

# ANSHUMAN SHARMA

Data Engineer | Lean Six Sigma Green Belt | Program Ambassador MDS | Top Computer Science Voice - LinkedIn  
Irvine, California | (949) 279-7105 | [anshums3@uci.edu](mailto:anshums3@uci.edu) | [linkedin.com/anshumansharma23/](https://www.linkedin.com/in/anshumansharma23/) | [Portfolio/Anshuman](#)

## EDUCATION

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

## EXPERIENCE

**City of Irvine | Data Engineer for Capstone Project | Irvine, CA, USA** Sep 2023 - Dec 2023

- Led the conceptualization, development, and successful deployment of a state-of-the-art Large Language Model (LLM) system, revolutionizing employee engagement and bolstering access to vital services for the City of Irvine
- Engineered and Implemented LLM solutions, utilizing a comprehensive tech stack including Llama2, Python, and Streamlit. This strategic approach streamlined query processing, markedly improving efficiency and user experience.
- Adopted and led Scrum practices to drive agile project management, resulting in an 80% improvement in query response times.
- Managed and optimized a massive dataset comprising over 1 million documents, ensuring robust system performance and scalability.

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

- Led over 50 in-depth stakeholder interviews to meticulously analyze requirements, identify inefficiencies, and devise strategic enhancements to team's data infrastructure.
- Engineered and deployed an innovative data integration framework, facilitating seamless data consolidation from 6+ disparate sources into a centralized data warehouse.
- Orchestrated the integration and analysis of insights from 9+ distinct reports, implementing a comprehensive data consolidation strategy. This initiative directly contributed to a remarkable annual savings of over 5,000 hours and elevated productivity by 12%.
- Led the implementation of BI dashboards, achieving an estimated annual time saving of over 10,000 hours in the PAN ISG initiative.

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

- Spearheaded cross-functional initiatives with Sales Executives and Data Engineers to establish and refine sales KPIs, identify trends, and present performance results using Power BI.
- Collaborated with senior Data Engineers to architect and implement robust, scalable data pipelines, adeptly handling millions of records daily and significantly boosting data processing efficiency.
- Achieved a notable 20% decrease in data transfer delays across the North America region through strategic optimization of data pipelines.
- Directed a high-performing team of five to surpass SLA benchmarks, achieving an impressive 98% on-time delivery rate for critical BI reports.
- Utilized SAP BW techniques, such as data partitioning and selective deletion, to optimize system performance and enhance analytics workflows.

**Samsung Research Institute (SRI-D) | Data Engineer Intern | Noida, UP, India** Feb 2020 - Apr 2022

- Engineered a real-time data ingestion and monitoring system using Apache Kafka to capture and analyze logs from tests executed via the Robot Framework in Python.
- Established an alert system to promptly identify failures resulting in a 80% reduction in system monitoring, streamlining QA processes.

## PROJECTS

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

- 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 the startup 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](#)

## 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
- Other:** Project Management (Agile), Stakeholder Communication, JIRA for Task Management, A/B Testing