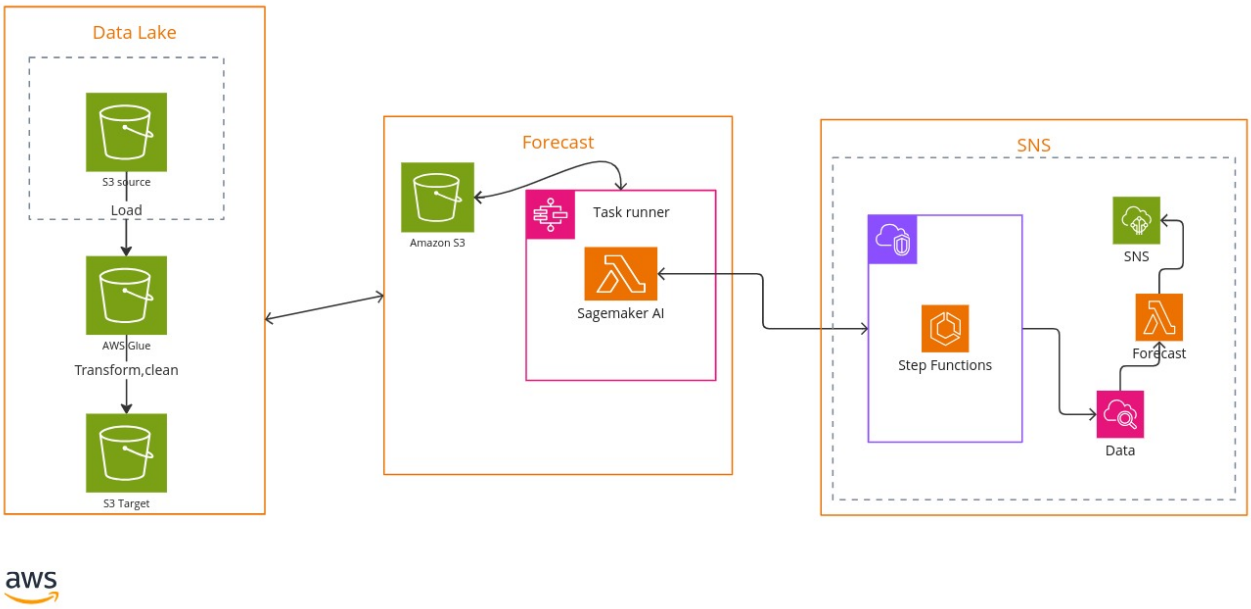
Marriage-Forecast Pipeline – Report

Author: Assadbek Doskaliyev (220103079)

Executive Summary

The Marriage-Forecast pipeline is a solution that ingests civil-registry data, cleans it, produces a one-year forecast for the number of marriages The preject demonstrates Glue-based ETL, Machine Learning Model, Step Functions orchestration, and SNS alerting..

1 System Architecture



Component Map

#	Service	Purpose	Key Configurations
1	**Amazon S3**	Data Lake	Load dataset to S3 bucket
2	**AWS Glue Job**	Spark ETL cleans raw CSV → Parquet	Transform and cleaning
3	**AWS Machine Learning**	Python forecast (LTSM)	Forecasting code in SageMaker AI notebook
4	**AWS Step Functions**	Orchestrates ETL → Forecast → Notify	Standard Workflow

#	Service	Purpose	Key Configurations
5	**Amazon SNS**	Success/failure email alerts	Email subscriptions

2 Data Flow

1. S3 Bucket → uploading the dataset, create folder for the cleaned data
2. **AWS (Glue) + ETL**
    - Create databases and table for the crawler
    - Cleans nulls, remove fields. transform.
3. **Forecast (Sagemaker AI)**
    - Downloads the clean dataset
    - Trains LTSM model
    - Predicts marriages for next year
4. **Notify (SNS)** – subscription to email

3 Key Learnings & Skills Gained

Area	What I Learned	In Project
AWS Glue	Authoring PySpark jobs, partitioning data, understanding how effictively clean data	Visual ETL, ETL notebook
S3	Knowing the best way the store data	Bucket
SageMaker AI	Work with forecasting model(LTSM)	Notebook(Jupyterlab)
Step Functions	Designing Standard workflows	State machine, code json
SNS	Subscribe the email for future reports/alerts	Notify, Subscription
Cost Optimization	Calculating DPH	Cost table §10