Wesley Lau

(571) 354-9369 • wesleylau.wcl@gmail.com • LinkedIn

PROFESSIONAL SUMMARY

Data Engineer with 6+ years of experience building scalable ETL pipelines and data warehouses using AWS, Python (PySpark), and SQL. Proven expertise in process automation, cost reduction, and leveraging big data technologies like Neo4j to support advanced analytics.

PROFESSIONAL EXPERIENCE

Data Engineer | National Cancer Institute **NOVEMBER** 2022 – January 2025

- Led the integration of two external systems into the existing data architecture, expand the existing Neo4j Graph data model, improving data accessibility and scalability.
- Developed and executed a data migration strategy to transfer 3TB of data into Neo4j, utilizing Python-based ETL scripts with *Pandas* and *NumPy* for data transformation and cleaning.
- Implemented CI/CD pipelines using GitHub Actions to automate the testing and deployment of Python data loaders, enabling continuous integration and delivery.
- Built event-driven ETL automation with *AWS Lambda* function with Python that triggered the ETL process based on S3 uploads, reducing cost by 20%.
- Developed and optimized Python ETL pipelines to parse and ingest terabytes of raw genomic sequence data from FASTA and FASTQ formats, enriching the patient data used to train a predictive cancer progression model.

Data Engineer | United States Department of Homeland Security **MAY** 2021 – August 2022

- Coordinated with SME and key stakeholders to define data policy and established governance framework; developed data dictionaries, data catalogs, and data lineage documentation.
- Designed an Enterprise Data Warehouse to consolidates disparate data sources from multiple DHS agencies into a harmonized schema, providing a single source of truth for DHS data assets.
- Engineered and automated data ingestion pipelines using Informatica PowerCenter, creating reusable mappings and workflows to extract and transform data from diverse agency systems into the central Enterprise Data Warehouse.

Data Analyst | FEMA **AUGUST** 2018 – May 2021

- Constructed data pipelines with *AWS Glue* to seamlessly ingest and transform legacy datasets from Oracle and Postgres into *AWS S3 data lake house*, leveraging *PySpark* for efficient data transformation at scale.
- Designed and implemented a complex *AWS Athena query* system to retrieve and process data from AWS S3 data lake, enabling faster insights for the data science team.
- Engineered a real-time data pipeline using Kafka to stream and ingest high-volume disaster data into a Hadoop (HDFS) data lake. Subsequently deployed Spark applications on AWS EMR for large-scale aggregation, reducing data processing time for critical reports by over 30%.
- Collaborated with stakeholders to create standardized data models and enforce consistency across relational databases (AWS Aurora, Oracle) and non-structural repository (AWS S3, DynamoDB).
- Constructed a layered data architecture with a normalized central Layer and a staging Area; tailored data marts to meet the unique analytical needs for different reports.

EDUCATION

OLIVET NAZARENE UNIVERSITY Bourbonnais, IL

Bachelor of Engineering, Major in Electrical Engineering 2012 – 2016

TECHNICAL SKILL

- **Cloud/Big Data:** AWS (S3, Lambda, Glue, EMR, Athena, Aurora), Spark, Hadoop (HDFS), Kafka
- **Programming & Scripting:** Python (PySpark, Pandas, NumPy, SQLAlchemy), SQL (T-SQL, PL/SQL), Node.js
- **Databases:** RDBMS (Postgres, Oracle, SQL Server, MySQL), NoSQL (Neo4j, MongoDB, DynamoDB), JSON, XML
- **Data Warehousing:** ETL/ELT Pipeline Design, Data Modeling, Data Governance, Informatica PowerCenter, Airflow
- **BI & Visualization:** Power BI, Tableau, Excel
- **DevOps & Tools:** Git, GitHub Actions (CI/CD), Jira, Docker, Agile/Scrum

CERTIFICATION

- Meta Database engineer - Jan 2025
- Google Advanced Data Analytics Specialization - Mar 2024