YOHAN M MARKOSE
Boston, MA | markose.y@northeastern.edu | (857) 867-5489 | LinkedIn | Git Hub | Website

EDUCATION

Northeastern University | Boston, MA | Sep 2024 – May 2026 (Expected)

- ***Master of Science in Information Systems (GPA: 4.0)***
  Relevant Courses*: Data Science Engineering Methods and Tools**,** Big-Data Systems and Intelligence Analytics**,** Neural Modeling Methods and Tools**,** Programming Structures and Algorithm (PSA)*

Mar Athanasius College of Engineering (MACE) | Kerala, India | Aug 2016 – Jun 2020

- ***Bachelor of Technology in Mechanical Engineering***
  Relevant Courses: *Calculus, Linear Algebra and Complex Analysis, Probability Distributions*

SKILLS & CERTIFICATION

- **Languages:** Python, SQL, Java, C++, HTML5, CSS3
- **Frameworks & Tools:** Apache Airflow, Docker, MCP, Git, GitHub, Microsoft Excel, Power BI, Pinecone, ChromaDB, React
- **Databases:** Snowflake, MySQL, dbt, MongoDB, Redis Streams
- **Cloude Technologies:** GCP, Google Cloud Run, AWS S3, GitHub Actions, Google Compute Engine
- **Python Libraries:** NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, FastAPI , SQLAlchemy ,Beautiful Soup, Selenium, LangChain, LangGraph, spaCy, PyMuPDF, snowflake-connector-python
- **Certifications:** Microsoft Certified: Power BI Data Analyst Associate

PROFESSIONAL EXPERIENCE

IQVIA | Kochi, Kerala, India
***Software Developer* | Oct 2020 – Apr 2023**

- Extracted, cleaned, and transformed large datasets (500K+ records) from multiple pharmaceutical data sources (API, Dashboards, Flat Files) using **Python (Pandas, NumPy)** to create training datasets for quarterly sales predictive models, ensuring data quality and **feature engineering** for optimal model performance
- **Led the automation team** in my department streamlining workflows in collaboration with DevOps teams by engineering automated ETL pipelines in Python with integrated statistical validation using scikit-learn and **CI/CD deployment** practices, reducing 250+ manual hours per month
- Designed and deployed automated weekly pharmaceutical deliverables using Python (Selenium) with Apache **Airflow** DAG **orchestration** for scheduled workflows, eliminating manual QA bottlenecks and saving more than 75 manual hours per month
- Engineered automated **ML data pipelines** using Python and **SQL** (**Snowflake**) with scikit-learn statistical validation and anomaly detection, reducing data preparation from **10 hours to 10 minutes**, ensuring data accuracy for model training
- Optimized database performance by maintaining and updating product database using SQL (**MySQL**), handling monthly additions of 100,000+ records, refreshes, and data quality monitoring using statistical methods
- **Collaborated with stakeholders** to deliver 15+ monthly reports through data analysis and validation using **Excel** and **Power BI** (integrating multiple data sources with Power Query), ensuring 100% on-time delivery

***Software Developer- Intern* | Jan 2020 – Apr 2020**

- Built **scalable QA automation** solutions and extracted critical business data from client dashboards using Python libraries (Beautiful Soup, Selenium), improving data accuracy by 95%
- Transformed raw data into actionable business insights using Python and Excel, enabling data-driven decision making for pharmaceutical client projects

PROJECTS

Snowflake Pipeline - FRED: *Snowflake, Snowpark, CI/CD Pipeline, Tasks(DAGs), Github Actions, AWS S3* | GitHub Link

- Engineered end-end orchestrated pipeline tracking U.S. Treasury yield curves using Federal Reserve data with real-time extraction, scheduled processing, and cloud storage integration
- Created interactive dashboard displaying yield curve inversions and economic indicators, enabling financial analysts to monitor market conditions and recession predictors effectively

Multi Agent - Agentic RAG (Venture-Scope): *MCP, LangGraph, FastAPI, Pincecone , CI/CD Pipeline, LLM*| GitHub Link

- Architected AI Ops platform with multi-agent orchestration for automated business intelligence, implementing CI/CD deployment pipelines and real-time monitoring workflows helping entrepreneurs make data-driven decisions
- Deployed scalable cloud infrastructure integrating automated workflows with AWS S3 storage, containerized architecture, and comprehensive system monitoring for nationwide business accelerators

Financial RAG Pipeline & Analytics Interface: *Scikit-learn, Pinecone, Airflow, GCP, Docker, Hugging Face* | GitHub Link

- Built RAGFlex system processing 5 years of NVIDIA financial reports plus custom PDFs through multiple parsing strategies and 3 vector database options (Pinecone, ChromaDB, manual)
- Deployed containerized solution with automated orchestration and intelligent metadata filtering, enabling analysts to extract insights from complex financial documents efficiently