Alex Chen

Brooklyn, NY | itsalexchen@gmail.com | LinkedIn | GitHub | Portfolio

EDUCATION

New York City College of Technology, CUNY Brooklyn, NY
Bachelor of Science in Data Science | GPA: 3.9 June 2024
Relevant Coursework: Database Fundamentals, Machine Learning for Physics, NoSQL Technologies, Information Retrieval

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript, TypeScript, HTML, CSS, Sass, Golang, Rust, Java, PHP, Bash
Libraries/Frameworks: Pandas, React, Scikit-learn, TensorFlow, Express, Next.js, Flask, Tailwind CSS, Apache Spark
Technologies: Git, Postgres, Docker, dbt, MongoDB, Redis, Terraform, Kafka, AWS, Neo4j, Kubernetes, Databricks, Excel

EXPERIENCE

Metropolitan Transportation Authority **New York, NY**
Tech Fellow (Data Engineer Intern) *June 2023 - Present*

- Architected and built ETL data pipelines for capturing and storing over **50MB** of transactional data per day about transit reports using Python, dbt, and Airflow, synchronizing data across multiple systems for 2+ teams
- Engineered a CDC ELT data pipeline to capture work request data into a data warehouse using, Delta Lake, dbt, and DuckDB, leading to a **75% decrease** in data latency and enabling immediate data analysis for business analysts
- Led the development of a PDF data pipeline using Apache Tika, TheFuzz, and Regex, capturing 25+ data points about work train requests and delivering data to stakeholders with accuracy **surpassing 70%**
- Collaborated with engineers and stakeholders to develop a mobile-friendly web application using Python, Flask, and Sqlite to streamline the work train request process, increasing workflow efficiency by **200%**
- Orchestrated the deployment and maintenance of containerized data-related services with **three-nine availability**, minimizing downtime and ensuring stakeholders have access to critical data, documentation, and applications
- Directed discussions with stakeholders to clarify project scope and attain consensus on deliverables for a mobile web application that collects transit service change requests, resulting in a **20% reduction** in project lead times

Develop for Good **Remote**
Software Engineer Volunteer (Data) *April 2023 - Aug 2023*

- Collaborated with 8 engineers and clients to implement a CDC data pipeline that moves website analytics data from BigQuery to a Postgres instance using Python and Airflow, resulting in a **40% reduction** in client costs
- Extracted data from BigQuery using Python and the BigQuery API, ensuring efficient and timely transfer to a virtual machine for downstream processing, resulting in improved data availability for the data pipeline

The City University of New York **New York, NY**
Energy Technology Intern (Backend) *July 2022 - Aug 2022*

- Led the end-to-end testing of an Excel app used to document appliances within all office and campus buildings using Python, resulting in over **30% faster** processing times and a significant reduction in app crashes
- Developed a Python script that automates data processing of raw energy data with over 100,000 data points into a report that can be referred back to later, saving time manually querying data by **50%** for 2 analysts
- Enhance software reliability by implementing Python unit tests, covering critical functionality of the Django app to manage building energy infrastructure, resulting in an increase in code coverage from **50% to 100%**

PROJECTS

NYC Taxi Data Pipeline - GitHub

- Designed and implemented an ELT pipeline to ingest and process over **3 million NYC taxi trip records monthly** using Databricks and Apache Spark, empowering data consumers with quick access to trip data for their analytical needs
- Architected and built a data lakehouse using dimensional modeling techniques and star schemas, **improving query performance by 25%** and enabling complex analytics on trip patterns and costs
- Accelerated decision-making processes by implementing a user-friendly data mart, enabling data consumers to perform ad-hoc queries and visualize KPIs through interactive dashboards and reports

Wikipedia Information Retrieval System - GitHub

- Engineered a full-stack web application using Python, Docker, and Streamlit that allows users to efficiently retrieve information from Wikipedia articles for users through an intuitive search engine and a large language model
- Implemented microservices utilizing Flask, Docker, and Ollama to optimize the web application's core functionality, resulting in improved scalability for requesting data from the information retrieval system and the llama2 model

Job Tracker Web Application - GitHub

- Developed a full-stack web app using TypeScript React and Express that allows users to manage items in their job tracker while seeing the data about the roles through data visualizations