MIA SHI

469-403-7557 ◇ minmiashi@gmail.com ◇ Personal Portfolio ◇ https://www.linkedin.com/in/min-mia-shi/

WORK EXPERIENCES

Data Scientist Jan. 2025 - Present
The Sunwater Institute *North Bethesda, MD / Remote*
Full-stack data professional transforming raw data into actionable insights through analytics dashboards, ETL pipelines, and search-based data modeling solutions.

- Engineered cloud-hosted interactive dashboards with Dash, integrating them into front-end applications for real-time data visualization.
- Developed ML/DL speaker identification models using sklearn and PyTorch, achieving 80-90% accuracy.
- Architected real-time data pipelines using AWS Glue, Lambda, Spark, and Python with OpenSearch integration for scalable data modeling and retrieval.
- Implemented automated, scalable AI-driven data pipelines for content generation through AWS Glue, Lambda, Streamlit, and LLMs.

Data Engineer Jun. 2024 - Oct. 2024
The Sunwater Institute *North Bethesda, MD / Remote*
Designed and optimized scalable data pipelines and big data workflows utilizing PySpark and AWS for high-performance data processing.

- Engineered scalable ETL pipelines for large-scale data ingestion, enabling real-time and batch workflows with AWS S3, Glue, Lambda, Spark, and Python.
- Implemented high-performance data workflows with PySpark and Glue for efficient big data querying and transformation.
- Automated speech-to-text and speaker identification pipelines using AWS Transcribe and Textract, achieving 90%+ accuracy with improved processing efficiency.

Data Analyst May 2020 - May 2024
The University of Texas at Dallas *Richardson, TX*
Took responsibility for data analysis for 10+ global health/policy projects using advanced statistical models.

- Managed data collection in diverse methods such as Qualtrics surveys and web scraping using R and Python.
- Developed 20+ robust statistical models (multi-variable and fixed-effect regression, difference-in-difference, time-series), utilized ML models and NLP skills to support correlation and causal inference in research.
- Led a team of five research assistants, ensuring timely project completion and publication.

PROJECTS

Twitter Clone: High-throughput Social Media Backend — *Python, Django* May 2024 - Present
Developing the backend for a social media platform using Django (Python) with HBase, MySQL, and Redis.

- Maximized query efficiency by storing objects with HBase, MySQL & Amazon S3 based on query complexity.
- Addressed N+1 slow query issues by implementing Redis caching and denormalization.
- Integrated Celery and RabbitMQ to establish asynchronous workers with varying priority levels.
- Implemented a push model for distributing news feeds to followers efficiently.
- Optimized memory and resource allocation using recursive small batches of asynchronous tasks.

AI-Powered Payment Service Chatbot for Enhanced Customer Engagement Aug. 2023 - Dec. 2023
Developed and deployed an AI-driven chatbot using Python and MySQL, leveraging advanced NLP techniques to enhance customer engagement for Onyx CenterSource.

- Leveraged NLP and MySQL for analyzing and querying an extensive database containing over 10 million entries.
- Improved response efficiency by 25% and achieved 99% accuracy using the XGBoost model.
- Enhanced user engagement, boosting customer satisfaction and strengthening the company's brand image.

Kaggle Plant Pathology Competition: Leveraging Deep Learning CNNs Nov. 2023 - Dec. 2023
Implemented deep learning models using Python and PyTorch to enhance disease identification accuracy in crops.

- Applied transfer learning on CNNs with 13,042 images across 12 categories, significantly improving accuracy.
- Performed image augmentation techniques (rotation, flipping, zooming, noise injection) to enhance data.

- Fine-tuned ConvNext DL models, achieving 86.8% accuracy and securing a Top 3 ranking in the competition.

Optimizing Big Data Risk Analysis for a Company with Hadoop and Tableau Aug. 2022 - Dec. 2022
Engineered data visualization dashboards using Tableau, linked to Hadoop, for business risk analysis.

- Processed and analyzed geospatial data with Hadoop, Hive, and Spark, reducing significant processing time.
- Developed Tableau visualizations linked to Hadoop and built interactive dashboards for business analysis.
- Conducted linear regression and multivariate analysis, contributing to predictive accuracy by 15%.

EDUCATION

Ph.D. in Political Science (Quantitative Methods & Data Science) Dec. 2024 (***GPA: 3.95/4***)
The University of Texas at Dallas
M.S. in Business Analytics (Data Science & Engineering Track) May 2024 (***GPA: 3.95/4***)
The University of Texas at Dallas

SKILLS

Programming & Tools: Python, SQL, Java, Javascript, AWS Web Services, Streamlit, Dash, R
AI & ML Modeling: Statistical Modeling, Deep Learning, Machine Learning, NLP, LLM, PyTorch, Tensorflow
Database & Big Data: SQL Server, MySQL, PostgreSQL, AWS RDS, Hadoop, Spark, Hive, Impala, Sqoop, Pig
Certificates: Certificate in Applied Machine Learning, AWS Certified Cloud Practitioner, Google Analytics