Sabeeha Syed

sabeeha.syed98@gmail.com | +1 (470)929-2750 | linkedin.com/in/sabeehasyed

EDUCATION

Georgia Institute of Technology, Atlanta **GPA 3.9**
Master's in Computer Science (Spl in Machine Learning) *Aug'22 - May'24*
Vellore Institute of Technology, Vellore, India **GPA 3.6**
Bachelors in Information Technology *June'16 - May'20*

SKILLS

Programming: Python, SQL, Java, R, Excel, VBA, HTML, CSS, C, C++, Alteryx, Django, Matlab
Visualization: Tableau, Power BI, Google Studio, Plotly, Excel Charts, Pandas, Scikit-Learn, Matplotlib
Database and Cloud: MySQL, Oracle, PySpark, AWS (S3, Lambda, API Gateway, Kinesis Data Stream, Firehose), Snowflake, Azure, DataBricks, Google Cloud Platform GCP, DynamoDB, Kafka, MongoDB, Agile
Data Engineering: ETL Processes, Data Pipeline Construction, Data Transformation, DBT (Data Build Tool), Git

WORK EXPERIENCE

Senior Data Engineer - Georgia Tech, Atlanta **Aug*'22 - present***

- Spearheaded a project for Dr. Feryal Ozel, building a data lake and multiple data pipelines for weather data from 9 EHT telescopes, serving 15 teams with varied output formats.
- Pioneered an **end-to-end data pipeline** using **DBT**, **Airflow and redshift** database, handling 25+ years of **46GB** EHT telescope historical and live data from VLBI. Implemented relational mapping and data warehousing technique
- Constructed an AWS-based **data ingestion** system with **API Gateway, Lambda** using **Python Boto3 SDK**, **S3, and Kinesis Firehose**, improving real-time analytics response by 20% and managing 5 million daily data points
- Configured **event-driven triggers** using **AWS S3 events, AWS SNS, IAM, AWS SQS, Lambda, AWS Step Functions**, and **CloudWatch** to automate the ingestion process and ensure real-time data availability
- Utilized **SQL** for **data extraction, transformation, and querying (ETL)** data from **Snwoflake**. Engineered complex **SQL** queries to retrieve and manipulate large datasets efficiently for SCD1 and SCD2.
- Designed a scalable data migration system for data transfer from **Snowflake** to **Google Cloud Platform** services including **BigQuery,** Cloud Composer (**Apache Airflow** on **GCP**), Cloud **Pub/Sub**, Cloud **IAM** and Cloud **Functions**

Data Science Intern - Digital Insomnia, Atlanta ***May'23 - Aug'23***

- Built a data pipeline for existing 8+ years of FB campaign data on Azure, using **Azure Data Factory**, **Data Lake Storage**, Azure **Blob Storage**, and Azure **Databricks** to process **180GB** of data from Facebook Marketing **API**
- Boosted **financial forecasting** and **risk assessment** with advanced ML models like **ARIMA, Logistic Regression, Monte Carlo Simulation** using **Python** for accurate predictions, optimizing FB ads targeting by 12% through audience segmentation and **A/B testing** leading to more accurate ad spend decisions and improved financial outcomes
- Boosted user click rates by 20% through enhanced data analysis & visualization with **Tableau dashboards** & **BI reports**
- Lead an assessment for a client and analyzed the data from Facebook AD api, and audited system architecture to identify key factors contributing to advertising behavior using **Power BI** and **PySpark**

Data Engineer - PricewaterhouseCoopers (PWC), Bengaluru, India ***Jun'20 - Aug'22***

Led DE and DS initiatives; deployed predictive models, optimized big data transformations for Fortune 500 clients

- Reduced data processing time by 16% by automating and optimizing big data transformation, handling over 2TB monthly via **PySpark ETL** operations and loading to **My SQL** servers from **AWS S3**.
- Applied **NLTK** for **NLP** tasks for tokenization and stemming, implementing **Word2Vec** and **SVM**. Achieved a notable 9% accuracy enhancement in sentiment analysis for customer service calls on **MongoDB** data using **Apache Cassandra**.
- Revised data integrity to 99.5% and cut discrepancies by 30% by intgrating **Google Pubsub** for efficient ingestion and processing of over 3M transactions daily, alongside rigorous **ETL** validation and quality assurance

PROJECTS

- **Financial Fraud Detection | Python, Scikit-Learn, TensorFlow, Power BI**
  Developed a financial fraud detection system using Random Forest and a multi-layer neural network with three hidden layers (ReLU activation). Processed 6GB of data, achieving 98% accuracy in identifying fraudulent transactions. Visualized fraud patterns and trends with Power BI, providing actionable insights to mitigate risk.
- **Adversial Attacks on LLM | PyTorch, NLTK**
  Applied a multi-model attack on MNLI dataset to flip the results of sentiment analysis using BERT, GPT models. Developed a counter defense using Adversial training to increase accuracy from 5% to 85%
- **Credit Card Approval Prediction | Python, Gradient Boosting Classifier, Smote, Tableau**
  Utilized Gradient Boosting Classifier with Python, SMOTE for data balancing, and Tableau for clear communication, achieving 90% accuracy on test set and 40% performance efficiency increase.