Xuhesheng Chen

4706597805 | chxuhesh@gmail.com | Location: OC, MI | linkedin.com/in/XuheshengChen | https://chxuhesh.github.io/

SKILLS

- **Programming:** Python (Pytorch, Tensorflow, Pandas, NumPy, Sk-learn, Beautifulsoup4), Java, SQL, Javascript
- **Databases**: PostgreSQL, MySQL, BigQuery, Athena(S3), Spark, Cassandra, MongoDB
- **Tools**: Apache Airflow, Mage, dbt, Kafka, Kubernetes, Docker, Terraform
- **Certification**: Google Cloud Professional Cloud Architect, Google Cloud Professional Data Engineer, Tableau Desktop

PROFESSIONAL EXPERIENCES

eMatrix Energy Systems. Inc **MI, US**

Data Engineer *May.2022 - Present*

Battery Inspection and EOL Report

- **Automated Quality Control**: Spearheaded the development of an advanced program that automates battery product inspections, integrating real-time data analysis and validation to ensure consistent quality and reduce manual intervention, boosting runtime efficiency by 300% and saving 140% in costs.
- **Cloud Deployment & CI/CD Automation**: Streamlined the deployment process by implementing a robust CI/CD pipeline with GitLab, enabling continuous integration and seamless updates to the program. Successfully deployed the solution on Google Cloud, utilizing Google Cloud Storage (GCS) for secure and scalable data management.
- **ETL Pipeline Engineering & Workflow Orchestration**: Architected and deployed ELT pipelines that efficiently transform and load test records into BigQuery and cloud-based SaaS data warehouse. Leveraged Apache Airflow to orchestrate workflows across GCS, Dataflow, and BigQuery, ensuring reliable, scalable, and automated data processing in a cloud environment.
- **Real-Time Monitoring & Insights**: Developed and integrated real-time dashboards in Looker, linked with BigQuery, to provide stakeholders with immediate access to actionable insights.

Microservices for ETL Automation and Remote Device Management

- **Microservices Architecture**: Designed and developed multiple microservices using the Flask framework to automate ETL of data from local manufacturing devices to a cloud-based SaaS data warehouse, ensuring scalability, modularity, and ease of maintenance.
- **Data Integrity & Automation**: Integrated robust error-handling mechanisms and data validation processes within the ETL pipelines, ensuring the integrity and accuracy of the data transferred to the SaaS platform. Automated data synchronization tasks, reducing manual oversight and ensuring that production data is consistently up-to-date and accessible for analysis.

Electrical Safety Testing Automation Program

- **Automated Testing & Inspection**: Architected and developed a robust electrical safety testing program leveraging Python and PLC integration to automate the inspection and production testing processes, reducing manual intervention, enhancing testing accuracy, and ensuring compliance with stringent safety and quality standards.
- **Safety & Efficiency Enhancement**: Designed operator-friendly interfaces using Tkinter to guide operators through testing procedures, significantly reducing human error and improving overall testing efficiency. Integrated real-time monitoring and alert systems to proactively identify and mitigate potential safety risks during testing.

Ernst & Young Parthenon **Shanghai, China**

Data Scientist *Feb.2021 - Jun.2021*

- Leveraged Python (Pandas, Numpy), SQL, and Apache Spark to collect and preprocess over 20 million rows of multi-source datasets related to the oil industry's development over the past 20 years, ensuring data integrity and consistency for robust analysis.
- Developed predictive models using scikit-learn, XGBoost, and TensorFlow, incorporating over 50 features, including oil prices, material, and labor costs, and city GDP to forecast industry revenue for the next 5 years.
- Utilized NLP techniques with SpaCy to extract insights from expert interviews and focused on upgrading nationwide and regional Hospital Management Information Systems, with a total projected investment of 60 billion dollars over 3 years.

PROJECT

Real-time Serverless Bike Analytics

- Spearheaded the end-to-end development of a data pipeline that processed over 30 months of Citi Bike trip data. This pipeline supported real-time and batch processing, enabling comprehensive analysis of over 50 million ride records.
- Implemented Docker for consistent environment management and Terraform to automate and manage cloud infrastructure. Leveraged Google Cloud Storage as a data lake and BigQuery as a data warehouse, enabling efficient storage, querying, and processing of large-scale datasets. Integrated dbt for streamlined data transformation, resulting in faster data processing times by 30%.
- Delivered actionable insights by analyzing data to uncover the most popular bike destinations, peak usage times, and trends in bike type preferences, visualized in an interactive Looker Studio dashboard, driving data-driven decisions for urban planning.

EDUCATION

University of North Carolina at Chapel Hill **NC, US**

Master of Information Science, GPA: 3.8/4.0 *Sep.2021 - May.2023*

- Core Course: Deep Learning, Machine Learning, Natural Language Processing, Information Retrieval, Data Mining

East China Normal University **Shanghai, China**

Bachelor of Engineering in Software Engineering *Sep.2010 - Jun.2014*

- Main Course: Probability Theory and Statistics, Database, Java Programming, Data Structure and Algorithms