AnushKrishna VenkataKrishnan

av3278@rit.edu | github.com/anushkrishnav | linkedin.com/in/anushkrishnav | (585)-981-1189 | New York

Education

Rochester Institute of Technology | Masters, Data Science | GPA: 3.94/4.0 **Expected Dec 2025**
Coursework: Statistical Machine Learning, High Performance Data Science(MPI & CUDA), Applied Statistics
Bharathiar University | Bachelors, Computer Science | GPA: 3.44/4.0 **Aug 2019 - Aug 2022**

Technical Skills

Python | C++ | JAVA | MongoDB | PostgreSQL | Redis | Flask | FastAPI | Spacy | Celery | Numpy | Keras | SQL | GraphQL | Apache Airflow | NLTK | Scikit-learn | Git | Docker | Kubernetes | Azure | Google Cloud | Pytorch | Tensorflow

Experience

ML Engineer (Co-op) August 2024 – December 2024
Metabob San Francisco, CA

- Set up**Prefect** workflow orchestrator infrastructure on **Azure Kubernetes Services** to enable data scientists to utilize cluster resources directly from Python notebooks, reducing experimentation time from months to days.
- Built a **GNN model training pipeline** using Prefect on Azure that processes and trains on over **100k language specific code snippets**, enhancing model's performance .
- Contributed to the creation of high-quality corpora for GNN bug detection models, with over **800k rows of data** extracted and cleaned every day by the pipeline.
- Created comprehensive documentation on the infrastructure to ensure seamless adoption.

Data Science Researcher September 2023 – Present
Rochester Institute of Technology Rochester, NY

- Co-authored a study analyzing **17,000+ developer-ChatGPT interactions** on GitHub, focusing on refactoring use cases and quality attributes, presented at the **Mining Software Repositories (MSR)** conference in Lisbon.
- Conducted research on leveraging **Large Language Models (LLMs)** for third-party library migrations, improving software maintenance and refactoring processes.

Lead Backend Engineer September 2023 – Present
RIT Student Government Rochester, NY

- Revamped the **PawPrints platform** serving **3,000+ RIT students**, enabling efficient petition creation, signing, and distribution.
- Optimized database design and restructured SQL queries, leveraging **Postgres full-text search** to improve search performance by **10x**.
- Spearheaded the migration from a **Django monolith** to **FastAPI microservices**, ensuring scalability and modularity.

Data Engineer Dec 2021 – Jul 2023
Metabob Mountain View, CA

- Worked Part-time remote during the final year of Undergrad
- Created a data collection pipeline by incorporating **Celery & Redis** for job queue management, **Postgres** for data storage and **Kubernetes** for scaling the pipelines in order to improve code, commits and textual data extraction .
- Developed a **Dask-based NLP pipeline** on **GKE and AKS**, resulting in a remarkable **52% reduction in preprocessing and training time**. This parallelization enhanced **batch training** efficiency.
- Oversaw the seamless **infra migration** of the Dask ML pipeline and Data Collection Pipeline from **Google Cloud Platform to Azure**, ensuring continued functionality and operational stability.

Publications

- E. A. AlOmar, **V. AnushKrishna**, M. W. Mkaouer, C. Newman, and A. Ouni. "How to refactor this code? An exploratory study on developer-ChatGPT refactoring conversations". In Proceedings of International Conferences on Mining Software Repositories, 5 pages, 2024 [MSR].

Projects

PayTrack | *Python, Postgres, Azure Kubernetes Service, Streamlit, GitHub Actions, Helm*

- Implemented an Airflow **ETL pipeline** that automates data retrieval from the time-clock API, parses it and loads it into a Postgres database in batches.
- **Developed a dashboard** in Streamlit to manage working hours for two of my on-campus jobs, providing in-depth statistics and weekly work status tracking.

DevGPT Pipeline | *Python, Dask, Spacy, SQLAlchemy, SQLite*

- Developed a robust data pipeline for DevGPT repository mining, utilizing Python, Dask, and Spacy to **process 60,000+ rows of data efficiently within two hours.**
- Implemented SQLAlchemy ORM models and SQLite Database for **effective storage of diverse data objects**.
- Conducted thorough data cleaning using **regular expressions** and **Spacy's stop word corpus**, ensuring consistent formatting across various text sources.