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

EDUCATION

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

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

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

- ***Bachelor of Technology in Mechanical Engineering***

TECHNICAL SKILLS

- **Languages:** Python, SQL, Java, C++, HTML5, CSS3
- **Python Libraries & Frameworks:** NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, Beautiful Soup, Selenium, LangChain, FastAPI, Streamlit, LiteLLM, LangGraph
- **Tools & Platforms:** Apache Airflow, Docker, Git, GitHub, Microsoft Excel, Power BI, Pinecone, ChromaDB, Mistral OCR
- **Database & Cloud:** Snowflake, MySQL, dbt, Google Cloud Run ,AWS S3, GitHub Actions, Redis Streams, Google Compute Engine

PROFESSIONAL EXPERIENCE

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

- Led the automation team in my department, streamlining processes and reducing manual QA tasks by 250+ hours/month
- Streamlined data processing workflows by implementing automated ETL pipelines in Python, reducing manual data preparation time from 10 hours to 10 minutes per cycle
- Automated weekly/monthly deliverables using Python (Selenium), saving more than 75 manual hours per month
- Developed Python scripts to integrate data from multiple environments (API, Dashboards, Flat Files) and analyze anomalies using NumPy, Pandas and Matplotlib libraries reducing data processing time by 80%
- Implemented server-based automation for end-to-end QA processes
- Performed data analysis and validation using Excel, Power BI, and SQL, consistently provisioning more than 15 reports each month
- Maintained and updated product database using SQL (Toad), including monthly additions and refreshes
- Mentored 3 interns in Excel, for data analysis, and 4 team members in Python automation

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

- Developed on-demand QA automation solutions and scraped data from client dashboards using Python libraries (Beautiful Soup, Selenium)
- Processed and visualized data using Python and Excel for actionable insights

PROJECTS

NVIDIA FINRAG: *LangGraph, Pincecone , serpapi, FastAPI , , LLM, AWS S3, GCP, Docker* | Mar 2025

- Agentic RAG pipeline that provides specific curated information based on the user's specific queries regarding Nvidia's financial dataset. It utilizes Langchain as its foundation and operates with three distinct tools to serve as its knowledge base. **Snowflake tool:** Serving its stocks and visual analysis, **RAG tool** - Working to server from the extracted quarterly reports using embedding and vector database, and **Web Search tool** - Provides the potential to fetch real time data analysis

Financial RAG Pipeline & Analytics Interface: *RAG, Scikit-learn, Pinecone, ChromaDB, Airflow, Docker* | Mar 2025

- Developed end-to-end Retrieval-Augmented Generation system for NVIDIA financial reports spanning 5 years, incorporating automated data extraction, custom PDF upload capabilities, and OpenAI vector embeddings with Pinecone/ChromaDB
- Implemented metadata filtering and context-aware retrieval achieving 95% accuracy across both pre-loaded and user-uploaded documents. Orchestrated multi-step document processing via Airflow DAGs and deployed containerized solution with Streamlit UI and FastAPI

Federal Reserve Economic Data- Snowflake Pipelines: *Snowflake, Snowpark, Tasks(DAGs), Github Actions, S3* | Feb 2025

- Using Snowflake's Snowpark for Python, the system efficiently extracts, transforms, and validates financial data for advanced analysis and reporting. It builds a pipeline to create a dashboard that displays the inverse treasury yield curve, derived from FRED's U.S. Treasury yield data for 10-Year and 2-Year bonds

Financial Statement Analytics Platform: *Snowflake, Airflow, DBT, Streamlit, FastAPI, AWS S3* | Feb 2025

- Architected an end-to-end data pipeline leveraging Airflow, Snowflake, and AWS S3 to automate extraction and processing of SEC financial datasets. Implemented automated data quality checks using DBT. Developed a full-stack solution with FastAPI backend and Streamlit dashboard for financial data exploration and visualization, enabling efficient analysis of market trends and company performance metrics**.**