Wesley Lau

(571) 354-9369 • [wesleylau.wcl@gmail.com](mailto:wesleylau.wcl@gmail.com) • [LinkedIn](https://www.linkedin.com/in/wesley-lau-34a8b374/)

**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

* Enhanced cancer research capabilities by leading the integration of two external data systems, expanding the Neo4j graph model to support a 50% increase in query load and provide researchers with unified access to previously siloed datasets.
* 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.
* Built event-driven ETL automation with *AWS Lambda* function with Python that triggered the ETL process based on S3 uploads, reducing cost by 20%.
* Fueled a predictive cancer progression model by engineering optimized Python ETL pipelines to ingest and process terabytes of complex genomic data (FASTA/FASTQ), increasing the richness and accuracy of the model's training dataset.

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

* Established a unified data governance framework by collaborating with SMEs, resulting in a 40% reduction in data inconsistencies and providing a clear data lineage that passed all compliance audits.
* Enabled cross-agency analytics by designing a centralized Enterprise Data Warehouse, consolidating data from 5 DHS agencies into a single source of truth and reducing report generation time for leadership from weeks to days.
* 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.
* Accelerated data science initiatives by implementing a high-performance AWS Athena query system, reducing data retrieval times by over 60% and empowering the data science team with near real-time access to disaster-related datasets.
* 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 Engineering2012 – 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