ANSHUMAN SHARMA

Lean Six Sigma Green Belt | Program Ambassador MDS | Top Computer Science Voice - LinkedIn
Irvine, California | (949) 279-7105 | anshums3@uci.edu | linkedin.com/anshumansharma23/ | Portfolio/Anshuman

EDUCATION

• **University of California, Irvine**	Master of Science in Data Science	GPA : 3.61/4.00	Sep 2022 - Dec 2023
• **Amity University, Noida**	B.Tech Computer Science	GPA : 3.88/4.00	Aug 2016 - Aug 2020

EXPERIENCE

City of Irvine | Machine Learning Engineer for Capstone Project | Irvine, CA, USA Sep 2023 - Dec 2023

- Conceptualized and led the development of a **large language model** (LLM) chatbot.
- Optimized a dataset of over **1 million documents** with the Llama Index, improving performance and scalability.
- Enhanced the Llama2 model via **prompt engineering** for targeted use cases, facilitating the retrieval of information from 750,000 contracts.
- Deployed the chatbot via **MLflow**, utilizing **Docker** and **Kubernetes** for streamlined model lifecycle management.
- Instituted **automated pipeline** monitoring, enabling auto-retraining with new data to maintain chatbot efficacy.

Dell Technologies | Data Engineer Intern | Round Rock, TX, USA Jun 2023 - Aug 2023

- Conducted **50+** stakeholder interviews to analyze requirements and identify inefficiencies, leading to strategic data infrastructure improvements.
- Orchestrated integration and analysis from 9+ reports, driving a strategy that saved over **5,000 hours** and boosted productivity by 12%.
- Developed and deployed a data integration framework, consolidating data from over **6 sources** into a centralized warehouse.
- Led the implementation of **BI dashboards**, achieving an estimated annual time saving of over 10,000 hours in the PAN ISG initiative.
- Coordinated the efforts to implement **Apache-Airflow** integration on the data platform to build the next-generation data orchestration.

Accenture | Data Engineer | Bangalore, KA, India Oct 2020 - Jul 2022

- Built **scalable** and **fault-tolerant** pipelines, ingesting millions of records daily by closely collaborating with lead data scientists.
- Ingested SAP HANA data into **Azure,** leveraging Azure data factory for copying and Azure data pipelines for movement management.
- Led cross-functional efforts to define and refine **sales KPIs**, analyze trends, and present outcomes in **Tableau**.
- Achieved a notable **20% decrease** in data transfer delays across the North American region through strategic optimization of data pipelines.
- Worked extensively on **data modeling** using **SAP BW** on HANA-optimized business content for sales and distribution process areas.
- Directed a high-performing team of five to surpass SLA benchmarks, achieving a **98% on-time** delivery rate for critical BI reports.

Samsung Research Institute (SRI-D) | Data Engineer Intern | Noida, UP, India Feb 2020 - Apr 2020

- Built a real-time data ingestion system using **Apache Kafka** to capture and analyze logs from Robot Framework tests in Python.
- Deployed a **TensorFlow**-based **CNN** model for anomaly detection in test outcomes, automating the identification of data irregularities.
- Achieved a **90% reduction** in monitoring workload through **process automation** and the establishment of dependable alert systems.

PROJECTS

Sparkify Music Streaming App | ETL Pipeline and Data Warehouse (Cassandra, Data Modeling, Redshift, AWS, Airflow)

- Developed ETL pipeline for the Sparkify Music Streaming App, utilizing Apache Cassandra for initial data modeling and preprocessing. Transformed raw event data into actionable insights, facilitating efficient analysis of music app history. GitHub
- Implemented an ETL pipeline on Amazon Redshift to accommodate the growing data volume. This enhancement bolstered analytics capabilities, enabling us to build dashboards for deeper insights and informed decision-making. GitHub

Match Me | (API Integration, ChatGPT, Python, Streamlit)

- Developed an innovative web tool leveraging OpenAI's API to automate resume-to-job matching, enhancing recruiters' efficiency by providing insightful, personalized evaluations of candidate-job alignment. Demo

Interchange.com | UCI - Database Development for an E-Commerce website (PostgreSQL, ER - Modelling)

- Crafted and fine-tuned an ER-Model database for an E-Commerce platform, elevating performance, scalability, and fortified security protocols to thwart unauthorized breaches.

PUBLICATION

Malicious URL Classification Using Machine Learning Algorithms and Comparative Analysis Link

- Presented significant research findings at the 3rd International Conference on Computational Intelligence, employing KNN, Naive Bayes, Decision Trees, and Random Forest classifiers to achieve superior accuracy in classifying internet traffic through meticulous analysis.

SKILLS

Programming:	Python, R, SQL.
Frameworks/ Packages:	Apache Airflow, Spark, Data Warehousing (Redshift, Snowflake, SAP BW), Pandas, Numpy, Docker, Hadoop, Git, NoSQL, Apache Kafka, Cassandra, Tableau, Power BI, Looker, Look ML, AWS, Azure, PySpark, Databricks.
Statistical Modeling:	Linear, Logistic & Multivariate Regression, Hypothesis Testing, Chi-Squared tests, Bayesian Methods, Time Series.
Machine Learning:	ML Algorithms, AutoML, Predictive Models, Cluster Models, Scikit-learn, PyTorch, TensorFlow, Keras.