Mustafa Alsaegh

Portfolio **812-361-0677** **malsaegh1@gmail.com** **LinkedIn** **GitHub**

SUMMARY

Data Scientist & Engineer skilled in building scalable data pipelines, automating workflows, and deriving actionable insights. Proficient in Python, SQL, Spark, BigQuery, and Power BI, with expertise in ML, NLP, and cloud infrastructure.

WORK EXPERIENCE

Data Scientist, Ellucian *Columbus, IN* Jun 2025 - Present

- Partnered with SMEs to design and implement rule-based data transformation logic for the Price FX platform, ensuring incompatible product combinations were excluded, thereby enabling the creation of valid, business-approved bundles.
- Built a Python-based validation framework on Databricks, leveraging Pandas and NumPy to cross-check LLM-generated numeric rule translations from complex AND/OR logic, achieving **100%** accuracy and automating compatibility checks.
- Engineered a dynamic product bundling algorithm leveraging PySpark to group compatible products based on business-defined thresholds, perform cost calculations, and ensure results remained within an acceptable **5%** deviation range.
- Automated CI/CD quality checks by integrating SonarQube with GitHub Actions, enabling bug and code smell detection during pipeline runs and reducing cognitive complexity by **60%**.

Graduate Research Assistant, Luddy School of Informatics. *Bloomington, IN* Nov 2024 - May 2025

- Employing Python and CNN models to analyze NOAA satellite imagery and geospatial features, such as elevation, to predict flood-prone areas to enhance disaster management strategies.
- Utilized Python's GDAL library to process and optimize large-scale satellite images exceeding **100 GB**, improving geospatial data analysis and rendering.
- Developed a custom web application using Node.js and Flask that enables visualization and annotation of geographic data, marking flooded regions, which helped in generating more than **600 training samples** for machine learning models.

Graduate Research Analyst, Kelly School of Business. *Bloomington, IN* May 2024 - Jul 2024

- Refactored and optimized ETL pipelines using T-SQL and PySpark to process **10+ million** ad campaign records, significantly improving large-scale data handling and computational performance on distributed systems.
- Applied causal inference techniques in R to assess the business impact of promoted ads on seller performance, generating insights that improved ROI, enabling eBay's Ads Ranking team to train ML models and optimize budget allocation and ad performance.

Sr. Data Analyst, IQVIA. *Kuwait City, Kuwait* Jul 2021 - Jul 2023

- Orchestrated the migration of **5+** TB of enterprise data from Oracle and on-prem databases to Data Lake, leveraging ADF, Synapse, and Databricks to automate workflows and improve scalability, accessibility, and integration of data pipelines.
- Enhanced data management efficiency by **80%** through process automation using Power Automate, and developed interactive dashboards in Power BI to visualize trends and reduce manual reporting efforts.
- Led a team of **3** IT professionals, providing round-the-clock client support, ensuring service availability, and driving team performance and client satisfaction through technical training and mentorship.

PROJECTS

GCP-Enhanced Airline Satisfaction Prediction.

- Designed and implemented an airline customer satisfaction analytics platform on Google Cloud, integrating Python-based ML models in Jupyter Notebook, containerized with Docker for scalable and consistent deployment.
- Managed cloud storage and large-scale data processing with BigQuery, achieving **98%** prediction accuracy using RandomForestClassifier, and delivered actionable insights through interactive visualizations to support data-driven service improvements.

Multi-Agent Stock Analysis and Portfolio Management System.

- Built a multi-agent stock analysis platform using Flask and React to deliver real-time investment insights, integrating **4** LLM-powered agents for trend analysis, risk assessment, and price forecasting, enabling informed portfolio decisions.
- Developed and deployed **5+** Flask RESTful APIs and React frontends to support user interaction and data querying, integrating the Finnhub API for news-based risk scoring and delivering actionable insights (BUY/SELL/HOLD) through automated analytics workflows.

LLM-based English-to-Spanish Translation Model.

- Fine-tuned a Google T5-based LLM on a dataset of **100K+** English–Spanish sentence pairs, leveraging LangChain to build agentic AI workflows for dynamic text generation and seamless API integration.
- Containerized training and inference pipelines with Docker to ensure environment consistency, and deployed the model as a Flask API for scalable access, integrating logging and monitoring mechanisms to optimize performance and reliability.

SKILLS AND EXPERTISE

Languages:	Python, SQL, Java
Big Data Technologies:	Spark, PySpark, Kafka, Hive, Databricks, Airflow
Cloud Platforms:	GCP (BigQuery, Bucket), AWS (Lambda, EC2), Azure (ADF, Synapse, Blob Storage, Log Analytics)
Tools & Frameworks:	Power BI, Flask, Node.js, QGIS, Git, Docker, Kubernetes
Concepts:	ETL/ELT Pipelines, Data Modeling, Data Governance, CI/CD, Machine Learning Pipelines

EDUCATION

Indiana University Bloomington Aug 2023 – May 2025
Master of Science in Data Science *Bloomington, IN*