Saumya Mehta

(812)-837-3956 | mehta.saumya29@gmail.com | linkedin.com/saumya | github.com/saumya | portfolio

Data Science masters student with 3+ years of experience in optimising data pipelines, architecting scalable and reliable software and deploying end-to-end machine learning pipelines on cloud platforms

EDUCATION

Indiana University — Bloomington, IN, USA
Masters of Science in Data Science CGPA: 3.93 — Aug 2021 – May 2023

Nirma University — Ahmedabad, GJ, INDIA
Bachelor in Technology in Computer Science — Jun 2014 – May 2018

TECHNICAL SKILLS

Python, C++, Java, PostgreSQL, R, Docker, Kubernetes, AWS, GCP, Kafka, Spark, Hadoop, Airflow, Grafana, Glue, PyTorch, Snowflake, Jenkins, Terraform, BigQuery, spacy, NLTK

WORK EXPERIENCE

Pearson(Savvas) — May 2022 – Aug 2022
Software Development Engineer Intern — Boston, MA

- Reduced query execution times by **5x** through migration and optimisation of PostgreSQL queries to Redshift SQL using Common Table Expressions and User Defined Functions.
- Streamlined the data analytics pipeline for the Learning Analytics team by creating external views in **Amazon Redshift DB** and **automated query scheduling** on AWS for **ETL** tasks
- Developed **Curriculum Recommendation** algorithms for the **SuccessMaker** engine using **Bayesian Item Response Theory** resulting in a **40%** improvement in student test scores

Playpower Labs — Jun 2018 – Aug 2021
Data Scientist — Remote

- Architected and deployed end-to-end machine learning pipelines on **Snowflake**, handling up to **1 million records per second** and delivering a **25%** improvement in model performance and a **50%** reduction in data processing time
- Designed and implemented scalable and fault-tolerant data processing systems using **Apache Spark** and AWS Glue for a data warehouse handling **petabyte-scale** data, delivering a **60%** reduction in ETL runtime
- Enhanced real-time data processing capabilities by utilising **Spark Streaming**, **R**, and **Kafka**, delivering a **50%** improvement in processing speed and scalability for a distributed computing setup handling over **100,000** records per second of streaming data
- Utilized **Jenkins** to design and automate **CI/CD** pipelines for deploying data pipelines and machine learning models into production, delivering a **40%** reduction in deployment time and a **20%** reduction in deployment errors
- Implemented and optimized **ETL** pipelines using **Apache Airflow** and **AWS Glue**, delivering a **30%** reduction in ETL runtime and improving data quality and reliability
- Spearheaded and directed cutting-edge **machine learning research** in Computer Vision and Natural Language Processing
- Optimised Machine Learning algorithms for performance and scalability using **Apache Fink** and Kafka streams and deployed in real-time environments using **MLFlow**
- Optimized data processing and **query performance** on Amazon Redshift by **3x** through the migration and optimization of SQL queries and implementing performance tuning

Indiana University — Jan 2022 – May 2023
Software Development Engineer - Graduate Research Assistant — Remote

- Worked on CompuCell3D, an open-source software extensively used for 3D simulations in computational biology
- Designed and implemented a **high-performance processing pipeline** for 3D cell simulations using CUDA, achieving a processing speedup of **10x**
- Created a scalable architecture to handle **100+** parallel simulations, resulting in a **90%** reduction in simulation time

PUBLICATIONS

Using Curriculum Pacing in Learnsphere to Visualize Student Learning Trajectories **[Paper]** — Mar 2019
Sharing and Reusing Data and Analytic Methods with LearnSphere conference

Advanced Chemical Transport Modeling in Dynamic Multicellular Contexts Using CompuCell3D
In preparation

PROJECTS

Abusive Language Detection in User Tweets (SemEval-2021) **[Github]**
Techstack: PyTorch, tensorboardX, RayTune, NumPy, scikit-learn, pandas,scipy, contractions, fair

- Developed and implemented a **tweet classification** model using Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and a combination of CNN and LSTM architectures.
- Improved model performance by **8%** in terms of **Macro-F1** scores compared to the baseline model using **CNN+LSTM** architecture.
- Leveraged **GloVE** and **FastText** embeddings to capture contextual information from tweets and learned more accurate word representations resulting in a **6% to 8%** improvement in **Macro-F1** score over the CNN+LSTM model.
- Utilized tokenization, stemming, and stop-word removal, to improve the data quality and achieve more accurate predictions

Paperflow(Smart Paper) **[Case Study]** **[Product Website]**
Techstack: Java, C++, Kotlin, Android Jetpack, OpenCV, JNI, Volley, Glide

- Designed and developed a robust Android app using Java, Kotlin, and Android Jetpack to enable adaptive assessment generation for students, resulting in improved learning outcomes.
- Utilised **multithreading** and IPC for efficient data transfer resulting in a **40% increase** in **UI responsiveness**
- Implemented **Lazy loading** and **Data Caching** resulting in **a 50% improvement** in app performance on **low-end** devices
- Developed and maintained data pipelines to ensure the efficient processing and ingestion of student data into the system