Vani Sai Kumar Komarraju

513-400-2992 | k.vanisaikumar@gmail.com | linkedin.com/in/vanisai

EXPERIENCE

Data Engineer — August 2022 – July 2023
Infosys — *Hyderabad, India*

- Designed and maintained scalable data pipelines on AWS using AWS Lambda, Amazon S3, and Amazon Redshift to support efficient ingestion and processing of provider, patient, and claims data across multiple healthcare systems.
- Built real-time data streaming solutions using Amazon Kinesis and Apache Kafka to enable timely operational monitoring across healthcare workflows, and utilized AWS Athena for ad-hoc querying of clinical and operational data stored in S3, enabling fast insights into customer behavior and operational metrics
- Developed interactive healthcare dashboards using Power BI, visualizing KPIs related to claims processing, provider performance, and patient engagement to assist executive decision-making.
- Implemented relational databases using Amazon RDS (PostgreSQL, MySQL) and Amazon Redshift to store and query large-scale provider and claims data, and processed high-volume healthcare transactions using Apache Spark on Databricks and dbt, improving reporting accuracy and enabling timely, data-driven clinical insights.
- Utilized Microsoft SQL Server, Oracle, and Teradata for advanced SQL development, building complex stored procedures and views to support regulatory reporting, risk stratification, and healthcare analytics.
- Integrated Power BI with AWS Redshift, SQL Server, and Snowflake to deliver real-time performance tracking and compliance reporting, and containerized healthcare microservices using Docker, orchestrating deployments with Kubernetes to enhance scalability, resilience, and operational efficiency.

SKILLS

Programming Languages: Python, SQL
Python Libraries: Scikit-learn, Matplotlib, Seaborn, TensorFlow, PyTorch, Keras, OpenCV, NLTK
Big Data & ETL: Apache Spark, PySpark, Apache Kafka, Airflow, Databricks, dbt, Snowflake
Data Visualization Tools: Tableau, Power BI
Cloud Platforms: AWS (EC2, Lambda, Redshift)
Databases & Querying: PostgreSQL, MySQL, Oracle, SQL Server, MongoDB
DevOps & Containers: Docker, Git/GitHub, CI/CD
Project Tools: JIRA, Microsoft Office Suite

PROJECTS

Sentiment Analysis of Product Reviews using Python and NLP | *Python, Pandas, NumPy, Scikit-learn, Matplotlib, Power BI, NLTK, Gemini AI, EDA*

* Conducted sentiment analysis on consumer reviews, mimicking user behavior tracking similar to analytics platforms like Google Analytics, and visualized insights via Power BI for strategic decision-making.

State Farm Distracted Driver Detection | *Python, TensorFlow, PyTorch, Data Augmentation, Keras, Flask, CNNs, OpenCV*

* Developed a deep learning model using TensorFlow and PyTorch to classify driver behaviors into 10 categories. Achieved 92% accuracy through CNN optimization, data augmentation, and API integration.

Student Dropout Prediction | *Python, Pandas, Scikit-learn, NumPy, Matplotlib, Seaborn, Logistic Regression, K-Means, RMSE*

* Built a predictive model using logistic regression and K-Means clustering on NC DPI dropout data, integrating socio-economic factors. High accuracy and low RMSE were achieved while identifying county-level risk patterns for targeted interventions.

EDUCATION

University of Cincinnati — Cincinnati, OH
Master of Engineering in Data Science, GPA 3.7 — *May 2025*