Waltham, MA 02451
United States

Andy (Xiangyu) Cui

GitHub | LinkedIn | Portfolio
xiangyucui@outlook.com | (402)-853-3000

3 year data engineer
5 years project manager
2 years software engineer
2 years user experience researcher

SKILLS

- Programing languages: Java, Python, JavaScript, HTML, CSS, C/C++, Assembly, VB
- Database: MySQL, PostgreSQL, MongoDB, Redis, SQLite, Hive
- Machine Learning: TF-IDF, Naïve Bayes, GPT-2, Bert, CNN, Transformer
- Framework/Libraries: Spring Boot, React.js, Node.js, jQuery, Django
- Version Control & CI/CD: GitHub, TFS, Azure DevOps, Jenkins, Jira
- Others: AWS, Docker, Maven, Tomcat, Axure RP, Servlet, Junit, Nginx, LaTeX, CAD design, Tableau

WORK EXPERENCE

CAC Auto Group Inc | Full-Time Southborough, MA, USA
Data Engineer | predictive maintenance system for e-commerce platform Feb. 2024-Present

- Develop and maintain a predictive price forecast system for vehicles listed on CarGurus using a serverless architecture on AWS. This system enhances market compatibility for vehicle pricing and optimizes operational efficiency. Leveraged core AWS services such as S3, Lambda, DynamoDB, SNS, CloudWatch, and Kinesis for real-time data processing, with Python as the primary programming language and AWS CloudFormation (YAML) for Infrastructure as Code (IaC).
- Implemented real-time data monitoring using AWS Kinesis to track market data fluctuations, enabling the system to detect and trigger S3 events when vehicle prices deviate from predefined ranges. This near real-time data handling detects discrepancies in vehicle market prices promptly, allowing quick corrective actions to keep vehicle listings competitive and accurate.
- Established Automated Alert notifications through Amazon SNS and integrated with AWS SQS and Lambda to automatically notify the team when specific events are triggered. This system improved the accuracy and market compatibility of vehicle prices displayed on the company's website, resulting in an 80% boost in daily operational efficiency and over 50% improvement in real-time price adjustments compared to industry standards.
- Designed and managed a data lake architecture on AWS, incorporating raw, cleansed, and application data layers stored in S3. This data lake supports the BI team with enhanced access to structured data, improving data processing efficiency by 50%.
- Orchestrated data processing workflows through AWS Step Functions, which automate the execution of multiple Glue Crawlers and ETL jobs to process data and update S3 layers. Enabled continuous monitoring through Glue Data Catalog and Athena to maintain data quality and consistency across all data lake layers.
- Enhanced infrastructure reliability by using AWS CloudFormation to deploy reproducible infrastructure configurations across environments, which streamlined deployment, reduced development costs by 70%, and supported faster product iteration for an agile startup environment.
- Automated SNS notifications to alert the team about Glue Crawler or ETL job failures, improving response time and reducing data pipeline downtime by 30%.
- Implemented a Lambda-based schema change detection system that uses Glue Crawler and Glue Data Catalog to track metadata changes in source data and trigger alert emails for schema inconsistencies. This setup led to a 60% reduction in troubleshooting time and a 50% decrease in data delivery delays.
- Managed Agile processes for development, participating in Scrum meetings and data pipeline structure design, while maintaining version control through Git and tracking project tasks on Jira. Established a continuous integration pipeline via Bamboo from Bitbucket, ensuring streamlined deployments.

AlpaLifeBio Inc | Internship Woburn, MA, USA
Software Engineer of Data Engineer | Analysis of Potential Consumer Demand Dec. 2022-June. 2023

- Built and managed a robust AWS streaming data pipeline to automate biomedical data ingestion from multiple public databases into Kinesis Data Stream. This system processed over 500,000 data entries daily, using Lambda Functions for real-time data transfer and S3 and DynamoDB for efficient, scalable storage and retrieval. This architecture allowed seamless handling of high-volume data with minimized latency and reduced operational costs.
- Configured and optimized a structured SQL database to integrate and process data from diverse biomedical sources. Implemented an efficient tag-processing system for enhanced search and retrieval operations, reducing data retrieval time by 80%. This improvement significantly boosted operational efficiency, making it easier to access and analyze critical information for downstream applications.
- Applied advanced data matching algorithms and TF-IDF (Term Frequency-Inverse Document Frequency) techniques to perform data comparison and accurately identify potential client profiles with a 95% match rate. This methodology directly supported targeted marketing strategies by enabling precise identification of high-value clients within the biomedical field.
- Enabled real-time notifications for data updates by setting up DynamoDB Streams to capture table modifications. Configured Lambda Functions to process these events and send automated alert emails through Amazon SES, enhancing response times by 30% and streamlining the data update workflow.

- Utilized Python and graphical web scraping tools to gather and preprocess extensive biomedical datasets from various online resources. This preprocessing step ensured data quality and consistency, enabling deeper analysis and accurate results for research teams.

Dutchgo LLC | Self-employed & Co-founder Omaha, NE, USA
Software Engineer | Analysis and Strategy Optimization Oct. 2020-Apr. 2022

- Developed a customer demand analysis pipeline during the Covid-19 pandemic using AWS S3 for secure data storage and Lambda for serverless processing, combined with SQL and Python. Built and deployed a Random Forest model on AWS SageMaker to identify high-potential user profiles for conversion, achieving an accuracy of 98%. This automated approach saved considerable processing time and improved prediction accuracy for strategic marketing efforts.
- Established a data ingestion and processing framework using AWS Glue to automate data ETL processes, enabling seamless integration of monthly customer market data from S3 into DynamoDB for structured storage. Employed SQL and Python to preprocess data and built a multiple regression model on SageMaker to predict revenue trends, achieving an 85% accuracy rate. This solution provided valuable insights for strategic decision-making regarding resource allocation and customer segmentation.
- Analyzed user behavior patterns to inform UX design by employing Human-Computer Interaction (HCI) principles. Utilized Axure RP to create interactive software prototypes aligned with user behavior insights. Additionally, used AWS Amplify to deploy and test front-end prototypes rapidly, ensuring seamless integration with back-end data and enhancing user experience.
- Constructed real-time monitoring dashboards using Tableau to track essential investment metrics, such as ROI (Return on Investment), growth rate, and portfolio balance. Integrated AWS Kinesis for real-time data streaming, enabling prompt anomaly detection and root cause analysis. This system provided stakeholders with actionable insights on data trends, supporting a 30% faster decision-making process for adjusting business strategies.
- Set up AWS CloudFormation templates to manage and automate infrastructure provisioning, ensuring efficient resource management and scalability of the analysis system. Leveraged AWS IAM to implement robust access control, securing data and model resources across the team.

EDUCATION | Relevant Coursework | Location & Date

Northeastern University Boston, MA, USA
M.S. in Artificial Intelligence of Khoury College Sept. 2021-Dec.2023

University of Nebraska-Lincoln Lincoln, NE, USA
B.S. in Computer Science of Arts Science College Sept. 2016-May. 2020