John L Mackin

Email: JohnLutherMackin@jlmw.xyz
Phone:

Professional Summary

Data Engineer and developer specializing in developing and deploying solutions in Amazon Web Services. Combines an in-depth understanding of data modeling with knowledge of the underlying architecture for various data engineering technologies to determine appropriate solutions for real-world problems. Well versed in optimized data solutions, and familiar with the evolution of effective technologies. Experience with optimizing data storage, processing, and analysis, implementing automation solutions, and as a devotee of the computer arts.

- Orchestrated and scheduled jobs using shell scripts, CRON, and Apache Airflow
- Utilized popular libraries including Pandas and NumPy in Python to build data pipelines.
- Designed and deployed a normalized relational database schema on Postgres.
- Selected appropriate list, range, and hash partitioning schemas based on a combination of the underlying database technology and business requirements.
- Deployed a cluster on MongoDB Atlas and developed functions for CRUD operations.
- Constructed SQL queries utilizing joins, sub-queries, and windowing functions to meet complex reporting requirements.
- Utilized knowledge of core Hadoop architecture for Hadoop, HDFS, YARN, and MapReduce to understand the evolution of the modern data engineering ecosystem.
- Certified practitioner of core AWS services and use cases, billing and pricing models, security concepts, and how cloud impacts a business.
- Created ETL (Extract / Transform / Load) processes using Spark on Elastic Map Reduce.
- Ingested a variety of batch and streaming datasets in AWS using Lambda and Kinesis.
- Deployed a variety of services on AWS to create a Data Lake solution using the Delta format employing raw, staging, and curated layers.
- Connected Databricks to an AWS account to deploy Apache Spark clusters using AWS resources and data.
- Experienced troubleshooting Spark application performance by analyzing the breakdown of jobs, stages, and tasks to identify bottlenecks in the application.
- Configured Spark configurations for the driver, executors, and cluster management with YARN to optimize performance of Spark Applications.

Technical Skills

- Python, SQL
- Pandas, Dask, Delta Lake, Parquet, etc.
- Relational Databases, MySQL, PostgreSQL
- NoSQL Databases, MongoDB
- Databricks
- Apache Spark, Hadoop
- Data Modeling
- Data Warehouse, Data Lake, and Data Lakehouse Architecture
- AWS Services, i.e. EC2, S3, Lambda, Glue, Redshift, QuickSight, CloudWatch and CloudTrail
- Java, Spring, and SpringBoot
- Google Python APIs, AppsScript, and Javascript

Professional Experience

SkillStorm **Jun 2022 – Feb 2023**

Data Engineer

- *Environmental Protection Agency*

The EPA received grant funding to develop a robust data engineering solution to process and analyze global air quality measurements. Measurements are taken by stations across the globe to record the levels of several pollutants at various intervals. Historical and real-time data must be unified so that it can be analyzed by several other teams to better understand and predict global air quality.

- Developed AWS Lambda functions to load historical data into the raw layer of an S3 data lake.
- Created Kinesis Data and Delivery Streams to ingest and format real-time data into the S3.
- Configured an AWS Glue Data Catalog on the S3 data lake to provide a centralized source of metadata for data governance and discovery.
- Analyzed business and reporting requirements to create an optimized data model.
- Integrated with Databricks and deployed a Spark cluster to transform and aggregate cleaned data into a snowflake schema for use in a data warehousing solution.
- Designed, implemented, and tested the data warehousing solution on Amazon Redshift.
- Orchestrated jobs across disparate systems using Apache Airflow.
- Utilized Amazon Quick Sight to develop a dashboard adhering to BI requirements.

- *Sparrow Insights*

Sparrow Insights was migrating their data infrastructure from on-premises to the AWS cloud. They specialize in analyzing version control data for software to provide insights into the habits of developers. They tasked the data engineering team to develop an ETL pipeline on Apache Spark on Databricks.

- Created and managed a data lake on AWS S3 using raw, conformance, and curated layers.
- Analyzed business and data requirements to appropriately size and configure an Apache Spark cluster on the Databricks platform.
- Developed multiple PySpark scripts for each stage of the ETL pipeline.
- Utilized the Databricks Jobs API to manage orchestration of Spark applications.
- Deconstructed Spark applications into jobs, stages, and tasks to identify bottlenecks and optimize performance.
- Separated data into fact and dimension tables using a star schema to optimize OLAP workloads.
- Aggregated data using business reporting requirements for use in visualization and BI tools.

- *Bastion Analytics*

A data analytics firm was tasked with providing an Extract, Transform, and Load (ETL) pipeline for the U.S. Customs and Border Protection. The shipping data was generated from an Automated Manifest System (AMS) for shipments from 2018-2020.

- Utilized the Python library Pandas to perform Exploratory Data Analysis of the dataset to determine an appropriate data model and data cleaning strategy.
- Determined an appropriate Data Lake structure with bronze, silver, and gold layers.
- Submitted a Proof of Concept (PoC) for the architecture of their proposed solution for the data lake, data warehouse, and ETL pipeline.
- Developed SQL and Python scripts to create partitioned and indexed tables in Postgres and automate loading the transformed data into the tables.
- Estimated the cost of the proposed solution in AWS using S3, Lambda, and Redshift.

Culture Management, Mckinleyville, CA **July 2020 - July 2022**

Data Specialist

- Managed and optimized worker and product data, using JavaScript and Google Apps Script
- Implemented improved automation solutions using Google Drive API, Apps Script, and Python
- Implemented systems for data analysis and storage using Google Drive API, Apps Script, and Python.

Certifications

- AWS Certified Cloud Practitioner
- Databricks Certified Associate Developer for Apache Spark 3.0 – Python
- CompTIA Data+

Education

Humboldt State University, Arcata, CA **January 2020**
Bachelor of Science in Computer Science

Relevant Projects

TTF Data analytics

- Data analytics tool for pulling data from documents in Google drive using Python and Drive API

JLM Site

- Personal website, developed using a PostgreSQL database, Java SpringBoot.
- Blog concept with user info storage, and blog entry storage.