SHARDUL CHAVAN

shardulc36@gmail.com ◇ (857)313-5138 ◇ Union City, CA ◇ linkedin/shardulchavan36

EDUCATION

Northeastern University, Master of Science in Information Systems, **GPA: 3.72/4.00** Expected December 2024
Relevant Coursework: Big Data Engineering and Intelligence Analytics, Data Science Engineering

Mumbai University, Bachelor of Computer Engineering July 2019 - June 2022

EXPERIENCE

Research Assistant, Northeastern University, Boston, MA September 2024 - Present
Tech Stack: Python, MSSQL, GCP, Docker, JWT, OpenAI, LLMs, Git CI/CD

- Developed Airflow pipeline to scrape data, generate **vector embeddings**, and store text in Azure SQL and embeddings in **Pinecone**, enabling multimodal architecture for precise query responses and real-time news delivery every 10 minutes
- Leveraged **Langchain** and NLP techniques with LLM models to provide precise query responses, optimizing retrieval through similarity search and text summarization, enriching content delivery by 30%
- Deployed AI-driven news application on **GCP** with Docker and FastAPI, leveraging **Git CI/CD pipeline** to ensure scalable, production-ready infrastructure

Data Engineer - Analytics, Skyworks Solutions Inc., Boston, MA January 2024 - June 2024
Tech Stack: Python, SQL, Docker, ETL Development, Azure SQL Server, Databricks, PowerBI

- Constructed 25+ data pipelines using **Python, SQL, and Airflow** to extract, transform, and load (ETL) data from Azure Data Lake, SQL Server, and flat files into SQL Server Data Warehouse
- Implemented Slowly Changing Dimensions (SCD) and advanced data transformation techniques with **Databricks SQL** and PySpark APIs, improving historical tracking and data quality by 30%
- Collaborated with engineering and business teams to create data models and **PowerBI dashboards** for OPEX, product analysis, and prototype builds, maximizing efficiency across 3+ teams in Business unit
- Engineered Semiconductor product test data analysis application pipeline to extract data from raw test data files into valuable information for RF engineers, by performing data transformation and report generation which optimizes more than 5 hours of manual work on weekly basis

Graduate Teaching Assistant, Northeastern University, Boston, MA September 2023 - December 2023
Tech Stack: Python, Pandas, scikit-learn, MLlib, HDFS, Predictive Analytics & Validation

- Spearheaded ML pipeline using **PySpark** on **Hadoop cluster** to optimize data loading, perform EDA, and automate feature selection using correlation analysis, and chi-squared tests
- Researched and benchmarked **generative AI models** (BART, Da-Vinci, Stable Diffusion), applying BLEU, ROUGE, and few-shot learning techniques, elevating performance by 10%

Data Analyst, Accion Labs, India January 2023 - July 2023
Tech Stack: Python, JavaScript, Snowflake, MySQL, ELT, REST APIs, Natural Language Processing

- Designed and executed ELT scripts processing 10GB+ daily data from **Snowflake data warehouse** to MySQL data warehouse, utilizing advanced SQL functions (CTEs, subqueries, window functions) to reduce query execution time and enable data access for downstream analytics dashboards
- Developed REST APIs using Python and JavaScript, integrated large language model APIs into ServiceNow, boosting virtual agent performance by 20%, and architected vector databases to enhance NLU in knowledge management

SKILLS

Programming Languages Python, Java, SQL, PL/SQL, R, Scala, VBA, C#, shell scripting (UNIX/Linux)
Data Processing & Cloud: Apache Spark, Kafka, Hadoop, Hive, MongoDB, AWS (S3, RDS), Docker, Kubernetes

PROJECTS

AWS-Based Scalable YouTube Data Analytics Pipeline (EC2, IAM, Spark) June 2024 - August 2024

- Architected scalable ELT pipeline to process large YouTube data, using **AWS S3** for storage and **AWS Glue** for cataloging, reducing data preparation time by 30%
- Automated workflows with **AWS Lambda** and optimized SQL transformations in **Athena**, increasing data processing efficiency by 25% and integrated **QuickSight** dashboards, enhancing insights into YouTube trends and performance for 100,000+ videos