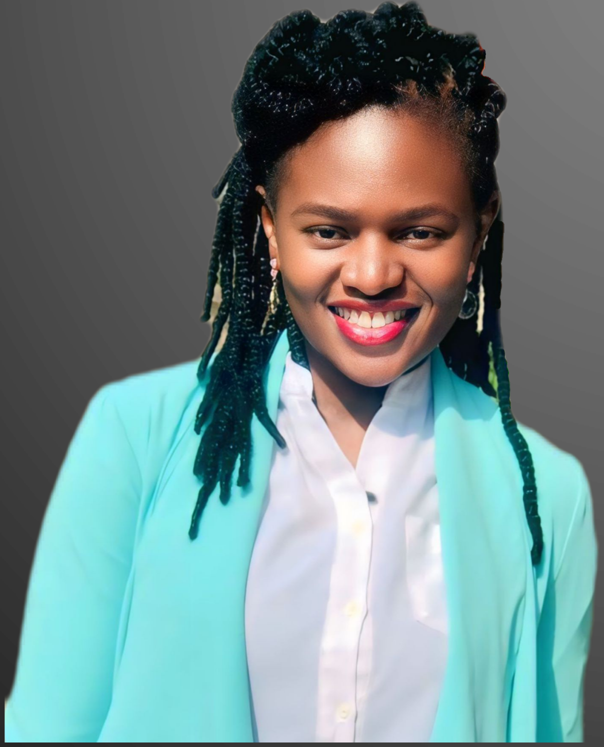
TENGU QUINTER



AWS SOLUTION ARCHITECT INSTRUCTOR

BUSINESS ENTHUSIAST 💼

BAKER 🧑‍🍳

INTRODUCTION

This project is a personal mission driven by a real-world problem in my community in Cameroon. As a mom, I'm highly aware of how easy it is to miss critical vaccination dates. The reality is that not every parent has the literacy or resources to keep meticulous medical records. This project was born from that need: to build an automated, reliable system that proactively reminds parents and healthcare providers about vaccination schedules, helping to prevent health crises like measles and hepatitis B.

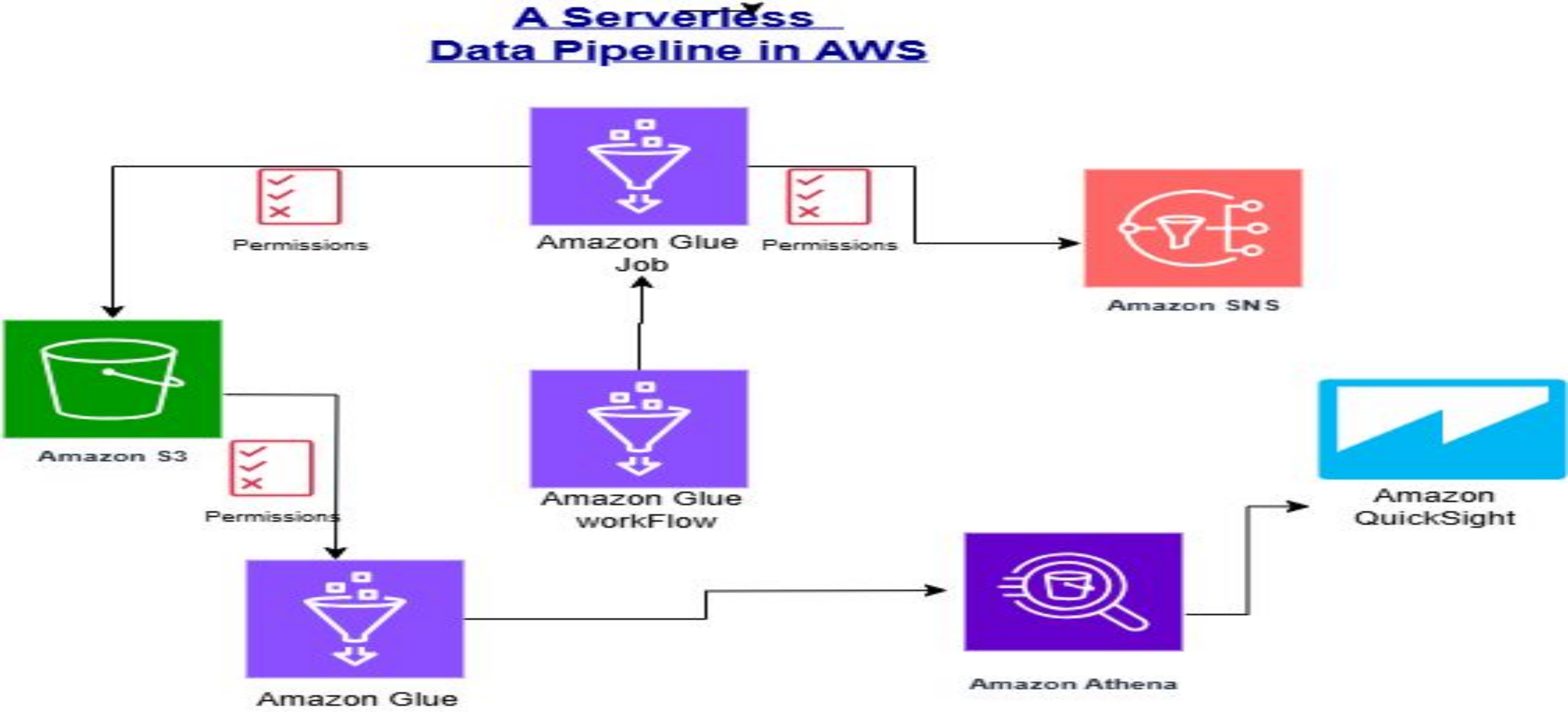
THE CHALLENGE

In many communities, a lack of consistent communication and record-keeping leads to children missing life-saving vaccinations. This project addresses the challenge of ensuring timely immunizations by creating a scalable, automated solution that doesn't rely on manual follow-up or expensive, high-maintenance infrastructure.

The Solution: A Serverless ETL Data Pipeline on AWS

This project is an end-to-end, serverless ETL (Extract, Transform, Load) data pipeline built entirely on the AWS cloud. It automates the process of identifying babies due for vaccination and notifies healthcare workers in real-time.

AUTOMATED VACCINATION REMINDER
A SERVERLESS ETL DATA PIPELINE ON AWS



AWS SERVICES: A DEEP DIVE

Amazon S3 (Simple Storage Service)

- Amazon S3 is a highly scalable and durable object storage service.
- **Role in Project**: Serves as the central data lake. We use it to securely store raw vaccination data in a scalable and highly available manner.
- **Key Function**: Provides object storage for data of any type. It's the foundation of our serverless architecture, giving us a single, reliable source of truth.

AWS SERVICES: A DEEP DIVE

AWS IAM (Identity and Access Management

A service that enables you to manage access to AWS services and resources securely. It allows you to create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources.

Role in Project: The backbone of our project's security. We used it to create a specific role for the Glue job, granting it only the necessary permissions to read from S3 and publish messages via SNS. This ensures our pipeline is both functional and secure.

Key Function: A global service that allows you to manage user permissions and access to your AWS resources.

AWS SERVICES: A DEEP DIVE

Amazon CloudWatch

A monitoring and observability service that provides you with data and actionable insights to monitor your applications, respond to system-wide performance changes, and optimize resource utilization.

Role in Project: Provides crucial monitoring and logging needed to troubleshoot and ensure our job runs successfully every time. We relied on its detailed logs to track the progress of our pipeline and debug any issues.

AWS SERVICES: A DEEP DIVE

AWS Glue

- AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data.
- **Role in Project:** Our ETL (Extract, Transform, Load) engine. We use a serverless Glue Job to read the raw data from S3, process it, calculate future dates, and identify today's reminders.
- **Key Function:** An automated, serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

AWS SERVICES: A DEEP DIVE

AWS Glue Workflow

- AWS Glue Workflow is a powerful orchestration service for creating and managing data pipelines.
- **Role in Project:** Acts as our scheduler and orchestrator. It manages the entire data pipeline, ensuring that the AWS Glue Job runs on a defined schedule and handles all dependencies in the workflow. This replaces the need for a separate EventBridge rule.
- **Key Function:** A fully managed orchestration service for data pipelines. It allows you to create and manage complex ETL workflows with dependencies, schedules, and triggers.

AWS SERVICES: A DEEP DIVE

Amazon SNS (Simple Notification Service)

- Amazon SNS is a fully managed messaging service for both application-to-application and application-to-person communication.
- **Role in Project:** The notification hub. Once the Glue Job identifies the reminders, it publishes a message to an SNS Topic. SNS then delivers this message to a subscriber, such as a Public Relations Officer via email.
- **Key Function:** A fully managed messaging service for both application-to-application (A2A) and application-to-person (A2P) communication.

AWS SERVICES: A DEEP DIVE

Amazon Athena

- Amazon Athena is a serverless query service that allows you to analyze data directly in Amazon S3 using standard SQL.
- **Role in Project:** Our interactive query service. We use Athena to run standard SQL queries directly against the data in S3. It allows us to perform powerful, on-demand analysis without having to load the data into a data warehouse.
- **Key Function:** A serverless analytics service that makes it easy to analyze data in Amazon S3 using standard SQL. It is a powerful tool for ad-hoc analysis and business intelligence.

AWS SERVICES: A DEEP DIVE

Amazon QuickSight

- Amazon QuickSight is a scalable, serverless, and embeddable machine learning-powered business intelligence (BI) service.
- **Role in Project:** The visualization and reporting tool. We can use QuickSight to create rich, interactive dashboards from the data queried by Athena. This allows stakeholders to easily visualize vaccination trends, forecast future needs, and monitor campaign effectiveness without writing any code.
- **Key Function:** A fully managed business intelligence service that allows you to create and publish interactive dashboards, providing a simple way for non-technical users to explore data.

How It Works: the Pipeline Flow

1. **Data Ingestion:** A healthcare worker uploads a CSV file containing patient data to an S3 bucket.
2. **Schema Discovery:** The **Glue Crawler** automatically detects the schema and registers a table in the Glue Data Catalog.
3. **Automated Processing:** A **Glue Workflow** triggers a **Glue Job** on a daily schedule.
4. **ETL:** The Glue Job processes the data, identifying babies with near due vaccinations.

How It Works: the Pipeline Flow

5. **Proactive Notifications:** The Glue Job publishes two separate messages to an **SNS topic**, one for babies ready for vaccination and another if there are no babies due for vaccination.

6. **Email Reminders:** SNS delivers these messages to the subscribed Public Relations Officer (PRO) via email.

7. **Data Analysis:** The PRO or other stakeholders can use **Athena** to run SQL queries or create a dashboard in **QuickSight** to analyze vaccination trends and forecasts.

SETUP AND INSTRUCTIONS

STEP 1: DATA AND STORAGE 📁

Preparing Our Dataset

Create the CSV File:

- Copy and save Google Sheet file in .csv format.
- While saving choose CSV(comma delimited)
- Save as `vaccination_records.csv`.

STEP 1: DATA AND STORAGE 📁

Preparing Our Dataset

- **Create an S3 Bucket:**
  - Go to **Amazon S3** and create a new bucket.
  - Name it descriptively (e.g., vaccination-data-2025).
- **Upload the File:**
  - Upload the vaccination records gotten from the hospital to the S3 bucket.
  - Name it descriptively (e.g., vaccination-data-2025).

STEP 2: NOTIFICATIONS 🔔

Setting up a Channel for Reminders

- **Create an SNS Topic:**
  - Go to **Amazon SNS** and click **Create topic**.
  - **Name:** pro-vaccination-reminders.
  - **Type:** Standard.

- **Add an Email Subscription:**
  - Within the topic, click **Create subscription**.
  - **Protocol:** Email.
  - **Endpoint:** pro@example.com (add PRO's email).

- **Confirm Subscription:** The PRO must check their email
- and click the confirmation link.

STEP 3: THE GLUE JOB 💻

The Engine of Our Pipeline

- **Create IAM Role:**
  - Attach policies for S3, SNS, and CloudWatch access.
  - Attach the AWSGlueServiceRole managed policy.
- **Modify the Glue script:**
  - **C**opy your **S3 Bucket Name,Key(csv file name)** and **SNS Topic ARN** and paste on the Glue script
- **Create the Glue Job:**
  - Go to **AWS Glue** > **Jobs**.
  - Select **Script editor.**
  - **Type: Spark** (critical for pyspark).
  - Uplaod script from your pc and create script.

STEP 3: THE GLUE JOB 💻

- Go to **Job details** and entire the details.
- **Name:** vaccination-reminder-job.
- **IAM Role:** Select the role you created.
- **Script Path:** s3://your-bucket/glue_script.py.
- **Worker Type:** G.1X.
- **Workers:** 2. we don’t need the default 10 workers for this job.
- Leave other default configurations and save.
- You can test the job by running it manually.

STEP 4: THE TRIGGER ⏰

The Daily Scheduler

- **Create Glue Workflow:**
  - Go to **AWS Glue** > **Workflows**.
  - Click **Add workflow**.
  - Name: vaccination-reminder-workflow.
- **Create the Trigger:**
  - Add a trigger to your workflow.
  - **Schedule Type:** Scheduled.
  - **Frequency:** choose **Custom** and add rate(5 minutes) OR Cron Expression: 0/5 * * * ? *.
  - **Add Node:** Connect your vaccination-reminder-job to this trigger.
- **Start the Workflow:** The workflow will now automatically execute the job every 5 minutes.

STEP 5: DATA ANALYSIS WITH ATHENA 📊

Insights from Our Data

- **Create a Glue Crawler:**
  - Go to **AWS Glue** > **Crawlers**.
  - Point it to your S3 bucket.
  - Create an IAM Role add GlueServiceRole as well as S3BucketFullAccess
  - Create the database vaccination-Reminder-db.
  - Run the crawler to create the table vaccination_records.
- **Query Your Data:**
  - Open the Table your crawler created,
  - Click on Action and click view data
  - Click Proceed
  - Run queries to get insights, like finding all upcoming reminders.
  - SELECT * FROM "vaccine-db2"."vaccination_records" WHERE CAST
  - (date_of_birth AS DATE) + INTERVAL ‘30' DAY > CURRENT_DATE;

STEP 5: DATA ANALYTICS WITH QUICKSSIGHT 📊

Create an analytical dashboard on Quicksight

- Go to the **QuickSight** console and click **Manage data**.
- Click **New data set**.
- Select **Athena** as your data source.
- Enter a name for your data source (e.g., VaccinationData).
- Select the vaccination-Reminder-db database.
- Choose the cameroon_vaccination_data table.
- Click **Publish and Visualize** to create a new analysis and build a dashboard.

TROUBLESHOOTING 💡

Common Issues and Solutions

- **ModuleNotFoundError: No module named 'pyspark'**
  - **Fix:** Ensure your Glue job **Type** is **Spark**, not Python Shell.
- **TABLE_NOT_FOUND**
  - **Fix 1:** Verify your Glue crawler is pointing to the correct S3 path (e.g., s3://bucket/, not s3://bucket/file.csv).
  - **Fix 2:** In Athena, ensure the database dropdown is set to the correct
    database (e.g., vaccine-db2).
  - **Fix 3:** Check for typos in the database or table name.

Delete all resources you created to save cost.

QUESTIONS & DISCUSSION

Thank you!