Shardul Chavan

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

Experienced Data Engineer skilled in designing and optimizing scalable data pipelines for data processing and analytics. Strong background in Data Modeling, Warehousing, and ETL/ELT development, with expertise in Semiconductor Data Analytics. Proficient in utilizing cloud platforms and big data technologies to deliver actionable business insights.

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 Science* July 2019 - June 2022

EXPERIENCE

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

- Constructed 25+ data pipelines to extract, transform, and load (ETL) data from Azure Data Lake, SQL Server, and flat files into MS SQL Server Data Warehouse
- Enhanced historical tracking and data quality by 30% through SCD implementation and advanced transformations using Databricks SQL and PySpark APIs
- Collaborated with engineering teams to create data models and **PowerBI dashboards** for product analysis, and prototype builds, maximizing efficiency across 3+ teams in Business unit
- Engineered pipeline to transform product test data into insights, saving RF engineers 5+ hours of manual work weekly
- Integrated validation checks, and retry mechanisms enhancing fault tolerance and data accuracy by 20% at scale
- Operated in an Agile environment, utilizing Kanban for effective project management and task tracking

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

- Facilitated weekly lab sessions for 40+ students, delivering instruction on data science techniques and A/B testing, leading to improved student performance and engagement
- Led ML pipeline using **PySpark** on **Hadoop** cluster to optimize data loading, perform EDA, and automate feature selection, improving risk assessment and pricing strategies for Prudential Life Insurance
- Benchmarked GenAI models (LLaMA, GPT-3), using BLEU and ROUGE scores to improve text generation quality

Data Analyst Intern, Accion Labs, India January 2023 - July 2023
Tech Stack: JavaScript, Snowflake, MySQL, Natural Language Processing

- Designed ELT scripts to process 10GB+ daily data from Snowflake to MySQL, leveraging advanced SQL (CTEs, window functions) to streamline data access for analytics dashboards
- Developed REST APIs, integrated large language model APIs into ServiceNow, boosting virtual chatbot performance by 20%, and architected vector databases to enhance NLU in knowledge management
- Compiled POCs, technical documentations outlining implementation processes to ensure detailed reference materials

SKILLS

Programming Languages	Python, Java, SQL, PL/SQL, R, Scala, VBA, C#, shell scripting (UNIX/Linux)
Big Data Tools & Databases	Apache Spark, PostgreSQL, Tableau, DBT, NoSQL, Redis, Hive, Oracle, MongoDB
Cloud & DevOps	GCP, Azure, AWS, Microservices, Kubernetes, Docker, Github actions, Jenkins

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 AWS Lambda workflows and refined SQL transformations in Athena, increasing data processing efficiency, and integrated QuickSight dashboards, improving insights into YouTube trends and performance for 1,000+ videos

AI-Driven News Aggregator, (Python, GCP, Docker, JWT, Git CI/CD) November 2023 - December 2023

- 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
- Deployed app on **GCP** with Docker and FastAPI, utilizing **Git CI/CD** for scalable, production-ready infrastructure