Sparsh Marwah

Boston, MA 02130 | marwah.sp@northeastern.edu | +1 (857) 225-9142 | LinkedIn | GitHub | Portfolio

Education

Northeastern University, Boston, MA **May 2025**
Master of Science in Data Analytics Engineering, GPA: 3.75/4.0
Relevant Coursework: Data Management in Analytics, Data Mining in Engineering, Machine Learning Operations, Financial Management for Engineers

SRM Institute of Science and Technology, Chennai, India **May 2021**
Bachelor of Technology in Computer Science Engineering
Relevant Coursework: Data Structures, Data Science and Big Data Analysis, Object Oriented Analysis and Design
Publication: AI Music Generator (Research paper)

Technical Skills

Programming & Databases: SQL (PostgreSQL, Hive, MySQL), Python (Pandas, NumPy, Matplotlib), NoSQL (MongoDB)
BI Tools: Power BI, Tableau, Microsoft Excel
Data Analysis: Data enrichment, automation, A/B testing, KPI analysis, and scalable insights development
Big Data & Tools: Hadoop, PySpark, Databricks, GitHub
Data Integrity & Reporting: Automated tools for gap identification, trend reporting, and data validation
Certifications: Python (Programming, Data Structures), Data Science & AI, Intro to Cloud Data Analytics, ETL in Python and SQL

Work Experience

Teaching Assistant, Northeastern University **Sep 2024 – Dec 2024**

- Instructed students in **Python**, database management and data analysis, offering tailored guidance towards data visualization
- Directed labs and workshops on **Tableau**/storytelling with data; resolved problems for students to deliver 20+ projects

Data Science Analyst, Tredence Analytics Solutions Pvt. Ltd., Bengaluru, India **Jun 2021 - Jul 2023**

- Conducted advanced data analysis on e-commerce data to provide actionable insights, influencing business strategies using **python** for a top US retail client
- Cleaned and organized large datasets to prepare them for analysis, ensuring increase in data accuracy through validation and quality checks by 20%
- Reduced query execution times by 30% for datasets exceeding 10M+ rows using optimized **SQL** and **PySpark** transformations
- Collaborated with cross-functional teams to enhance data integrity and security processes
- Managed code in **GitHub** for efficient version control and streamlined development processes

Data Integration Intern, SJVN Ltd., Shimla, India **Jun 2019 - Aug 2019**

- Gathered information about their different energy forms, analyzed their powerhouse tools inventory data by developing **SQL** queries to understand the stock levels and sales trends.
- Developed data integration workflows documentation to decide the entire lifecycle of the project, ensuring seamless dataflow.
- Performed data quality audits and troubleshooting to ensure data accuracy, integrity, consistency, contributing to improved decision making and operational efficiency.

Project Experience

Air Quality Prediction (View Project) **Sep 2024 – Dec 2024**

- Developed **machine learning** models to predict PM2.5 and PM10 levels using OpenAQ API
- Applied advanced **data preprocessing, feature engineering,** & **model selection** techniques to create a reliable prediction system
- Designed and implemented a comprehensive **MLOps** pipeline using **Airflow**, automating data ingestion, model retraining, and deployment of new data seamlessly through **Google Cloud Platform**
- Leveraged **MLflow** for model tracking, drift detection, and version control on **GitHub**, automating drift detection to flag accuracy deviations and enable timely retraining, maintaining performance standards across deployments

Sales Performance Optimization Dashboard **Jan 2024 – Apr 2024**

- Built an interactive **Power BI** dashboard integrating data from **SQL** and **MongoDB**, visualizing KPIs such as revenue, territory alignment, and sales trends.
- Automated data pipelines for real-time updates using Python, improving reporting efficiency by 30%.

Cricket Auction Player Performance Tracking System (View Project**)** **Sep 2023 – Dec 2023**

- Collected, cleaned, and optimized player stats, linking performance data with team outcomes for key relationship analysis
- Developed **SQL** queries and used **Python** with **Matplotlib** and **Seaborn** for visualizations for performance insights like batting average, top 10 batsmen, top 10 bowlers & bowling average.
- Utilized **NoSQL** database for unstructured data in the dataset