invalid	null	---	null	??

The screenshot shows the DataBrew interface under the 'Projects' tab. It displays a list of projects with one entry: 'nexus-cleaning-project'. The project details are as follows:

Project name	Associated dataset	Attached recipe	Jobs	Create date	Created by	In use by	Tags
nexus-cleaning-project	nexus-dataset	nexus-cleaning-project-recipe	-	2 months ago June 21, 2025, 11:53:35 am	-	-	-

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the Nexus DataLens interface. On the left, there's a sidebar with 'PROJECTS' selected, showing a tree view of datasets, recipes, DQ rules, jobs, and what's new. The main area displays a dataset named 'nexus-dataset'. A progress bar indicates 'Your session will be ready soon! Provisioning compute' at 0%. The right panel shows the 'Recipe (8)' for the dataset, listing steps such as removing duplicates from Name, Emp_ID, Age, Dept, Join_Date, and Salary, and deleting empty rows with missing values in Emp_ID and Name. The status bar at the bottom right shows '© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences'.

The screenshot shows the dataset sample with 2 rows. The columns are labeled '# Emp_ID', 'ABC Name', 'ABC Age', 'ABC Dept', and 'ABC Join_Date', 'ABC Salary'. The data shows two employees: Ali (Age 28, Dept HR, Join Date 2020-01-15, Salary 50000) and Zara (Age 31, Dept IT, Join Date 2019-05-10, Salary 60000). The status bar at the bottom right shows '© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences'.

The screenshot shows the dataset sample after cleaning, with 2 rows. The columns are labeled '# Emp_ID', 'ABC Name', 'ABC Age', 'ABC Dept', 'ABC Join_Date', and 'ABC Salary'. The data remains the same as the previous screenshot. The status bar at the bottom right shows '© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences'.

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the DataBrew interface with the 'RECIPES' tab selected in the sidebar. The main area displays a list of recipes under 'Recipes (1) Info'. A single recipe, 'nexus-cleaning-project-recipe', is listed with its details: Version 4.0, Associated project 'nexus-cleaning-project', Published date '2 months ago' (July 10, 2025, 5:23:13 pm). Action buttons include 'Download as YAML', 'Download as JSON', 'Create job with this recipe', 'Actions', and 'Upload recipe'.

Recipe name	Version description	Associated projects	Published date
nexus-cleaning-project-recipe	Published version 4.0	nexus-cleaning-project	2 months ago July 10, 2025, 5:23:13 pm

This screenshot shows the detailed view of the 'nexus-cleaning-project-recipe'. It includes tabs for 'Recipe steps' and 'Data lineage'. The 'Recipe details' section shows the recipe name 'nexus-cleaning-project-recipe', associated project 'nexus-cleaning-project', and published date '2 months ago by arn:aws:iam::242201271328:root'. The 'Recipe steps (8)' section lists eight steps: Remove duplicates from Name, Emp_ID, Age, Dept, Join_Date, and Salary.

The screenshot shows the DataBrew interface with the 'JOBS' tab selected in the sidebar. The main area displays a list of jobs under 'Recipe jobs (1) Info'. A single job, 'csv-cleaning-job-final', is listed with its status 'Succeeded', Job input 'nexus-dataset', Job output '1 output', Last run 'a month ago' (July 12, 2025, 2:04:22 pm), and Created on '2 months ago' (July 10, 2025, 5:35:10 pm). Action buttons include 'View details', 'Run job', 'Actions', and 'Create job'.

Job name	Status	Job input	Job output	Last run	Created on
csv-cleaning-job-final	Succeeded	nexus-dataset	1 output	a month ago July 12, 2025, 2:04:22 pm	2 months ago July 10, 2025, 5:35:10 pm

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the DataBrew interface with the 'Jobs' section selected. A specific job named 'csv-cleaning-job-final' is displayed, which uses the dataset 'nexus-dataset' and the recipe 'nexus-cleaning-project-recipe'. The 'Job run history' tab is active, showing two successful runs. The first run, 'csv-cleaning-job-final_2025-07-12-14:02:24', completed successfully in 1 minute and 47 seconds, producing 1 output. The second run, 'csv-cleaning-job-final_2025-07-10-17:55:11', also completed successfully in 1 minute and 44 seconds, producing 1 output. Both runs were started by 'arnaws:iam::242201271328:root'.

7.4. Second Lambda Invocation

Once the Glue Job completes, it **invokes a second Lambda function** to automate the following:

- Trigger a second Glue Crawler for the cleaned data
- Configure Athena dataset
- Automatically generate a QuickSight dashboard using a predefined template

❖ Technologies Used:

- AWS Lambda
- AWS Glue Crawler
- Athena dataset creation via Boto3
- QuickSight automation via SDK

💡 Screenshot Suggestions:

- Lambda code snippet invoking dashboard generation
- Glue Crawler for cleaned output

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the AWS Glue interface with the 'Crawlers' section selected. There are two crawlers listed:

Name	State	Last run	Last run timestamp	Log	Table changes from last run
cleaned-output-crawler	Ready	Succeeded	July 21, 2025 at 12:40:35	View log	45 created
nexus-crawler	Ready	Succeeded	August 26, 2025 at 18:46:02	View log	-

The screenshot shows the 'cleaned-output-crawler' properties page. It includes sections for Crawler properties, Crawler runs, and a detailed table of recent runs.

Crawler properties:

- Name: cleaned-output-crawler
- IAM role: GlueDataBrewRole_1
- Description: -
- Security configuration: -
- Database: nexusdb_3
- Lake Formation configuration: -
- Table prefix: -
- State: READY

Crawler runs (5):

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
July 21, 2025 at 12:40:35	July 21, 2025 at 12:41:55	01 min 20 s	Completed	0.066	45 table changes, 5 partition changes
July 18, 2025 at 11:25:04	July 18, 2025 at 11:25:57	52 s	Completed	0.071	3 table changes, 4 partition changes
July 18, 2025 at 10:41:31	July 18, 2025 at 10:42:24	52 s	Completed	0.059	-
July 18, 2025 at 10:29:33	July 18, 2025 at 10:30:30	56 s	Completed	0.083	2 table changes, 0 partition changes
July 12, 2025 at 15:14:02	July 12, 2025 at 15:14:54	52 s	Completed	0.079	1 table change, 0 partition changes

8. Data Storage & Querying

8.1 Cleaned Data Output in S3

Name	Type	Last modified	Size	Storage class
about.html	html	July 22, 2025, 12:59:18 (UTC+05:30)	16.0 KB	Standard
athena-output/	Folder	-	-	-
cleaned-output/	Folder	-	-	-
contact.js	js	July 22, 2025, 12:59:18 (UTC+05:30)	1.0 KB	Standard
dashboard.html	html	July 26, 2025, 13:49:35 (UTC+05:30)	4.8 KB	Standard
Data/	Folder	-	-	-
faq.html	html	July 22, 2025, 12:59:18 (UTC+05:30)	12.1 KB	Standard
features.html	html	July 22, 2025, 12:59:18	6.7 KB	Standard

Name	Type	Last modified	Size	Storage class
people_100.csv	csv	August 6, 2025, 09:54:49 (UTC+05:30)	11.1 KB	Standard
people_101.csv	csv	July 21, 2025, 23:37:41 (UTC+05:30)	11.1 KB	Standard
people-100/...	Folder	-	-	-
people-101.csv	csv	July 30, 2025, 13:55:32 (UTC+05:30)	11.1 KB	Standard
Projects/...	Folder	-	-	-
quickstarts-embedding-js-sdk.csv	csv	July 22, 2025, 11:02:27 (UTC+05:30)	311.0 B	Standard
s_8.csv	csv	July 21, 2025, 19:28:45 (UTC+05:30)	47.0 B	Standard
s_8/...	Folder	-	-	-
sahil/...	Folder	-	-	-
text_1_e_F.csv	csv	July 21, 2025, 19:47:04	47.0 B	Standard

8.2 Querying via Athena

- Athena enables serverless querying using standard SQL.
- Cleaned data in S3 is queried via the Glue Catalog tables.

Example query:

```
SELECT * FROM nexusdb.cleaned_<filename> LIMIT 10;
```

Screenshot Suggestions:

- Athena query editor with sample query

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

- Preview of query results

The screenshot shows the Amazon Athena Query editor interface. On the left, the 'Data' sidebar displays the 'Data source' as 'AwsDataCatalog', 'Catalog' as 'None', and 'Database' as 'nexusdb'. Below this, the 'Tables and views' section lists 'people' and 'Tables (5)'. Under 'Tables (5)', there are three entries: 'people_100' (Partitioned), 'people_100_csv', and 'people_100.csv_796046d9754dff7d59e 4bf849511a231'. The main workspace shows 'Query 1' with the SQL query: 'SELECT * FROM nexusdb.people_100 LIMIT 10;'. Below the query, the status bar indicates 'SQL Ln 1, Col 33'. At the bottom of the editor, there are buttons for 'Run again', 'Explain', 'Cancel', 'Clear', and 'Create'. A 'Query results' tab is selected, and a note says 'Reuse query results up to 60 minutes ago'. The footer includes links for CloudShell, Feedback, Privacy, Terms, and Cookie preferences.

The screenshot shows the results of the query 'SELECT * FROM nexusdb.people_100 LIMIT 10;' in the Amazon Athena Query editor. The results are displayed in a table titled 'Results (10)'. The table has 10 rows and 6 columns, labeled 'col0' through 'col6'. The columns represent attributes such as 'User Id', 'First Name', 'Last Name', 'Sex', 'Email', and 'Phone'. The data includes various names like Elijah, Shelby, Terrell, Phillip, Summers, Kristine, Travis, etc., along with their corresponding email addresses and phone numbers. The footer includes links for CloudShell, Feedback, Privacy, Terms, and Cookie preferences.

#	col0	col1	col2	col3	col4	col5	col6
1	Index	"User Id"	"First Name"	"Last Name"	Sex	Email	Phone
2	1	88F7B33d2bcf9f5	Shelby	Terrell	Male	elijah57@example.net	001-084-906-7849x73518
4	2	f90cD3E76f1A9b9	Phillip	Summers	Female	bethany14@example.com	214.112.6044x4913
5	3	DbeAb8CcdefC2c	Kristine	Travis	Male	bthompson@example.com	277.609.7938
6	4	A31Bee3c201ef58	Yesenia	Martinez	Male	"kaitlinkaiser@example.com"	584.094.6111
7	5	1bA7A3dc874da3c	Lori	Todd	Male	"buchananmanuel@example.net"	689-207-3558x7233
8	6	bfDD7CDEF5D865B	Erin	Day	Male	tconner@example.org	001-171-649-9856x5553
9	7	bE9EEF34cB72AF7	Katherine	Buck	Female	conniecowan@example.com	"+1-773-151-6685x49162"
10	8	2EFC6A4e77FaEaC	Ricardo	Hinton	Male	wyattbishop@example.com	001-447-699-7998x88612

9. Visualization and Reporting

9.1 Data Visualization – QuickSight & Power BI

- In **Amazon QuickSight**, dashboards are generated automatically using Athena datasets.
- In **Microsoft Power BI**, users can connect via the **Athena ODBC driver** for custom reporting.

❖ Technologies Used:

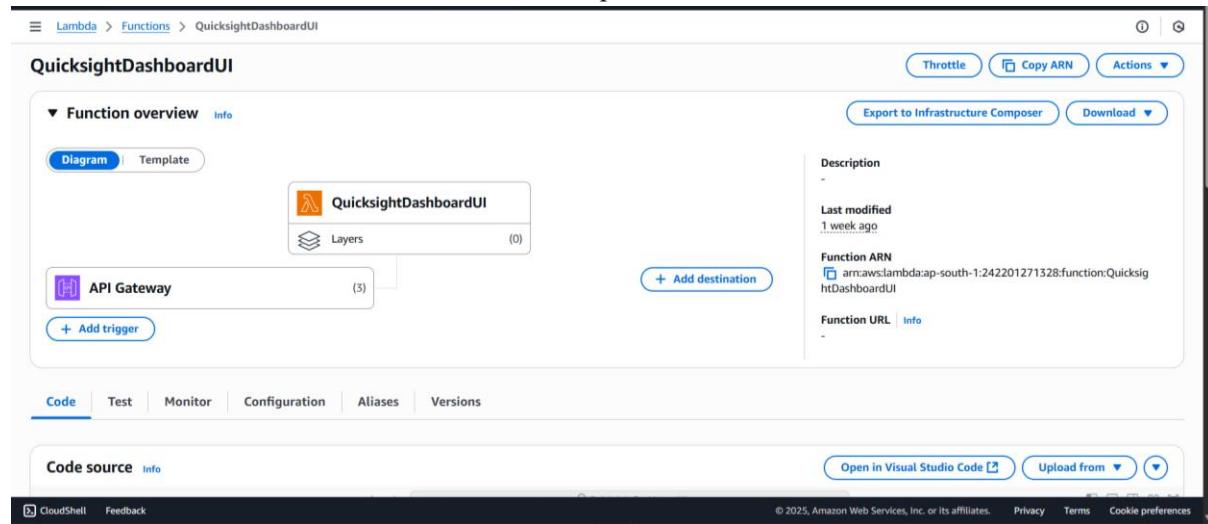
- Amazon QuickSight
- Microsoft Power BI (Optional)
- Athena ODBC Connector

📸 Suggested Screenshots:

- QuickSight dashboards (bar, pie, line)
- Power BI report preview (same dataset)
- ODBC config panel in Power BI

📸 Screenshot Suggestions:

- QuickSight dashboard with graphs/charts
- Power BI connected to Athena with a sample visual



Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The image displays three screenshots of the QuickSight interface, illustrating the Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS.

Screenshot 1: Dataset Management

This screenshot shows the "people_100" dataset summary page. It includes sections for About (Size: 34.4 KB), Refresh (Status: Completed, Last successful refresh: July 21, 2025 at 11:46 PM GMT+5:30), Access Settings (Sharing: Owners (1) Viewers (0), Row-level security: No restrictions, Column-level security: No restrictions), Schema (Unique key: Learn more, Disabled), Sources (nexusdb), and Usage (Analyses (1), Dashboards (1), Datasets (0)).

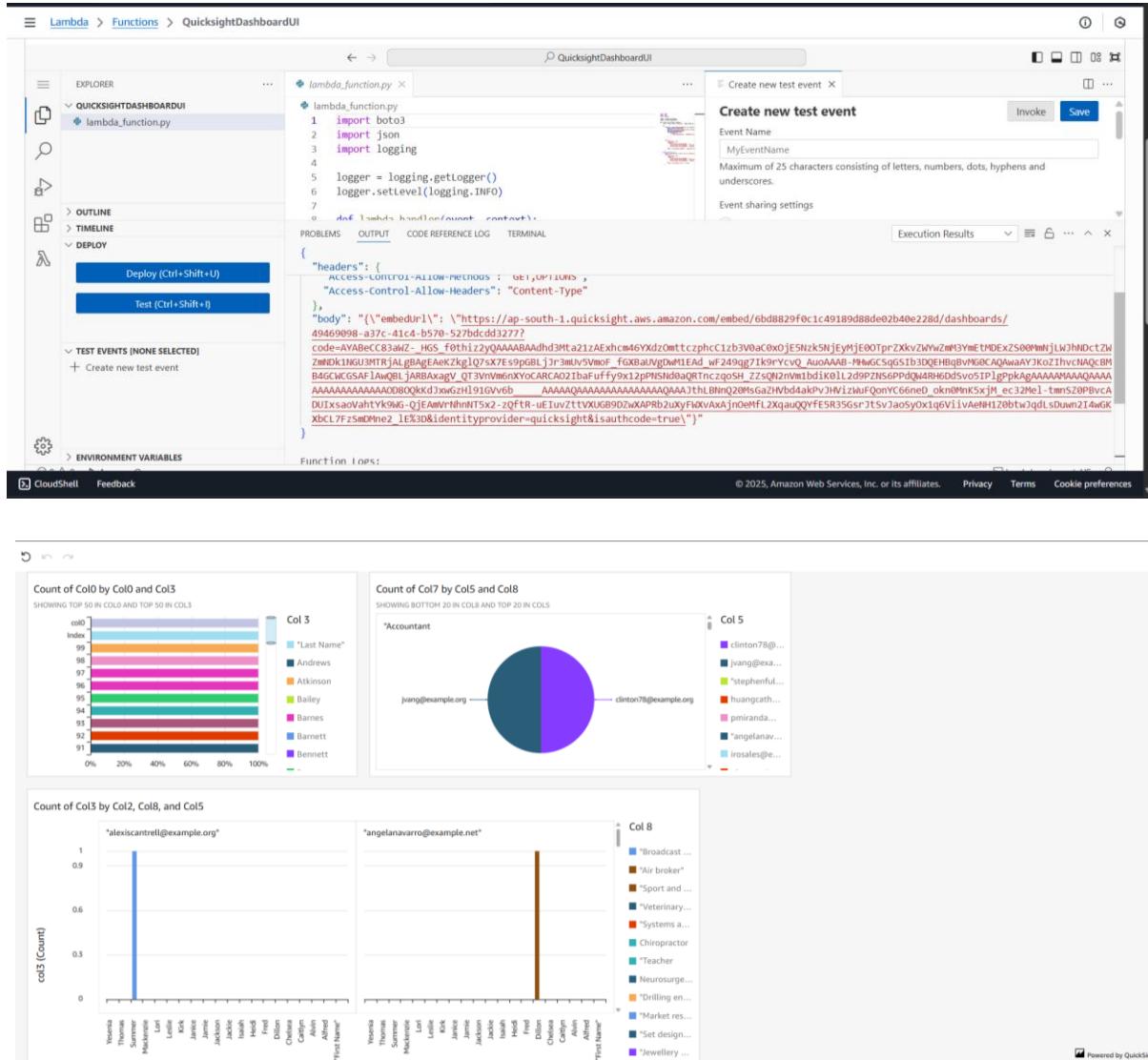
Screenshot 2: Analysis Creation

This screenshot shows the "people_100 analysis" page. It features a Data panel on the left listing fields like col0 through col8 and partition_0. The Visuals panel on the right contains four visualizations: a horizontal bar chart titled "Count of Col0 by Col0 and Col3" showing top 102 in Col0 and bottom 25 in Col3; a pie chart titled "Count of Col7 by Col5 and Col8" showing bottom 20 in Col8 and top 20 in Col5; a bar chart titled "Count of Col3 by Col2, Col8, and Col5" for "alexiscantrell@example.org"; and a bar chart titled "Count of Col3 by Col2, Col8, and Col5" for "angelanavarro@example.net".

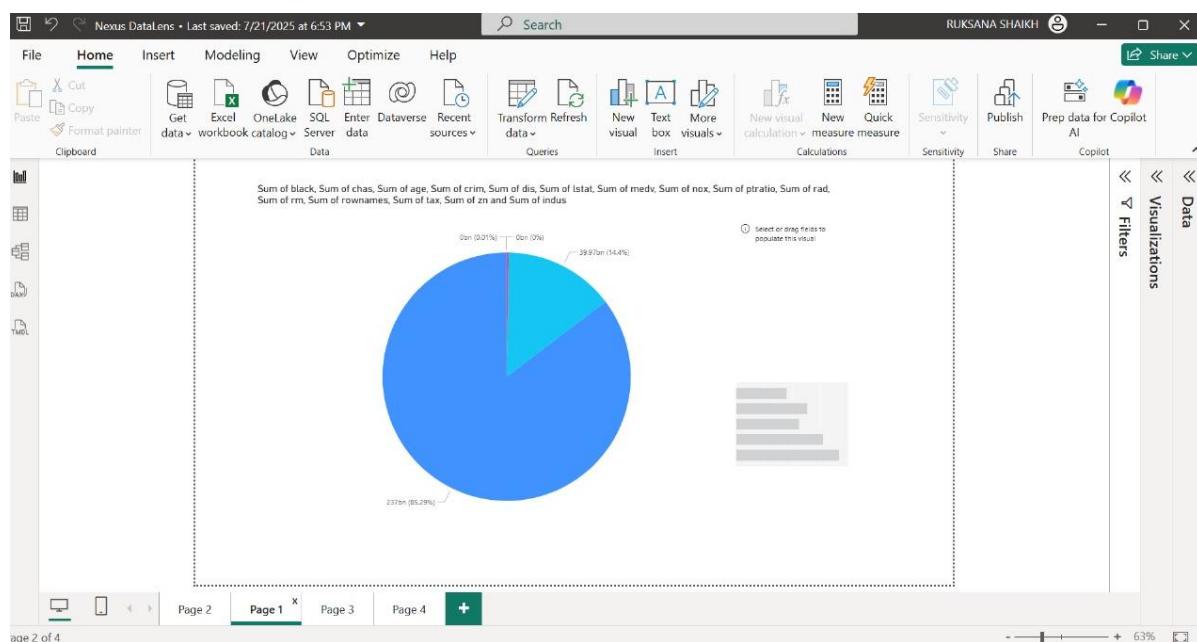
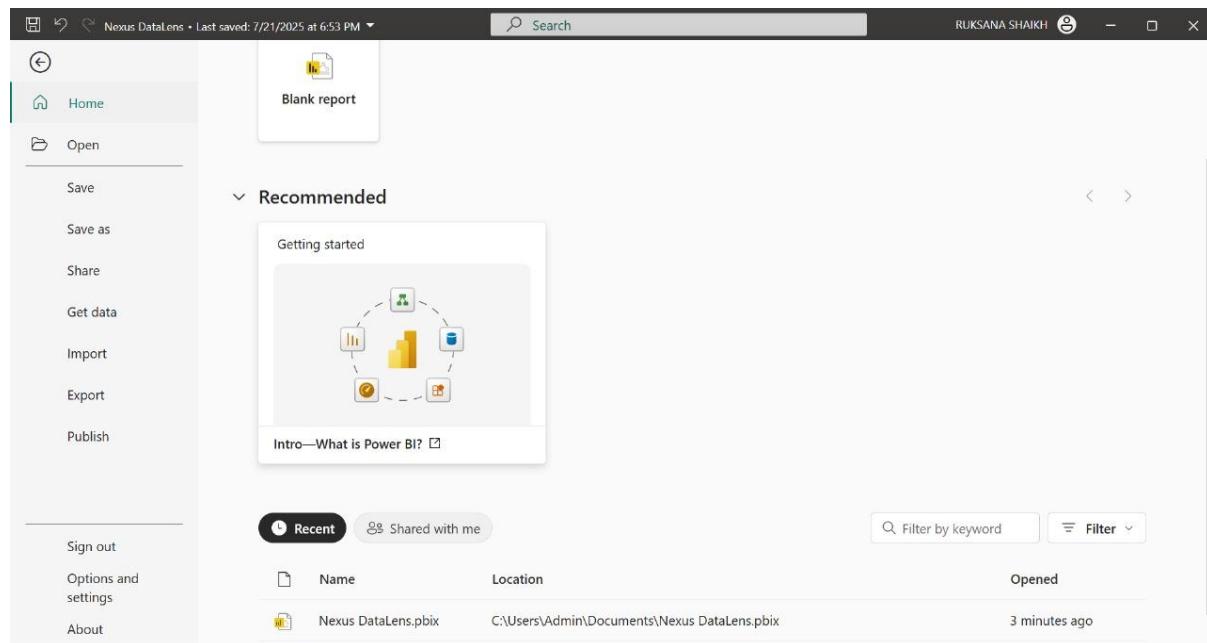
Screenshot 3: Dashboard Creation

This screenshot shows the "Original dashboard" page. It features two main visualizations: a treemap titled "Average Tenure and Monthly Compensation" showing monthly compensation by business function (APAC, EMEA, US) and tenure (4.74K to 10.12K); and a stacked bar chart titled "Education profile by Business function" showing education levels (PHD, MS, BS, Associate, ASSOC) across business functions (Sales, Operations, Marketing, Development, Corporate).

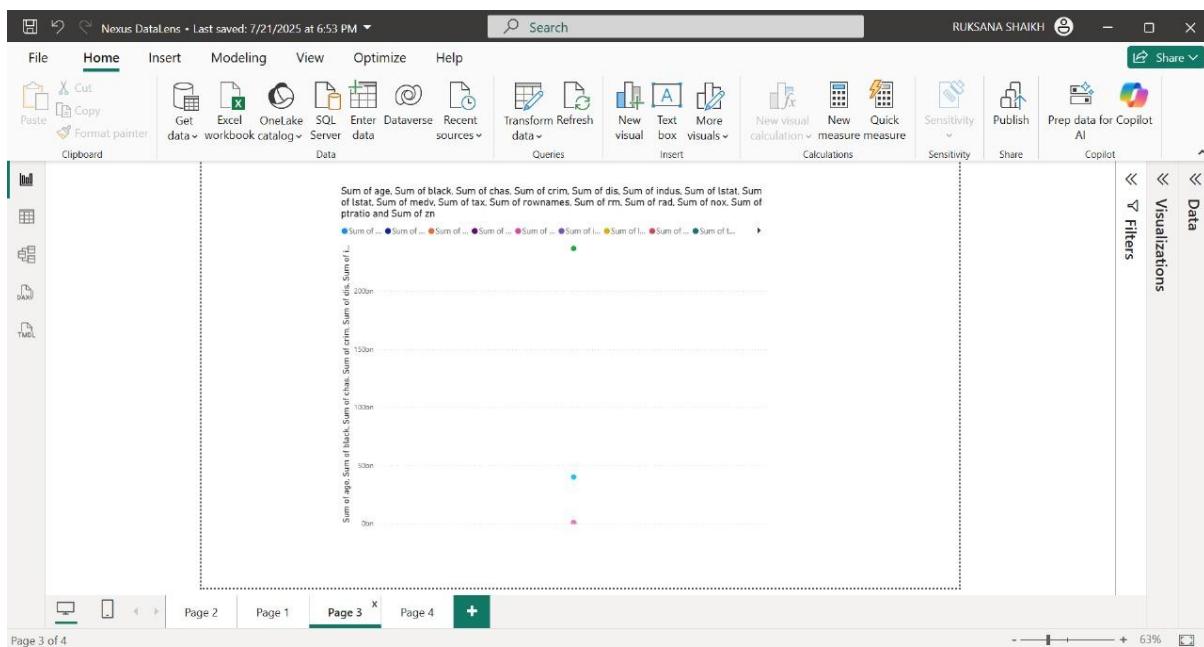
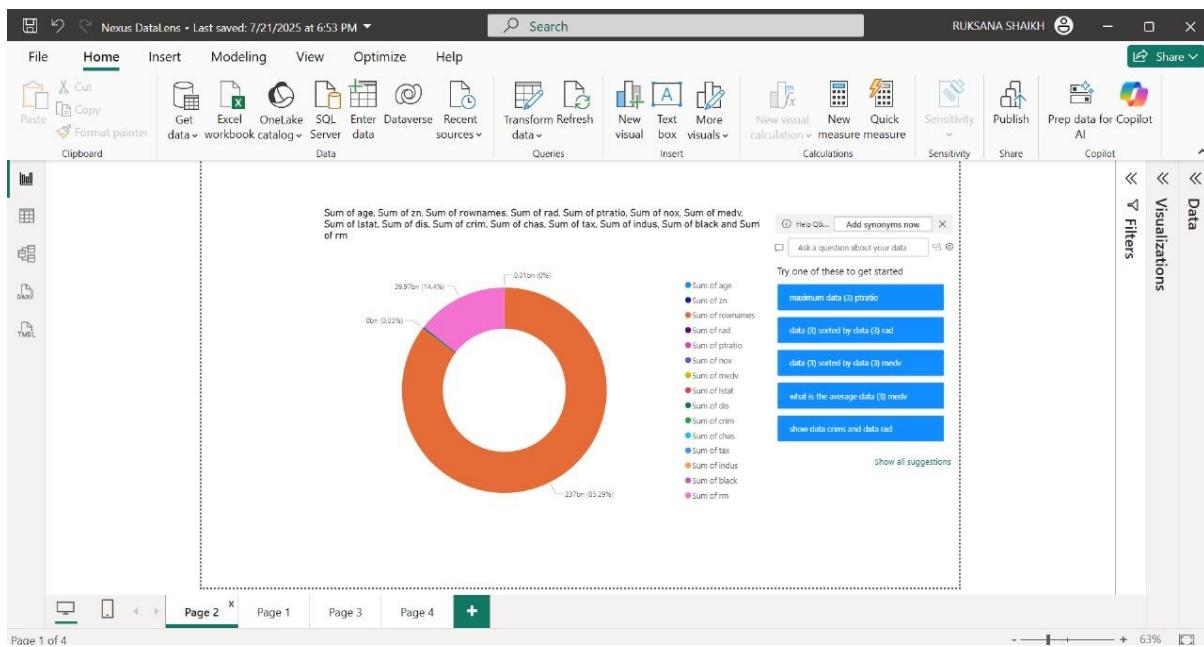
Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



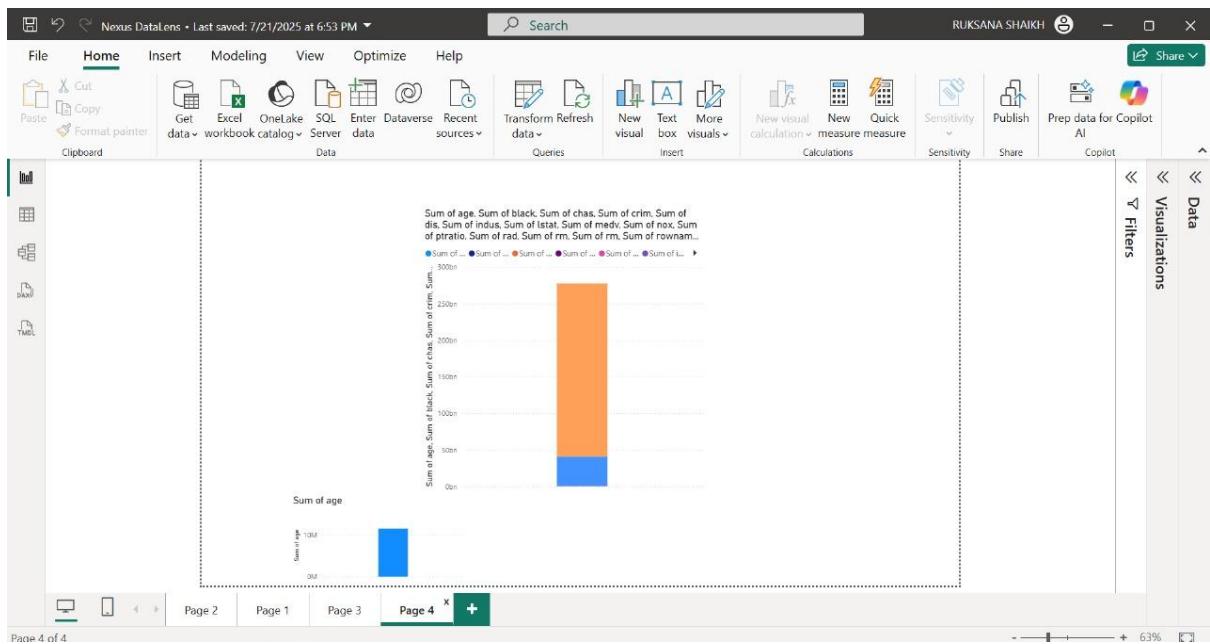
9.2 Power BI Integration (Optional)



Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



10. Challenges and Resolutions

Throughout the development of *Nexus DataLens*, several technical and architectural challenges were encountered. Each of these obstacles presented an opportunity to strengthen the system's robustness and reliability. Below is a summary of the key challenges faced and the corresponding solutions implemented.

1. Incorrect S3 Path in Glue Crawler Configuration

Challenge:

Initially, the AWS Glue Crawler was configured to scan an outdated S3 bucket (nexus-web-front), which resulted in the creation of metadata tables pointing to obsolete data locations. This caused downstream services like Athena and QuickSight to display empty or outdated datasets.

Resolution:

The crawler configuration was updated to target the correct S3 path (nexus-web-frontend1/data/). Existing Glue tables linked to the old location were manually deleted from the Data Catalog, and a fresh crawl was initiated to reflect accurate schema metadata in Athena.

2. Frontend Upload Failure Due to S3 Bucket Policy

Challenge:

CSV file uploads from the frontend to S3 were initially unsuccessful due to restrictive CORS policies and insufficient S3 permissions.

Resolution:

The S3 bucket policy was carefully updated to allow secure PUT access from the frontend domain. Cross-Origin Resource Sharing (CORS) rules were configured to allow the appropriate HTTP methods. This enabled seamless file transfers while maintaining security boundaries.

3. Lambda Execution Timeout During Large File Processing

Challenge:

When handling larger datasets, the Lambda function occasionally exceeded its default timeout, disrupting the automated trigger for Glue Crawlers and Jobs.

Resolution:

The function timeout setting was increased within AWS Lambda to accommodate longer processing durations. Logging and exception handling were also enhanced to improve reliability and traceability for future executions.

4. Duplicate Schema Entries in Glue Catalog

Challenge:

Multiple crawls on the same S3 location led to the creation of duplicate tables in the Glue Catalog, causing confusion in downstream Athena queries and QuickSight reports.

Resolution:

To resolve this, a naming convention and scheduled crawler execution strategy were implemented. Old tables were cleaned up, and table creation settings were restricted to avoid duplication unless schema changes were detected.

5. QuickSight Dataset Not Refreshing Automatically

Challenge:

After successful data cleaning and schema updates, QuickSight dashboards were not reflecting the latest data due to the absence of automatic dataset refresh.

Resolution:

A manual refresh was triggered initially. Later, this was resolved by integrating QuickSight dataset update logic via Lambda post Glue Job execution. This ensured the dashboard always presents the most recent data.

11. Key Features & Benefits

Nexus DataLens offers a robust set of features designed to simplify the data analytics lifecycle while ensuring scalability, automation, and enterprise-grade security.

Key Features:

- **End-to-End Automation** – From CSV upload to dashboard generation, the entire pipeline runs automatically without manual intervention.
- **Serverless Architecture** – Built entirely on AWS managed services (S3, Lambda, Glue, Athena, QuickSight), ensuring scalability and cost-efficiency.
- **User-Friendly Frontend** – A simple web interface for non-technical users to upload CSV files securely.
- **Flexible Data Cleaning** – Supports both code-based cleaning via AWS Glue Jobs and no-code transformations using AWS Glue DataBrew.
- **Seamless Visualization** – Provides dynamic dashboards in Amazon QuickSight with optional integration into Microsoft Power BI.
- **Enterprise-Grade Security** – Implements AWS IAM for access control, KMS for encryption, and CloudWatch for monitoring.

Benefits:

- **Reduced Effort & Time** – Eliminates manual data preparation, accelerating insights.
- **Scalability on Demand** – Automatically adapts to different workloads without infrastructure management.
- **Improved Data Accuracy** – Automated cleaning minimizes human error and ensures consistent datasets.
- **Accessibility for All Users** – Enables non-technical users to generate insights effortlessly.
- **Cost-Effective Solution** – Pay-as-you-go pricing model of AWS serverless services reduces operational costs.

12. Use Cases of Nexus DataLens

The *Nexus DataLens* platform is designed as a scalable, serverless data analytics solution that caters to a variety of real-world business and analytical needs. Its architecture allows organizations to ingest, clean, query, and visualize data with minimal manual effort. Below are key use cases where Nexus DataLens can deliver significant value:

1. Business Intelligence & Reporting Automation

Problem: Businesses often deal with scattered raw data that requires time-consuming preprocessing before analysis.

Solution: Nexus DataLens enables automated transformation of raw CSV files into clean, queryable datasets, and visualizes them through Amazon QuickSight or Power BI dashboards.

Impact: Saves analyst time, enables faster decision-making, and improves data accuracy.

2. Operational Monitoring & KPI Dashboards

Problem: Monitoring KPIs (Key Performance Indicators) across departments can be difficult without centralized visualization.

Solution: Uploads from various teams can be cleaned and visualized automatically into live dashboards using QuickSight or Power BI.

Impact: Enables department heads to monitor sales, operations, customer satisfaction, or finance in real time.

3. Data Democratization for Non-Technical Users

Problem: Non-technical users struggle to query data or generate insights.

Solution: With a simple frontend upload UI, anyone can upload CSV files and get visualized insights without needing to write code.

Impact: Makes data-driven decision-making accessible to marketing, HR, or sales teams.

4. Academic or Research Data Analysis

Problem: Researchers need a quick way to analyze large CSV-based datasets without complex setups.

Solution: Researchers can upload their raw data, clean it, and immediately query or visualize results using built-in analytics tools.

Impact: Accelerates data-driven research and facilitates reproducibility.

5. Client/Partner Data Integration for Enterprises

Problem: Receiving external data from clients or partners often requires preprocessing before use.

Solution: External partners can upload CSVs via the secure frontend; Nexus DataLens automatically prepares it for analysis.

Impact: Speeds up B2B data integration and fosters seamless collaboration.

6. Government or NGO Reporting Dashboards

Problem: Public agencies and NGOs handle data from multiple sources and require transparent reporting.

Solution: Nexus DataLens provides automated processing and visualization, ideal for transparency

and compliance.

Impact: Facilitates evidence-based reporting to stakeholders and donors.

7. Startup MVP for Analytics-as-a-Service

Problem: Building a custom analytics backend from scratch is time- and resource-intensive.

Solution: Startups can white-label Nexus DataLens as a backend for providing analytics services to their users.

Impact: Reduces go-to-market time and cost for SaaS products.

8. Internal Tool for ETL Process Validation

Problem: Data engineers often need to test ETL transformations before deployment.

Solution: Engineers can use Nexus DataLens to validate ETL logic on sample CSVs and visualize the output.

Impact: Ensures data quality and reduces post-deployment errors.

13. Results and Outcomes

The implementation of **Nexus DataLens** successfully demonstrated the value of an automated, serverless data analytics pipeline on AWS. The project delivered measurable improvements across multiple dimensions:

- **Seamless Automation** – Achieved a fully automated workflow where CSV files uploaded via the frontend are processed, cleaned, and visualized without manual intervention.
- **Data Accuracy** – Cleaning logic in AWS Glue Jobs ensured removal of null values, inconsistent headers, and invalid entries, producing high-quality datasets ready for analysis.
- **Faster Time to Insights** – Reduced the end-to-end process (from file upload to dashboard generation) from hours of manual effort to just a few minutes.
- **Scalability** – Validated that the serverless architecture can handle growing data volumes without additional infrastructure management.
- **Visualization Success** – Generated interactive dashboards in **Amazon QuickSight** and validated **Power BI integration**, confirming flexibility for different BI environments.
- **Enhanced Accessibility** – Enabled non-technical users to generate insights through a simple upload interface, removing dependency on specialized data engineers.

Overall Outcome: Nexus DataLens proved to be a secure, cost-effective, and extensible solution that can be applied across business, research, and organizational use cases to accelerate data-driven decision-making.

14. Conclusion and Future Enhancements

14.1 Recommendations

To further enhance the robustness, usability, and scalability of the Nexus DataLens platform, several strategic improvements are proposed. These recommendations aim to optimize performance, reduce manual overhead, and extend functionality to support enterprise-scale analytics needs.

1. File Validation Prior to Processing

Implementing pre-ingestion validation checks for uploaded files—such as verifying size limits, acceptable formats (e.g., CSV only), and schema conformity—would help prevent processing failures and ensure data consistency.

2. Automated Cleanup of Obsolete Data

Introduce lifecycle management for raw and processed datasets stored in Amazon S3. Automating the deletion of outdated files, logs, and dashboards would not only reduce storage costs but also maintain a clutter-free data environment.

3. Event-Based Notifications

Incorporating Amazon SNS or SES to notify users upon pipeline completion, failure, or anomalies would provide real-time operational transparency. This is particularly useful in multi-user or team environments where timely updates are crucial.

4. Integration with Redshift or RDS

Extending data storage and querying capabilities by integrating with Amazon Redshift or Amazon RDS would enable persistent and structured storage for large-scale data warehousing and advanced analytics.

14.2 Future Scope

Looking ahead, the Nexus DataLens platform can evolve significantly to support broader use cases and provide deeper insights:

- **Support for Additional File Formats**

Extend compatibility beyond CSV to include formats such as Excel (XLSX) and JSON, thereby enabling a wider range of data sources to be processed seamlessly.

- **Machine Learning-Based Anomaly Detection**

Integrate Amazon SageMaker or similar ML services to automatically detect anomalies and outliers during the data cleaning process, enabling smarter and more proactive data quality assurance.

- **Role-Based Dashboards and Access Controls**

Enhance user experience by enabling personalized dashboards based on roles, departments, or teams. AWS Cognito and IAM policies can be used to implement granular access controls.

- **Comprehensive Notification System**

Expand the notification mechanism to support both email (SES) and SMS-based alerts (SNS), offering flexible communication options tailored to user preferences.

15. References

1. **Amazon S3 – Object Storage Service**
<https://docs.aws.amazon.com/s3>
2. **AWS Lambda – Serverless Compute**
<https://docs.aws.amazon.com/lambda>
3. **AWS Glue – Data Integration & ETL**
<https://docs.aws.amazon.com/glue>
4. **Amazon Athena – Query Service**
<https://docs.aws.amazon.com/athena>
5. **Amazon QuickSight – Business Intelligence**
<https://docs.aws.amazon.com/quicksight>
6. **Amazon API Gateway – Managed API Service**
<https://docs.aws.amazon.com/apigateway>

17. Appendix

17.1 Source Code Links (GitHub)

The complete source code, including frontend files, Lambda functions, and Glue job scripts, is available in the project's GitHub repository:

👉 [GitHub Repository – Nexus DataLens 2.0](#)

17.2 Sample Input/Output

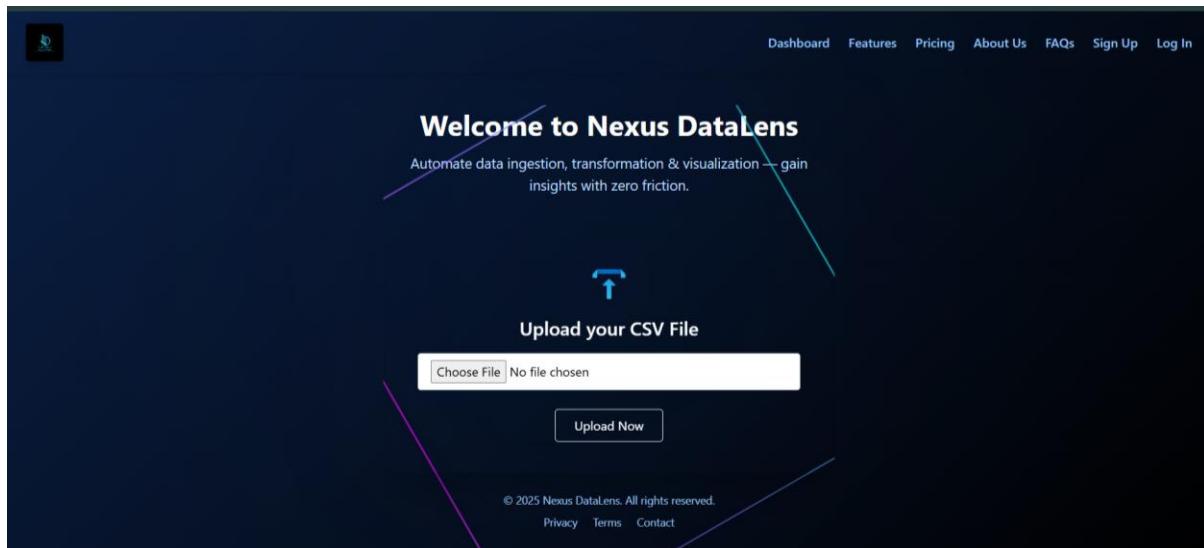
- **Sample Input (Raw CSV):** A noisy employee dataset with missing values, inconsistent headers, and placeholder entries.
- **Sample Output (Cleaned CSV):** The processed dataset generated by the Glue Job, containing standardized headers, filtered values, and a single flat CSV file stored in s3://nexus-web-frontend1/cleaned-output/.

17.3 Additional Screenshots

Below are supplementary screenshots to illustrate the end-to-end pipeline:

- Frontend upload interface
- S3 bucket structure (raw data vs. cleaned output)
- AWS Lambda configuration and CloudWatch logs
- Glue Crawler and Glue Job setup
- Athena query editor with sample SQL results
- Amazon QuickSight dashboards (bar, line, and pie charts)
- Power BI report (optional integration demo)

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



This screenshot shows the Amazon S3 console. The left sidebar includes sections for General purpose buckets (with options like Directory buckets, Table buckets, Vector buckets, Access Grants, etc.) and Storage Lens (Dashboards, Storage Lens groups). The main area displays a list of objects in a bucket named 'nexus-web-frontend1'. The list includes several CSV files: 'people_100.csv', 'Projects.csv', 's_8 (1).csv', 's_8.csv', 'test data.csv', 'test_1_s_2.csv', 'test_1_s_3.csv', 'test_1_s_4.csv', and 'test_1_s_5.csv'. Each entry shows details like name, type, last modified date, size, and storage class (Standard). A toolbar at the top provides actions like Copy S3 URI, Copy URL, Download, Open, Delete, Actions, Create folder, and Upload. The bottom of the page includes standard AWS footer links.

This screenshot shows the same Amazon S3 bucket listing as the previous one, but with specific items selected. The 'people_100.csv' file is highlighted with a blue border. The rest of the interface is identical to the first screenshot, including the sidebar, object list, and toolbar.

Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

Crawler properties

- Name: cleaned-output-crawler
- IAM role: GlueDataBrewRole_1
- Description: -
- Database: nexusdb_3
- State: READY

Crawler runs (5)

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
July 21, 2025 at 12:40:35	July 21, 2025 at 12:41:55	01 min 20 s	Completed	0.066	45 table changes, 5 partition changes
July 18, 2025 at 11:25:04	July 18, 2025 at 11:25:57	52 s	Completed	0.071	3 table changes, 4 partition changes
July 18, 2025 at 10:41:31	July 18, 2025 at 10:42:24	52 s	Completed	0.059	-
July 18, 2025 at 10:29:33	July 18, 2025 at 10:30:30	56 s	Completed	0.083	2 table changes, 0 partition changes
July 12, 2025 at 15:14:02	July 12, 2025 at 15:14:54	52 s	Completed	0.079	1 table change, 0 partition changes

Editor | Recent queries | Saved queries | Settings

Data

- Data source: AwsDataCatalog
- Catalog: None
- Database: nexusdb

Tables and views

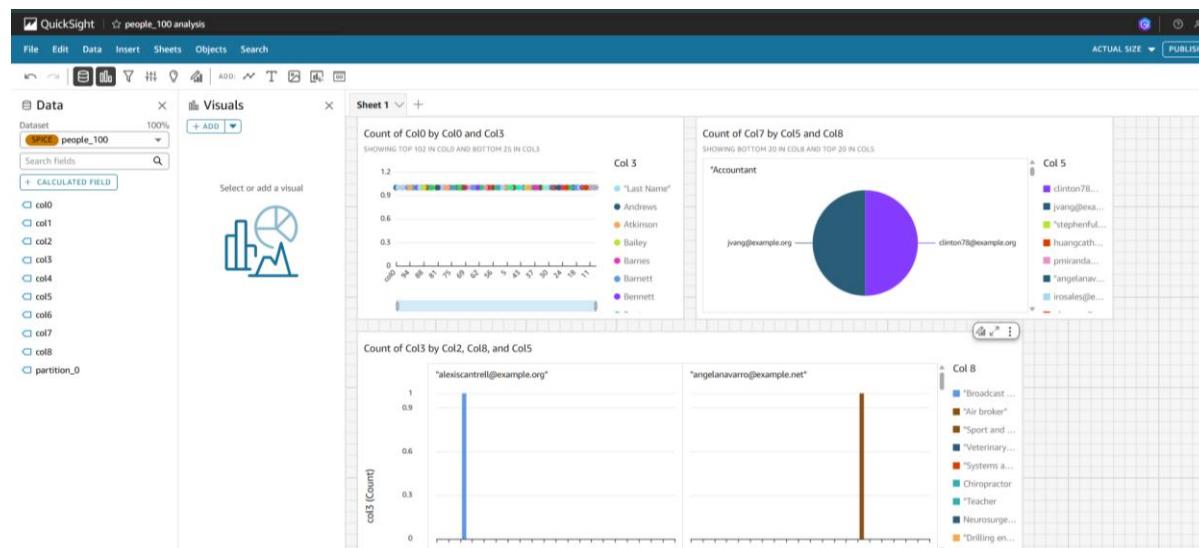
- Tables (5): people_100, people_100.csv, people_100.csv_796046d9754dff7d59e, 4bf849511a231

Query 1 :

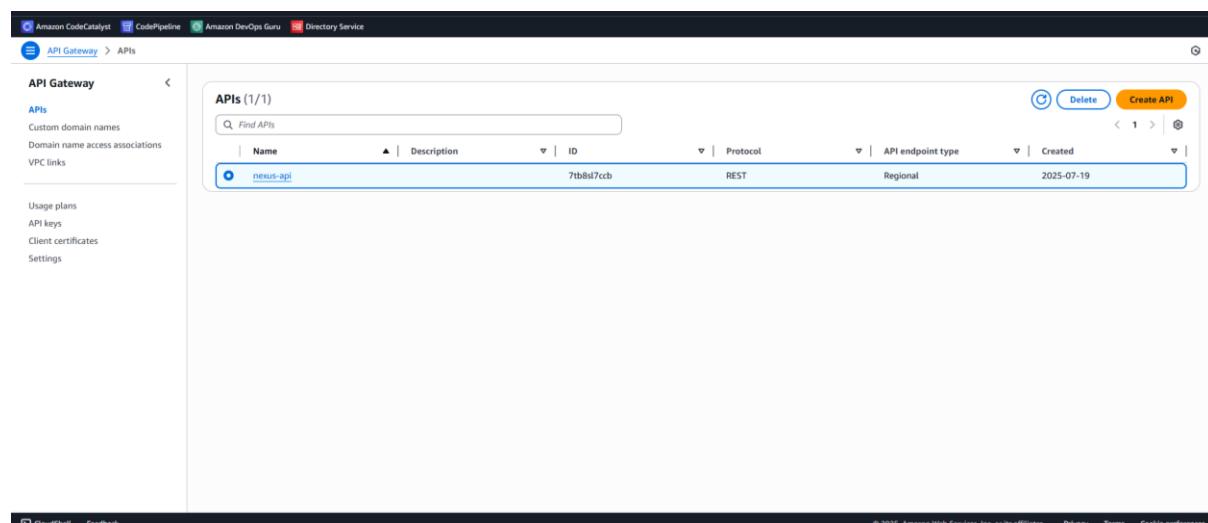
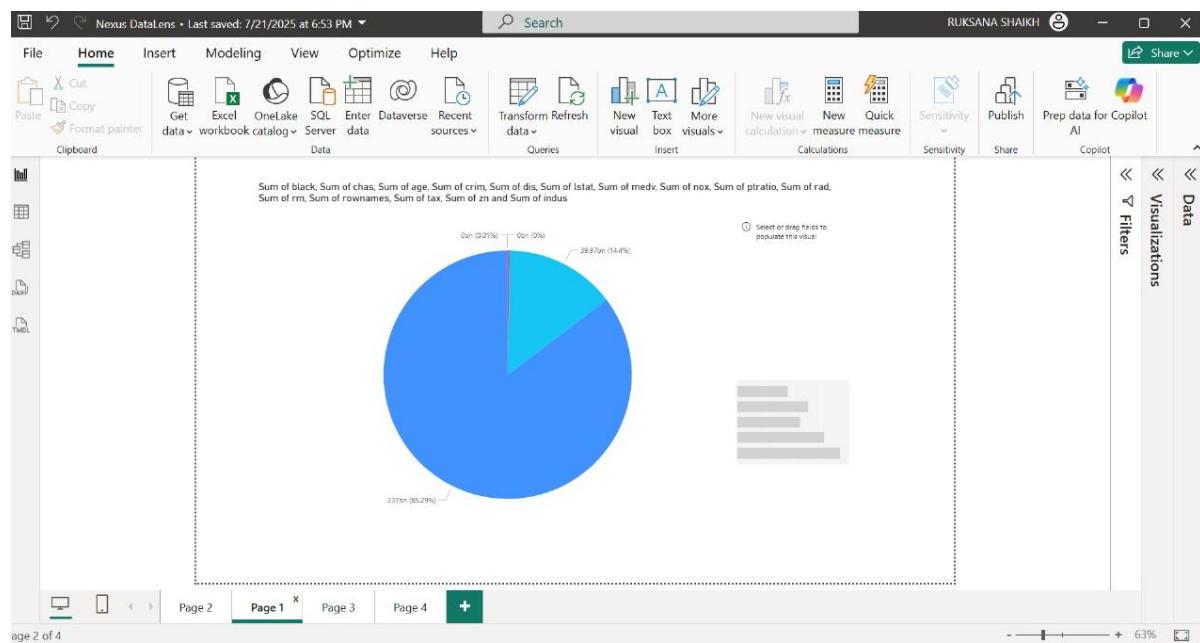
```
1 SELECT * FROM nexusdb.people_100 LIMIT 10;
```

SQL | Run again | Explain | Cancel | Clear | Create | Reuse query results up to 60 minutes ago

Query results | Query stats



Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS



Nexus DataLens 2.0 – Automated Data Analytics Pipeline on AWS

The screenshot shows the AWS API Gateway console with two API resources defined:

- /QuicksightDashboardUI**: A resource with a single method, **ANY**, which integrates with a Lambda function. Authorization is set to "None" and API key is "Not required".
- /user**: A resource with five methods: **DELETE**, **GET**, **OPTIONS**, **POST**, and **PUT**. All methods integrate with a Lambda function, and both authorization and API key requirements are set to "Not required".

The left sidebar shows the navigation path: API Gateway > APIs > Resources - nexus-api (7b8sl7ccb). The right sidebar includes standard AWS links like CloudShell, Feedback, and footer links for Privacy, Terms, and Cookie preferences.