Rajoju Sai Charan

Denton, TX | rajojusaicharan1@gmail.com | +1 940-300-2732 | linkedin.com/in/sai-charan-71b104206

Summary

Data Science Master's student at the University of North Texas with a **4.0 GPA** and over two years of professional experience in data engineering and analytics. Proven ability to develop end-to-end data pipelines, validate large-scale datasets (**120M+ records**), and build dashboards that drive business decisions for global clients like **Nike** and **Lloyds Banking Group**. Eager to contribute advanced skills in **machine learning**, **cloud computing** (GCP, Azure, AWS), and predictive modeling to a challenging Data Science or AI/ML internship.

Technical Skills

Programming & Scripting: Python (Pandas, NumPy, scikit-learn, PySpark), **SQL**, Java, HTML, CSS, JavaScript

Machine Learning & AI: Classification, NLP (TF-IDF, Word2Vec, GloVe, **BERT**), **Deep Learning**, Predictive Modeling, Feature Engineering

Big Data & Cloud Platforms: Databricks, **Apache Spark**, **GCP** (BigQuery, Dataproc), **Azure** (ADF, Data Lake, SQL DB), **AWS** (ML Associate Certified)

Data Visualization & BI Tools: Tableau, Looker, Looker Studio, **Power BI**, Matplotlib, Seaborn

Data Engineering & ETL: Data ingestion pipelines, Data migration, **Data quality validation**, Automation workflows

Version Control & Tools: Git/GitHub, Azure Repos, Jira, Agile/Scrum

Professional Experience

Project Engineer (Client: Nike), Wipro Limited — June 2023 – July 2024

- Supported the deployment of new regional size curve models in Spring 2024 by performing data discovery, pre-processing, and data validation in **Databricks**; these models are projected by Nike's analytics team to drive over **$11M** in incremental annual value.
- Utilized **Databricks** and **Python/PySpark** to automate data aggregation and feature engineering workflows, supporting the team's product assortment and price optimization models.
- Managed the weekly monitoring of allocation forecasting models, conducting **SQL** analyses to validate model performance against KPIs and provide data insights to Global store allocators.
- Supported a price elasticity analysis by preparing datasets and validating results, which identified key opportunities in discount strategy. Collaborated with senior analysts to present findings and data-driven pricing recommendations that were subsequently adopted to improve profit margins.

Project Engineer (Client: Lloyds Banking Group), Wipro Limited — September 2022 – May 2023

- Developed and maintained **Looker dashboards** to track the migration of **15+ TB** of customer and transaction data from **Teradata** to **Google BigQuery**, streamlining ETL reporting and reducing **8+ hours** of manual work weekly.
- Validated over **120M+** customer account records post-migration by executing systematic **SQL** checks in **BigQuery**, documenting and resolving data quality issues within an Agile framework (**Jira**) to help achieve a **99.8% data integrity rate**.
- Prepared **5M** rows of customer communication log data using **Python (Pandas)**, delivering a clean, labeled dataset to the Data Science team for customer risk identification models.

Trainee, Wipro Limited — March 2022 – June 2022

- Assisted in building and testing data ingestion pipelines using **Azure Data Factory (ADF)** to move data from **Azure Data Lake Storage** into an **Azure SQL Database**.
- Contributed to a data migration project by performing data quality checks and writing **SQL queries** to validate data for accuracy and consistency, ensuring a reliable dataset for downstream **Power BI** reporting.

Projects

Real-Time Patient Monitoring: A Big Data Pipeline on GCP

- Architected and deployed an end-to-end big data pipeline on **Google Cloud Platform (GCP)**, processing both batch clinical records and simulated real-time patient vitals for a predictive health monitoring system.
- Utilized **PySpark** on **Dataproc** to perform large-scale ETL, joining 5 static EHR datasets with a Python-generated stream of vital signs and engineering critical monitoring features like rolling averages and alert flags.
- Loaded the enriched data into **BigQuery** for high-speed querying and developed a **Looker Studio dashboard** to visualize key patient metrics and real-time alerts, providing a proof-of-concept for a clinical monitoring interface.
- **Technologies Used:** GCP (Dataproc, BigQuery), Apache Spark (PySpark), Python, Looker Studio

Predicting Mental Health Disorders Using Social Media Data

- Developed and compared **5 machine learning models** to predict mental health disorders, training on a corpus of over **50,000 posts** from Reddit and Twitter.
- Engineered features using **TF-IDF**, **GloVe/Word2Vec**, and **BERT embeddings**; leveraging **BERT transformers** boosted classification accuracy by **12%** compared to baseline TF-IDF models.
- Achieved a peak **F1-score of 0.82** with the fine-tuned **BERT model**, demonstrating a significant lift over traditional algorithms like logistic regression.
- **Technologies Used:** Python, BERT, scikit-learn, Twitter API

Fake News Detection Using Machine Learning

- Built a **Python**-based classification model to detect fake news, achieving **85% accuracy** on a Kaggle dataset.
- Performed data preprocessing, including text cleaning and **TF-IDF feature extraction**, reducing noise by **40%**.
- Developed a user-friendly interface to display results, improving user interaction time by **50%**.
- **Technologies Used:** Python, scikit-learn

UNT Hydro Insight 2024 – Kaggle Competition

- Collaborated with a team of 4 to analyze hydrological data, securing a **top 5 finish**.
- Conducted EDA to identify 5 key features, improving model performance by **12%** on 2 evaluation metrics.
- Analyzed competition leaderboard trends, providing insights that contributed to the team's top 5 finish.
- **Technologies Used:** Python, pandas, NumPy

Certifications

AWS Certified Machine Learning – Associate (2025)

Education

University of North Texas (UNT), Master of Science in Data Science   May 2026

- **GPA: 4.0/4.0**
- **Relevant Coursework:** Deep Learning for Big Data, Generative AI & Large Language Models, Applied Machine Learning, Data Modeling, Data Visualization, Data Storage & Retrieval.

Kakatiya Institute of Technology & Science (KITSW), Bachelor of Mechanical   May 2022
Engineering

- **CGPA: 8.09/10**